Performance Evaluation of Consumer Decision Support Systems

Jiyong Zhang, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Pearl Pu, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Consumer decision support systems (CDSSs) help online users make purchasing decisions in e-commerce Web sites. To more effectively compare the usefulness of the various functionalities and interface features of such systems, we have developed a simulation environment for decision tasks of any scale and structure. Furthermore, we have identified three criteria in an evaluation framework for assessing the quality of such CDSSs: users' cognitive effort, preference expression effort, and decision accuracy. A set of experiments carried out in such simulation environments showed that most CDSSs employed in current e-commerce Web sites are suboptimal. On the other hand, a hybrid decision strategy based on four existing ones was found to be more effective. The interface improvements based on the new strategy correspond to some of the advanced tools already developed in the research field. This result is therefore consistent with our earlier work on evaluating CDSSs with real users. That is, some advanced tools do produce more accurate decisions while requiring a comparable amount of user effort. However, the simulation environment will enable us to efficiently compare more advanced tools among themselves, and indicate further opportunities for functionality and interface improvements.

Keywords: *consumer decision support systems; decision accuracy; electronic catalogs; elicitation effort; extended effort-accuracy framework; multi-attribute decision problem; performance-based search; performance evaluation; product recommender systems.*

INTRODUCTION

With the rising prosperity of the World Wide Web (WWW), consumers are dealing with an increasingly large amount of product and service information that is far beyond any individual's cognitive effort to process. In early e-commerce practice, online intermediaries were created. With the help of these virtual storefronts, users were able to find product information on a single Web site that gathers product information from thousands of merchants and service suppliers. Examples include shopping.yahoo.com, froogle.com, shopping.com, cars.com, pricegrabber.com, and so forth.

However, due to the increasing popularity of electronic commerce, the amount of online retailers proliferated. As a result, there are now easily millions (or 16-20 categories) of brand-name products available on a single online intermediary Web site. Finding something is once again difficult, even with the help of various commercially available search tools.[1] Recently, much attention in e-commerce research has focused on designing and developing more advanced search and product recommender tools (Burke, Hammond, & Young, 1997; Pu & Faltings, 2000; Reilly, McCarthy, McGinty, & Smyth, 2004; Shearin & Lieberman, 2001; Shimazu, 2001; Stolze, 1999). However, they have been not employed in large scales in practicing e-commerce Web sites. Pu and Kumar (2004) gave some reasons as to why this is the case and when such advanced tools are expected to be adopted. This work was based on empirical studies of how users interact with product search tools, providing a good direction as to how to establish the true benefits of these advanced tools. However, insights gained from this work are limited. This is mainly due to the lack of a *large* amount of *real* users for the needed user studies and the high cost of user studies, even if real users were found. Each of the experiments reported in Pu and Kumar (2004) and Pu and Chen (2005) took more than 3 months of work, including the design and preparation of the study, the pilot study, and the empirical study itself. After the work was finished, it remains unclear whether a small amount of users recruited in an academic institution can forecast the behavior of the actual user population, which is highly diverse and complex.

Our main objective in this research is to use a simulation environment to evaluate various search tools in terms of interaction behaviors: what users' effort would be to use these tools and what kind of benefits they are likely to receive from these tools. We base our work on some earlier work (Payne, Bettman, & Johnson, 1993) in terms of the design of the simulation environment. However, we have added important elements to adapt such environments to online e-commerce and consumer decision support scenarios. With this simulation environment, we hope to more accurately forecast the acceptance of research tools in the real world, and curtail the evaluation of each tool's performance from months of user study to hours of simulation and a week of fine tuning the simulation results against a small but diverse amount of real users. This should allow us to evaluate more tools and, more importantly, discover design opportunities of new tools.

Our initial work of measuring the performance of various decision support strategies in e-commerce environments was reported in a conference paper (Zhang & Pu, 2005). The current article is an extended version of the conference paper. Besides adding significantly more details on the work already reported, there are a number of important and new contributions:

- In the conference paper, we only reported the performance evaluation results of various decision strategies such as the lexicographical (LEX) strategy, the elimination-by-aspects (EBA) strategy, and so forth; in this paper, we consider the evaluation of a consumer decision support system as an integral unit compromising decision strategies, user interfaces, and the underlying product catalog;
- In the extended effort-accuracy framework described in the conference paper, we only used a classical definition of decision accuracy; here we propose two new definitions of decision accuracy that correspond more precisely with a user's choice behavior in e-commerce situations rather than the classical choice problem in decision literature;
- Based on the new definitions, we were able to draw more conclusions from the simulation results: not only can we establish that hybrid decision approaches can reduce user's effort while achieving a high level of decision accuracy, but we can also see some opportunities for improving consumer decision support systems by designing better interfaces and decision approaches, courtesy of the simulation environment.

This paper is organized as follows: the second section reviews some related research work; the third section defines the consumer decision support system (CDSS) and clarifies its relationship with our earlier published concepts on multi-attribute decision problem (MADP) and various decision strategies; the fourth section describes in detail the simulation environment for the performance evaluation of CDSSs; the fifth section describes the extended effort-accuracy framework consisting of three performance criteria: cognitive effort, elicitation effort, and decision accuracy; the sixth section reports the performance evaluation of various CDSSs with respect to a set of simulated MADPs and user preferences; the seventh section discusses the main research results obtained, followed by the conclusion section.

RELATED WORK

In traditional environments where no computer aid is involved, behavioral decision theory provides adequate knowledge describing people's choice behavior, and presents typical approaches of solving decision problems. For example, Payne et al. (1993) established a well-known effort-accuracy framework that described how people adapted their decision strategies by trading off accuracy and cognitive effort to the demands of the tasks they faced. The simulation experiments carried out in that work were able to give a good analysis of various decision strategies that people employ, and the decision accuracy they would expect to get in return.

In the online electronic environment where the support of computer systems is pervasive, we are interested in analyzing users' choice behaviors when tools are integrated into their information processing environments. That is, we are interested in analyzing when given a computer tool with its system logic, how much effort a user has to expend and how much decision accuracy he or she is to obtain from that tool. On one hand, though the decision maker's cognitive effort is still required, it can be significantly decreased by having computer programs carry out most of the calculation work automatically; on the other hand, the decision makers must expend some effort to explicitly state their preferences to the computer according to the requirements of the underlying decision support approach employed in that system. We would like to call this extra user effort (in addition to the cognitive effort) preference elicitation effort. We believe that elicitation effort plays an important role in the new effort-accuracy model of users' behavior in online choice environments.

Many other researchers have carried out simulation experiments in evaluating the performance of their systems or approaches. Payne et al. (1993) introduced a simulation experiment to measure the performance of various decision strategies in off-line situations. Recently, Boutilier (Boutilier, Patrascu, Poupart, & Schuurmans, 2005) carried out their experiments by simulating a number of randomly generated synthetic problems, as well as user responses to evaluate the performance of various query strategies for eliciting bounds of the parameters of utility functions. In Reilly et al. (2005), various users' queries were generated artificially from a set of off-line data to analyze the recommendation performance of the incremental critiquing approach. These related works generally suggest that simulating the interaction between users and the system is a promising methodology for performance evaluation. In our work, we go further in this direction and propose the general simulation environment that can be adopted to evaluate the performance of various CDSSs systematically within the extended effort-accuracy framework. To the best of our knowledge, our simulation work is the first attempt in systematically evaluating the performance of various CDSSs with simulation methodology.

CONSUMER DECISION SUPPORT SYSTEM

In a scenario of buying a product (such as a digital camera), the objective of consumers is to choose the product that most closely sat-

isfies their needs and preferences (decision result), and furthermore, they are not likely to regret the products that they have bought (how accurate their decision is). They usually face a large amount of product alternatives (or options) and need a decision support system in order to process the entire product catalog without having to examine all items exhaustively. Therefore, a consumer decision support system consists of three components: (1) the product catalog that is accessible to consumers via an interface; (2) the underlying decision support approach that helps a consumer to choose and determine the product most satisfying his or her preferences; (3) the user interface with which a consumer interacts in order to state his or her preferences. We will now introduce these three components respectively.

The product catalog or more precisely an electronic product catalog (EPC) (Palmer, 1997; Torrens, 2002) provides a list of products, each one represented by a number of attributes. The process of determining the most preferred product from the EPC can be formally described as solving a multi-attribute decision problem[2] $\mathbf{\Psi} = \langle \mathbf{X, D, O, P} \rangle$, where $\mathbf{x}=\{x_1, ...,x_n\}$ is a finite set of attributes the product catalog has, $\mathbf{D}=D_1\times...\times D_n$ indicates the space of all possible products in the catalog (each $D_i(1\le i\le n)$ is a set of possible domain values for attribute X_i), $\mathbf{O}=\{O_1,...,O_m\}$ is a finite set of available products (also called alternatives or outcomes) that the EPC offers, and $\mathbf{P}=\{P_1,...,P_t\}$ denotes a set of preferences that the decision maker may have. Each preference P_i may be identified in any form as required by the decision methods. The solution of a MADP is an alternative O most satisfying the decision maker's preferences.

In traditional decision making environments, consumers usually adopt various decision strategies such as EBA or LEX to obtain decision results (Payne et al., 1993). In a computer assisted scenario, the distribution of work is quite different. It is the CDSS that will perform these decision strategies to help the consumer to make decisions. The consumer is only required to input his or her preferences as required by the specific decision strategy; and then the solution can be chosen for the consumer automatically. When a decision strategy is adopted in the consumer decision support system, we also say it is a decision support approach for that system. In this paper, we focus on the following decision strategies and study the performance of CDSSs based on these decision strategies.

1. The weighted additive (WADD) strategy. It is a normative approach based on multi-attribute utility theory (MAUT) (Keeney & Raiffa, 1993). In our simulation experiment, we use it as the baseline strategy.
2. Some basic heuristic strategies. They are the equal weight (EQW) strategy, the elimination-by-aspects strategy, the majority-of-confirming dimensions (MCD) strategy, the satisficing (SAT) strategy, the lexicographic (LEX) strategy and the frequency of good and bad features (FRQ) strategy. Their detailed definitions can be found in Payne et al., 1993 and Zhang & Pu, 2005.
3. Hybrid decision strategies. Besides the basic heuristic strategies, people may also use a combination of several of them to make a decision to try to get a more precise decision result. These kinds of strategies are called hybrid decision strategies. As a concrete example of hybrid decision strategies, The C4 strategy (Zhang & Pu, 2005), which is a combination of four basic heuristic strategies: EBA, MCD, LEX, and FRQ, is also studied in this paper.

In a CDSS, the user-interface component is used to obtain the consumers' preferences. However, such preferences are largely determined by the underlying decision support approach that has been adopted in the system. For example, the popular ranked list interface is in fact an interface implementing the lexicographical strategy. Also, if we adopt the weight additive strategy in a consumer decision support system, the user interface will be designed in the manner of asking the user to input corresponding weight and middle values for each

attribute. In our current work, we assume the existence of a very simple user interface. Thus, we regard the underlying decision support approach as the main factor of the consumer decision support system.

SIMULATION ENVIRONMENT FOR PERFORMANCE EVALUATION

Our simulation environment is concerned with the evaluation of how users interact with consumer decision support systems (CDSSs), how decision results are produced, and the quality of these decision results.

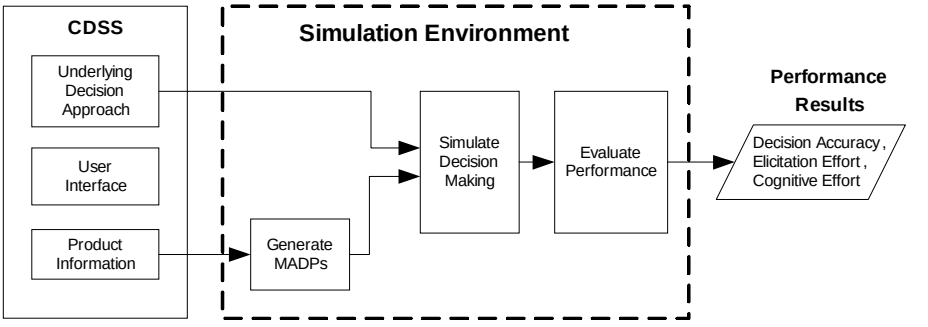
The consumer first interacts with the system by inputting his or her preferences through the user interface. With the help of decision support, the system generates a set of recommendations to the consumer. This interactive process can be executed multiple times until the consumer is satisfied with the recommended results (i.e., a product to purchase) or gives up due to loosing patience.

As shown in Figure 1, for a given CDSS, we evaluate its performance in a simulated environment by the following procedure: (1) we generate a set of MADPs using Monte Carlo method to simulate the presence of an electronic catalog up to any scale and structure characteristics; (2) we generate a set of consumer preferences also with the Monte Carlo method, taking into account user diversity and scale; (3) we carry out the simulation of the underlying decision approach of the CDSS to solve these MADPs; (4) we obtain associated decision results for the given CDSS (which product has been chosen given the consumer's preferences); and finally, (5) we evaluate the performance of these decision results in terms of cognitive effort, preference elicitation effort, and decision accuracy under the extended accuracy-effort framework (detailed discussion of this framework in the next section).

The simulation environment can be used in many ways to provide different performance measures of a given CDSS. For instance, if both the detail product information of CDSS and the consumer's preferences are unknown, we can simulate both the alternatives and the consumer's preferences, and the simulation results would be the performance of the CDSS independently of users and the set of alternatives; if the detail product information of the CDSS is provided, we then only need to simulate the consumer's preferences, and the alternatives of the MADPs can be copied from the CDSS instead of being randomly generated. The simulation results would be the performance of the CDSS under the specified product set.

As a concrete example to demonstrate the usage of such a simulation environment,

Figure 1. An architecture of the simulation environment for performance evaluation of a given consumer decision support system



we will show a procedure in evaluating the performance of various CDSSs in terms of the scale of the MADPs, which is determined by two factors: the number of attributes n and the number of alternatives m. Since we are trying to study the performance of different CDSSs (currently built on heuristic decision strategies) in different scales of MADPs, we assume that users and alternatives are both unknown and they are simulated to give results independently of the user and the system. More specifically, we classify the decision problems into 20 categories according to the scales of n (the number of attributes) and m (the number of alternatives): n has five values (5, 10, 15, 20, and 25), and m has four (10, 100, 1,000, and 10,000). To make the performance evaluation result more accurate, each time we randomly generate 500 different MADPs in the same scale and use their average performance as the final result. The detail simulation result will be reported in the experimental result section.

THE EXTENDED EFFORT-ACCURACY FRAMEWORK

The performance of the system can be evaluated by various criteria such as the degree of a user's satisfaction with the recommended item, the amount of time a user spends to reach a decision, and the decision errors that the consumer may have committed. Without real users' participation, the satisfaction of a consumer with a CDSS is hard to measure. However, the other two criteria can be measured.

The amount of time a user spends to reach a decision is equivalent to the amount of time he or she uses to express preferences and process the list of recommended items in order to reach a decision. The classical effort-accuracy framework mainly investigated the relationship of decision accuracy and cognitive effort of processing information by different decision strategies in the off-line situation. In the online decision support situation, however, the effort of eliciting preferences must be considered as well.

Furthermore, most products carry a fair amount of financial and emotional risks. Thus, the accuracy of users' choices is extremely important. That is, there is a posterior process where users evaluate the search tools in terms of whether the products they have found via the search tool are really what they want and whether they had enough decision support. This is what we mean by decision accuracy.

We therefore propose an extended effort-accuracy framework by explicitly measuring three factors of a given consumer decision support system: cognitive effort, elicitation effort, and decision accuracy. In the remainder of this section, we first recall the measurement of cognitive effort in the classical framework, we give various definitions of accuracy, and then we detail the method of measuring elicitation effort. Finally, the cognitive and elicitation effort of these decision strategies are analyzed in section 5.4, in an online situation.

Measuring Cognitive Effort

Based upon the work of Newell and Simon (1972), a decision approach can be seen as a sequence of elementary information processes (EIPs), such as reading the values of two alternatives on an attribute, comparing them, and so forth. Assuming that each EIP takes equal cognitive effort,[3] the decision maker's cognitive effort is then measured in terms of the total number of EIPs. Conformed with the classical framework, a set of EIPs for the decision strategies is defined as (1) READ: read an alternative's value on an attribute into short-term memory (STM), (2) COMPARE: compare two alternatives on an attribute, (3) ADD: add the values of two attributes in STM, (4) DIFFERENCE: calculate the size of the difference of two alternatives for an attribute, (5) PRODUCT: weight one value by another, (6) ELIMINATE: eliminate an alternative from consideration, (7) MOVE: move to next element of the external environment, and (8) CHOOSE: choose the preferred alternative and end the process.

Measuring Decision Accuracy

Accuracy and effort form an important performance measure for the evaluation of consumer decision support systems. On one hand, consumers expect to get highly accurate decisions. On the other hand, they may not be inclined (or able) to expend a high level of cognitive and elicitation effort to reach a decision. Three important factors influence the decision accuracy of a consumer decision support systems: the underlying decision approach used; the number of interactions required from the end user (if a longer interaction is required, a user may give up before he finds the best option); the number of options displayed to the end user in each interaction cycle (a single item is likely to miss the target choice compared to a list of items; however, a longer list of items requires more cognitive effort to process information). In our current framework, we investigate the combined result of these three factors (i.e., decision approach as well interface design components) of a given consumer decision support system.

In the following sections, we start with classical definitions of decision accuracy, analyze their features and describe their weaknesses for the online environments, and then we propose two definitions, which we have developed, that are likely to be more adequate for measuring decision accuracy in e-commerce environments. To eliminate the effect of a specific set of alternatives on the decision accuracy results, in the following definitions we assume that there is a set of N different MADPs to be solved by a given consumer decision support system that implements a particular decision strategy *S*. The accuracy will be measured in average among all those MADPs.

Accuracy Measured by Selection of Nondominated Alternatives

This definition comes from Grether and Wilde (1983). After adapting it to decision making with the help of a computer system, this definition says that a solution given by CDSS is correct if and only if it is non-dominated by other alternatives. So the decision accuracy can be measured by the numbers of solutions which are *Pareto optimal* (i.e., not dominated by other alternatives, see also Viappiani, Faltings, Schickel-Zuber, & Pu, 2005). We use O_S^i to represent the optimal solution given by the CDSS with strategy *S* when solving $MADP_i (1 \le i \le N)$. The accuracy of selection of nondominated alternatives $Acc_{NDA}(S)$ is defined as the following:

$$Acc_{NDA}(S) = \frac{N - \sum_{i=1}^{N} \text{Dominated}(O_S^i)}{N}, \quad (1)$$

where

$$\text{Dominated}(O_S^i) = \begin{cases} 1 & \text{if } O_S^i \text{ is dominated} \\ & \text{in } MADP_i \ (1 \le i \le N) \\ 0 & else \end{cases}$$

According to this definition, it is easy to see that a system employing the WADD strategy has 100% accuracy because all the solutions given by WADD are *Pareto optimal*. Also, this definition of accuracy measurement is effective only when the system contains some dominated alternatives, otherwise the accuracy of the system is always 100%.

This definition of accuracy can distinguish the errors caused by choosing dominated alternatives of the decision problems. However, measuring decision accuracy using this method is limited in e-commerce environments. In an efficient market, it is unlikely that the consumer products or business deals are dominated or dominating. That is, it is unlikely an apartment would be both spacious and less expensive compared to other available ones. We believe that although this definition is useful, it is not helpful to distinguish various CDSSs in terms of how good they are for supporting users to select the best choice (not just the nondominated ones).

Accuracy Measured by Utility Values

This definition of measuring accuracy was used in the classical effort-accuracy framework (Payne et al., 1993). Since no risk or uncertainty is involved in the MADPs, the expected value of an alternative is equivalent to the utility value of each alternative. The utility value of each alternative is assumed to be in the weight additive form. Formally, this accuracy definition can be represented as:

$$Acc_{UV}(S) = \frac{\sum_{i=1}^{N} \frac{V(O_S^i)}{V(O_{WADD}^i)}}{N}, \qquad (2)$$

where $V(O_S^i)$ is the value function given by the WADD strategy in $MADP_i$. In this definition, a system employing the WADD strategy is also 100% accurate because it always gives out the solution with the maximal utility value.

One advantage of this measure of accuracy is that it can indicate not only that an error has occurred, but also the severity of the error of the decision making. For instance, a system achieving 90% accuracy indicates that an average consumer is expected to choose an item that is 10% less valuable from the best-possible option. While this definition is useful for choosing a set of courses to take for achieving a particular career objective, it is not most suitable in e-commerce environments. Imagine that someone has chosen and purchased a digital camera. Two months later, she discovers that the camera that her colleague has bought was really the one she wanted. She did not see the desired camera, not because the online store did not have it, but because it was difficult to find and compare items on the particular e-commerce Web site. Even though the camera that she bought satisfied some of her needs, she is still likely to feel a great sense of regret, if not outright disappointment. Her likelihood of returning to that Web site is in question. Given that bad choices can cause great emotional burdens (Luce, Payne, & Bettman, 1999), we have developed the following definition of decision accuracy.

Accuracy Measured by Selection of Target Choice

In our earlier work(Pu & Chen, 2005), we defined decision accuracy as measured by the percentage of users who have chosen the right option using a particular decision support system. We call that option the target choice. In empirical studies with real users, we first asked users to choose a product with the consumer decision support system, and then we revealed the entire database to them in order to determine the target choice. If the product recommended by the consumer decision support system was consistent with the target choice, we said that the user had made an accurate decision.

In simulation environment, we take the WADD strategy as the baseline. That is, we assume the solution given by WADD is the user's final most-preferred choice. For another given strategy *S*, if the solution O_S^i is the same as the one determined by WADD, then we count it as one *hit* (this definition is called the *hit ratio*). The accuracy is measured statistically by the ratio of hit numbers to the total number of decision problems:

$$Acc_{HR}(S) = \frac{\sum_{i=1}^{N} Hit(O_S^i)}{N}, \qquad (3)$$

where

$$Hit(O_S^i) = \begin{cases} 1 & \text{if } O_S^i = O_{WADD}^i \text{ in } MADP_i \\ 0 & \text{else} \end{cases}$$

This measure of decision accuracy is ideally consistent with the consumers' attitude towards the decision results. However, by this definition, it is assumed that the consumer decision support system only recommends one product to the consumer each time. In reality, the system may show a list of possible products to the consumer, and the order of the product list is also important to the consumer: the products displayed at the top of the list are

more likely to be selected by the consumer. Therefore, we have developed the following definition to take into account that a list of products is displayed, rather than a single product.

Accuracy Measured by Selection of Target Choice Among K-best Items

Here we propose measuring the accuracy of the system according to the ranking orders of the K-best products it displays. This is an extension of the previous definition of accuracy. For a given $MADP_i$, instead of using strategy S to choose a single optimal solution, we can use it to generate a list of solutions with ranking order $L_S^i = \{O_{S,1}^i, O_{S,2}^i, \ldots, O_{S,K}^i\}$, where $O_{S,1}^i$ is the most-preferred solution according to the strategy S, and $O_{S,2}^i$ is the second-preferred solution, and so on. The first *K-best* solutions consist of the solution list. If the user's final choice (which is assumed to be given by the WADD strategy O_{WADD}^i) is in the list, we assign a *rank value* to the list according to the position of O_{WADD}^i in the list. Formally, we define this accuracy of choosing K-best items as:

$$Acc_{HR\_in\_Kbest}(S) = \frac{\sum_{i=1}^{N} RankValue(L_S^i)}{N}, \quad (4)$$

where

$$RankValue(L_S^i) = \begin{cases} 1 - \dfrac{k-1}{K} & \text{if } O_{S,k}^i = O_{WADD}^i \text{ in } MADP_i \\ & (1 \le k \le K, 1 \le i \le N) \\ 0 & else \end{cases}$$

According to this definition, the WADD strategy still achieves 100% accuracy and is used as the baseline. A special case of this accuracy definition is that when $K=1$, it degenerates to the previous definition of *hit ratio*. In the simulation experimental results that we will show shortly, we have set K to 5.

In practice, it is required to eliminate the effect of random decision, and we expect that the strategy of *random choice* (selecting an alternative randomly from the alternative set, denoted as *RAND* strategy) could only produce *zero* accuracy. By doing so, we define the *relative accuracy* of the consumer decision support system with strategy S according to different definitions as:

$$RA_Z(S) = \frac{Acc_Z(S) - Acc_Z(RAND)}{1 - Acc_Z(RAND)}, \quad (5)$$

where

Z= NDA, UV, HR, or HR_in_Kbest.

For example, $RA_{HR}(LEX)$ denotes the relative accuracy of the *LEX* strategy under the accuracy measure definitions of *hit ratio*.

From the previous definitions, we can see that each definition represents one aspect of the accuracy of the decision strategies. We think that the definitions of *hit ratio* and *K-best items* are more suitable to measure the accuracy of various consumer decision support systems, particularly in e-commerce environments. In the sixth section, we will study the performance of various decision strategies with these accuracy measurement definitions.

Measuring Elicitation Effort

In computer-aided decision environments, a considerable amount of decision effort goes into preference elicitation since people need to "tell" their preferences explicitly to the computer system. So far, no formal method has been given to measure the preference elicitation effort. An elicitation process can be decomposed into a series of basic interactions between the user and the computer, such as selecting an option from a list, filling in a blank, answering a question, and so forth. We call these basic interaction actions elementary elicitation processes (EEPs). In our analysis, we define the set of EEPs as follows: (1) SELECT: select an item from a menu or a dropdown list,

(2) FILLIN: fill in a value to an edit box, (3) ANSWER: answer a basic question, (4) CLICK: click a button to execute an action.[4]

It is obvious that different EEPs require different elicitation effort (for instance, the EEP of one CLICK would be much easier than an EEP of FILLIN a weight value for a given attribute). For the sake of simplification, we currently assume that each EEP requires an equal amount of effort from the user. Therefore, given a specific decision approach, elicitation effort is measured by the total amount of EEPs it may require.

This elicitation effort is a new factor for the online environment. The main difference between cognitive effort and elicitation effort lies in the fact that cognitive effort is a description of the mental activities in processing information, while the elicitation effort is about the interaction effort between the decision maker and the computer system through predesigned user interfaces. Even though the decision makers already have clear preferences in their mind, they must still state their preferences in a way that the computer can "understand." With the help of computer systems, the decision maker is able to reduce the cognitive effort by compensating with a certain degree of elicitation effort.

Let us consider a simple decision problem with three attributes and four alternatives. When computer support is not provided, the cognitive effort of solving this problem by the WADD strategy will be 24 READS, 8 ADDS, 12 PRODUCTS, 3 COMPARES, and 1 CHOOSE. The total number of EIPs is therefore 48.[5]

However, with the aid of a computer system, the decision maker could get the same result by spending two units of elicitation effort (FILLIN the weight value of first two attributes) and one unit of cognitive effort (CHOOSE the final result).

Analysis of Cognitive and Elicitation Effort

With the support of computer systems, the cognitive effort for WADD, as well as the basic heuristic strategies, is quite low. The decision maker inputs his or her preferences, and the decision support system executes that strategy and shows the proposed product. Then the decision maker chooses this product and the decision process is ended. Thus, the cognitive effort is equal to one EIP: CHOOSE the final alternative and exit the process. For the *C4* strategy, the cognitive effort of solving an MADP with *n* attributes and *m* alternatives is equal to that of solving a problem with *n* attributes and *4* alternatives in the traditional situation, the cognitive effort of which has been studied in (Payne et al., 1993).

According to their definitions, various decision strategies require that preferences with different parameters be elicited. For example, in the WADD strategy, the component value function and the weight for each attribute must be obtained, while for the EBA strategy, the importance order and cutoff value for each attribute are required. The required parameters for each strategy are shown in Table 1.

For each parameter in the aforementioned strategies, a certain amount of elicitation effort is required. This elicitation effort may vary with different implementations of the user interface. For example, to elicit the weight value of an attribute, the user can just FILLIN the value to an edit box, or the user can ANSWER several questions to approximate the weight value. In our analysis and the following simulation experiments, we follow the *at least* rule: the elici-

Table 1. Elicitation effort analysis of decision strategies

Strategy	Parameters required to be elicited
WADD	Weights, component value functions
EQW	Component value functions
EBA	Importance order, cutoff values
MCD	*None*
SAT	Cutoff values
LEX	Importance order
FRQ	Cutoff values for good and bad features
C4	Cutoff values, importance order

tation effort is determined by the least number of EEP(s). In the above example, the elicitation effort for obtaining a weight value is measured as 1 EEP.

SIMULATION RESULTS

In this section, we report our experimental results of the performance of various consumer decision support systems under the simulation environment that was introduced earlier. To simplify the experiments, we only evaluate those CDSSs built on the decision strategies listed in Table 1. Without loss of generality, we will also use the term *decision strategy* to represent the CDSS built on that decision strategy.

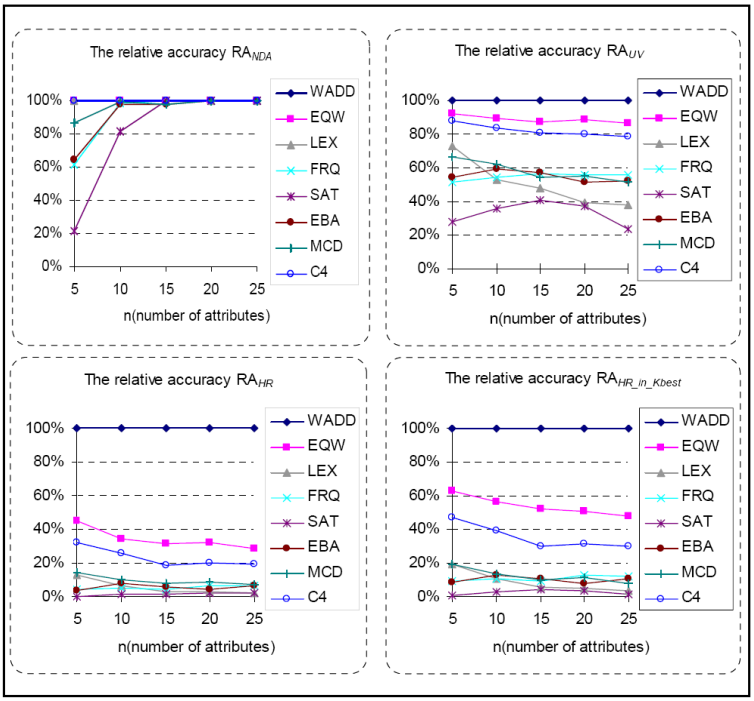
For each CDSS, we first simulate a large variety of MADPs, and then run the corresponding decision strategy on the computer to generate the decision results. Then the elicitation effort and decision accuracy are calculated according to the extended effort-accuracy framework. For each MADP, its domain values for a given attribute are determined randomly: the lower bound of each attribute is set to 0, and the upper bound is determined randomly from the range of 2 to 100. Formally speaking, for each attribute X_i, we define $D_i = [0,z_i]$ where $z_i \in [2,100]$.

As shown in Table 1, each decision strategy (except MCD) requires the elicitation of some specific parameters such as attribute weights or cutoff values to represent the user's preferences. To simulate the component value functions required by the WADD strategy, we assume that the component value function for each attribute is approximated by three midvalue points, which are randomly generated.[6] Thus, each component value function requires three units of EEPs. Other required parameters, such as the weight and cutoff value (each requires one unit of EEP) for each attribute are also simulated by the random generation process. The order of importance is determined by the weight order of the attributes for consistency.

In our simulation experiments, the WADD strategy is appointed as the baseline strategy, and the relative accuracy of a strategy is calculated according to equation (5). The elicitation effort is measured in terms of the total number of EEPs required by the specific strategy, and the cognitive effort is measured by the required units of EIPs. Since the relationship between accuracy and cognitive effort has already been studied and analyzed by Payne et al. (1993), in this section, we only focus on the performance of each strategy in terms of decision accuracy and elicitation effort.

Figure 2 shows the changes in *relative accuracy* with four different accuracy measure definitions for the listed decision strategies as *the number of attributes* increases in the case that each MADP has 1,000 alternatives. In all cases, the WADD is the baseline strategy; thus it achieves 100% accuracy. When measured by the selection of nondominated alternatives (RA_{NDA}), the relative accuracy of each heuristic strategy increases as the number of alternatives increases. This is mainly because the alternatives are more likely to be *Pareto optimal* when more attributes are involved. Furthermore, the RA_{NDA} of all strategies could achieve 100% accuracy when the attributes number is 20 or 25. This shows that the RA_{NDA} is not able to distinguish the decision errors occurred with the heuristic strategies in the simulated environment. When the accuracy is measured under the definitions of RA_{UV}, RA_{HR} and $RA_{HR\_in\_Kbest}$, the EQW strategy achieves the highest accuracy besides the baseline WADD strategy, and the SAT strategy has the lowest relative accuracy. The four basic heuristic strategies EBA, MCD, LEX, and FRQ are in the middle-level range. The LEX strategy, which has been widely adopted in many consumer decision support systems, is the least accurate strategy among the EBA, FRQ, and MCD strategies when there are over 10 attributes. When the accuracy is measured by RA_{UV}, the EQW strategy could gain over 90% relative accuracy, while it could only achieve less than 50% relative accuracy when measured by RA_{HR}. This comparison generally suggests that most of the decision results given by EQW strategy may be very close to a user's target choice (which is determined by the WADD strategy), but are not identical. Also, in all cases, the accuracy

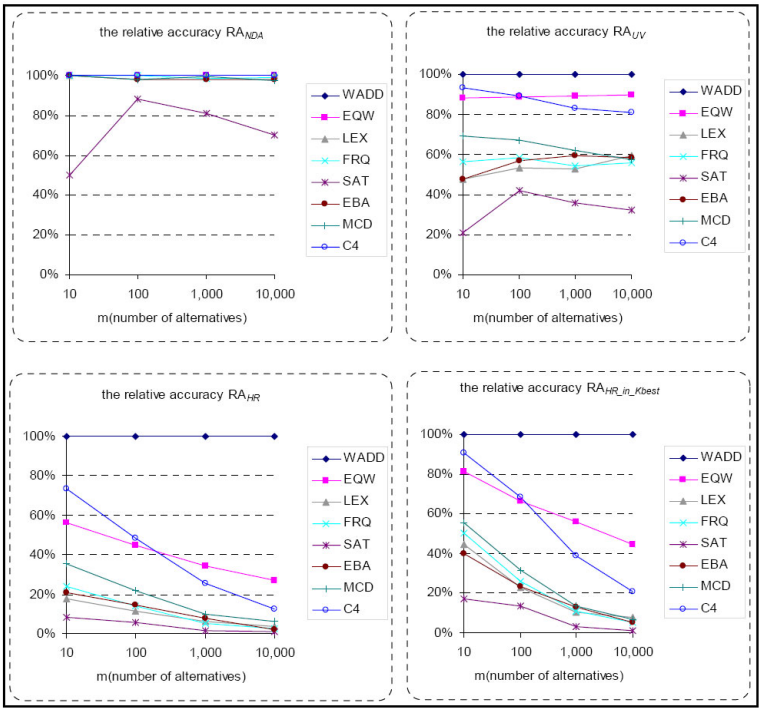
Figure 2. The relative accuracy of various decision strategies when solving MADPs with different number of attributes, where m*(number of alternatives)* = 1,000



measured by $RA_{HR\_in\_Kbest}$ (where $K=5$ in the experiment) is always higher than that measured by RA_{HR} (which is a special case of $RA_{HR\_in\_Kbest}$ when $K=1$). This shows that under this definition, the possibility of containing the final target choice in a K-item list is higher when K is larger. Of particular interest is that the proposed *C4* strategy, which is a combination of the four basic strategies, could achieve a much higher accuracy than any of them alone. For instance, when there are 10 attributes and 1,000 alternatives in the MADPs, the relative accuracy of *C4* strategy could exceed the average accuracy of the four basic strategies by over 27% when the definition of $RA_{HR\_in\_Kbest}$ is adopted.

Figure 3 shows the relationship between *relative accuracy* and *the number of alternatives* (or the number of available products in a catalog) for the listed decision strategies. When the accuracy is measured by the selection of nondominated alternatives (RA_{NDA}), all strategies except SAT could gain nearly 100% of relative accuracy without a significant difference. This generally shows that the RA_{NDA} is not a good definition of accuracy measurement in the simulated environment. When the accuracy is measured by the utility values (RA_{UV}), the accuracy of the heuristic strategies remains stable as the number of alternatives increases. With the definitions of *hit ratio* (RA_{HR}) and *hit ratio in K-best items* ($RA_{HR\_in\_Kbest}$), however, the heuristic strategies strongly descend into a lower range of accuracies as the size of a catalog increases. This corresponds to the fact that consumers have increasing difficulties finding the best product as the number of alternatives in the catalog increases. The *C4* strategy, though its accuracy decreases when the number of alternatives increases, could still maintain a considerably higher relative accuracy than that of

Figure 3. The relative accuracy of various decision strategies when solving MADPs with a different number of alternatives, where n*(number of attributes)* = 10



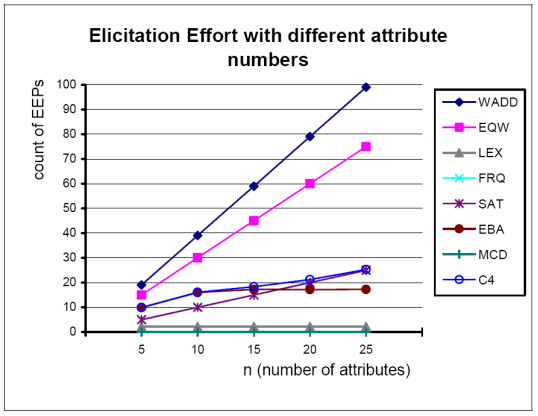
the EBA, MCD, LEX, and FRQ strategies when using the accuracy definition of RA_{HR} and $RA_{HR\_in\_Kbest}$.

The effect of the number of attributes on elicitation effort for each strategy is shown in Figure 4. As we can see, the elicitation effort of the heuristic strategies increases much slower than that of the WADD strategy as the number of attributes increases. For instance, when the number of attributes is *20*, the elicitation effort of the FRQ strategy is only about *25%* of that of WADD strategy. The FRQ and SAT strategies require the same level of elicitation effort, since both of them require the decision maker to input a cutoff value for each attribute. Except the MCD strategy, which requires no elicitation effort in the simulation environment, the LEX strategy is the one that requires the least elicitation effort in all cases among the listed strategies. The combined C4 strategy, which could share preferences among its four underlying basic strategies, requires only a slightly higher elicitation effort than the EBA strategy.

Figure 5 shows the relationship between *elicitation effort* and *the number of alternatives* for each strategy. As the number of alternatives increases exponentially, the level of elicitation effort for WADD, EQW, MCD, SAT, and FRQ strategies remains unchanged. This shows that the elicitation effort of these strategies is unrelated to the number of alternatives that a decision problem may have. For the LEX, EBA, and *C4* strategies, the elicitation effort increases slowly as the number of alternatives increases. As a whole, Figure 5 shows that the elicitation effort of the studied decision strategies is quite robust to the number of alternatives that a decision problem has.

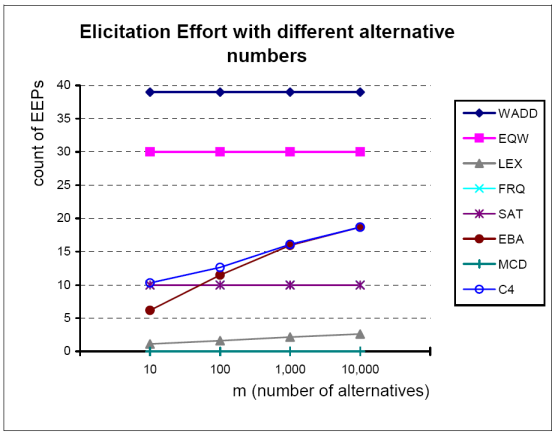
A combined study from Figure 2 to Figure 5 can lead to some interesting conclusions.

Figure 4. The elicitation effort of various decision strategies when solving MADPs with a different number of attributes, where m*(number of alternatives)* = 1,000



For each category of MADPs, some decision strategies, such as WADD and EQW, could gain relatively high-decision accuracy with proportionally high-elicitation effort. Other decision strategies, especially *C4*, MCD, EBA, FRQ, and LEX, could achieve a reasonable level of accuracy with much lower elicitation effort compared to the baseline WADD strategy. Figure 6 illustrates the relationship between elicitation effort and $RA_{HR\_in\_Kbest}$ for various strategies when solving different scales of decision problems. For the MADPs with 5 attributes and 100 alternatives, the MCD strategy could achieve around *35%* relative accuracy without any elicitation effort. The *C4* strategy, in particular, could achieve over *70%* relative accuracy while only requiring about 45% elicitation effort compared to the WADD strategy.

Figure 5. The elicitation effort of various decision strategies when solving MADPs with a different number of alternatives, where n*(number of attributes)* = 10



For all the decision strategies we have studied here, we say that a decision strategy *S* is *dominated* if and only if there is another strategy *S'* that has higher relative accuracy and lower cognitive and elicitation effort than *S* in the decision problem. Figure 6 shows that when the MADPs have 10 attributes and 1,000 alternatives, the WADD, EQW, *C4*, and MCD are nondominated approaches. However, for a smaller scale of MADPs (5 attributes and 100 alternatives), only the WADD, *C4*, and MCD strategies have the possibility of being the optimal strategy. This figure also shows that if the user's goal is to make decisions as accurately as possible, WADD is the best strategy among the listed strategies; while if the decision maker's goal is to have reasonable accuracy with a certain elicitation effort, then the *C4* strategy may be the best option.

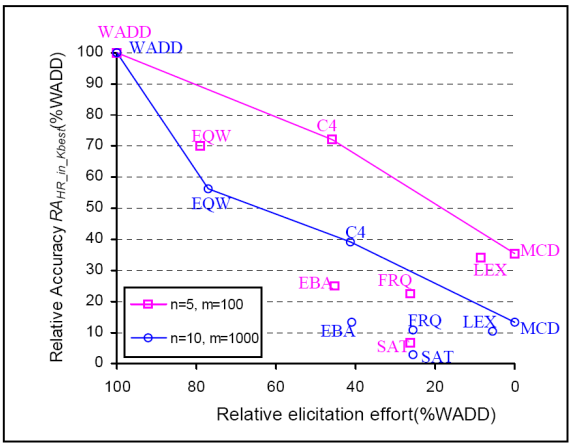
DISCUSSION

The simulation results suggest that the tradeoff between decision accuracy and elicitation effort is the most important design consideration for inventing high-performance CDSSs. That is, while advanced tools are desirable, we must not ignore the effort that users are required to make when stating their preferences.

To show how this framework can provide insights to improve user interfaces for the existing CDSSs, we have demonstrated the evaluation of the simplest decision strategies: WADD, EQW, LEX, EBA, FRQ, MCD, and SAT (Payne et al., 1993). The performance of these strategies was measured quantitatively in the proposed simulation environment within the extended effort-accuracy framework. Since the underlying decision strategy determines how a user interacts with a CDSS system (preference elicitation and result processing), the performance data allowed us to discover better decision strategies and eliminate suboptimal ones. In this sense, our work provides a new design method for developing user interfaces for consumer decision support systems.

For example, LEX is the underlying decision strategy used in the ranked list interface that many e-commerce Web sites employ (Pu & Kumar, 2004). However, our simulation results show that LEX produces relatively low-decision accuracy, especially as products become more complex. On the other hand, a hybrid decision strategy, *C4*, based on any combinations of LEX, EBA, MCD, and FRQ was found to be more effective. Combining LEX and EBA together, for example, we can derive an interface that looks like SmartClient. EBA (elimination by

Figure 6. Elicitation effort/relative accuracy trade-offs of various decision strategies



aspect) corresponds to eliciting constraints from users, and this feature was implemented as a constraint problem-solving engine in SmartClient (Torrens, Faltings, & Pu, 2002). After users have eliminated the product space by preference constraints, they can use the LEX strategy (ranked list) to further examine the remaining items. Even though this hybrid strategy does not include any interface features to perform trade-off navigation, the simulation results are still consistent with our earlier empirical work on evaluating CDSSs with real users (Pu & Chen, 2005; Pu & Kumar, 2004). That is, advanced tools such as SmartClient can achieve a higher accuracy while requiring users to expend slightly extra cognitive and elicitation effort than the basic strategies it contains.

The strongest implication of the simulation results is that we will be able to efficiently evaluate more-advanced tools and compare them in terms of effort and accuracy. Our plan for the future therefore includes evaluating SmartClient (Pu & Faltings, 2000) and its trade-off feature, FindMe (Burke et al., 1997), Dynamic Critiquing (Reilly et al., 2004), and Scoring Trees (Stolze, 2000), and comparing their strengths and weaknesses. We will also perform more simulations using different scales of K with the accuracy definition of $RA_{HR\_in\_Kbest}$. Because K is the size of the result set displayed in each user interaction cycle, we hope to gain more understanding on the display strategy used for CDSSs.

Finally, we do emphasize that the simulation results need to be interpreted with some caution. First of all, the elicitation effort is measured by approximation. As mentioned earlier, we assumed that each EEP requires an equal amount of effort from the users. Currently, it is unknown whether this approximation would affect the simulation results largely. In addition, when measuring the decision accuracy, the WADD strategy is chosen as the baseline, assuming that it contains no error. However, this is not the case in reality. Moreover, as the MADPs in the simulation experiments are generated randomly, there is a potential gap between the simulated MADPs and the product catalog in real applications. We are currently addressing these limitations, and fine-tune some of the assumptions with real user behaviors.

CONCLUSION

The acceptance of an e-commerce site by consumers strongly depends on the quality of the tools it provides to help consumers reach a decision that makes them confident enough to purchase. Evaluation of these consumer decision support tools on real users has made it difficult to compare their characteristics in a controlled environment, thus slowing down the optimization process of the interface design of such tools. In this paper, we described a simulation environment to evaluate the performance of CDSSs more efficiently. In this environment, we can simulate the underlying decision support approach of the system based on the consumers' preferences and the product catalog information that the system may have. The decision results can then be evaluated quantitatively in terms of decision accuracy, elicitation effort, and cognitive effort described by the extended effort-accuracy framework. More importantly, we were able to discover new decision strategies that led to interface improvements of existing CDSSs. Even though this is the first step, we hope to be able to evaluate and design new user interfaces for high performance CDSSs, and forecast users' acceptance of a new interface based on benefits such as effort and accuracy.

ACKNOWLEDGMENTS

Funding for this research was provided by the Swiss National Science Foundation under grant 200020-103490.

REFERENCES

Boutilier, C., Patrascu, R., Poupart, P., & Schuurmans, D. (2005). Regret-based utility elicitation in constraint-based decision problems. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI'05)*, Edinburgh, Scotland, UK (pp.

929–934).

Burke, R. D., Hammond, K. J., & Young, B. C. (1997). The FindMe approach to assisted browsing. *IEEE Expert, 12*(4), 32–40.

Grether, D. M., & Wilde, L. L. (1983). Consumer choice and information: New experimental evidence. *Information Economics and Policy, 1*, 115-144.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value tradeoffs*. Cambridge, UK: Cambridge University Press.

Luce, M. F., Payne, J. W., & Bettman, J. R. (1999). Emotional trade-off difficulty and choice. *Journal of Marketing Research, 36*(2), 143–159.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Palmer, J. W. (1997). Retailing on the WWW: The use of electronic product catalogs. *International Journal of Electronic Markets, 7*(3), 6–9.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, UK: Cambridge University Press.

Pu, P., & Chen, L. (2005). Integrating tradeoff support in product search tools for e-commerce sits. In *Proceedings of the ACM Conference on Electronic Commerce (EC'05)*, Vancouver, Canada (pp. 269–278).

Pu, P., & Faltings, B. (2000). Enriching buyers' experiences: The SmartClient approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, The Hague, The Netherlands (pp. 289–296).

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

Reilly, J., McCarthy, K., McGinty, L., & Smyth, B. (2004). Dynamic critiquing. In *Proceedings of the 7th European Conference in Case-Based Reasoning (ECCBR 2004)*, Madrid, Spain (pp. 763–777).

Reilly, J., McCarthy, K., McGinty, L., & Smyth, B. (2005). Incremental critiquing. *Knowledge Based Systems, 18*(2-3), 143–151.

Shearin, S., & Lieberman, H. (2001). Intelligent profiling by example. In *Proceedings of the International Conferences on Intelligent User Interfaces (IUI'01)*, Santa Fe, New Mexico (pp. 145–151).

Shimazu, H. (2001). Expertclerk: Navigating shoppers buying process with the combination of asking and proposing. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI'01)*, Seattle, Washington (pp. 1443–1448).

Stolze, M. (1999). Comparative study of analytical product selection support mechanisms. In *Proceedings of INTERACT 99*, Edinburgh, UK (pp. 45–53).

Stolze, M. (2000). Soft navigation in electronic product catalogs. *International Journal on Digital Libraries, 3*(1), 60–66.

Torrens, M. (2002). *Scalable intelligent electronic catalogs*. Doctoral dissertation no. 2690, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.

Torrens, M., Faltings, B., & Pu, P. (2002). Smart clients: Constraint satisfaction as a paradigm for scaleable intelligent information systems. *CONSTRAINTS: An International Journal, 7*(1), 49–69.

Viappiani, P., Faltings, B., Schickel-Zuber, V., & Pu, P. (2005). Stimulating preference expression using suggestions. In *Workshop Notes of the IJCAI-05 Multidisciplinary Workshop on Advances in Preference Handling*, Edinburgh, UK (pp. 186–191).

Zhang, J., & Pu, P. (2005). Effort and accuracy analysis of choice strategies for electronic product catalogs. In *Proceedings of the 20th ACM Symposium on Applied Computing (SAC-2005)*, Santa Fe, New Mexico (pp. 808–814).

ENDNOTES

[1] Search tools and consumer decision support systems are two terms used interchangeably throughout this article although the latter is considered to be more advanced in terms of its implementation and interface features.

[2] It is also known as multicriteria decision making (MCDM) problem (Keeney & Raiffa, 1993). Our definition emphasizes on the term *attribute*, which is an objective aspect of products, not related to the decision maker's preferences.

[3] Though this assumption is obviously imprecise, more studies by assigning different weighting of effort for the various EIPs show that the key relationships between the decision strategy and the decision environments were largely unchanged. See page 137 of (Payne et al., 1993).

[4] We assume that the actions are in their basic forms only. For example, the FILLIN operation is not allowed to elicit more than one value or even an expression. Otherwise a usability issue will arise.

[5] The detail analysis is given at pages 80–81 of Payne et al. (1993). This example assumes the values of all attributes are numeric and consistent with the decision maker's preferences.

[6] The procedure of assessing component value functions with midvalue points is introduced in page 120 of Keeney and Raiffa (1993).

Mr. Jiyong Zhang (jiyon.zhang@epfl.ch) obtained both his BS (1999) and MS (2001) in computer science from Tsinghua University, Beijing, China. Currently, he is a PhD student and research assistant in the Human Computer Interaction Group, School of Computer and Communication Sciences, Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland. His research interests include intelligent user interfaces, automatic decision making, constraint based problem solving, preferences modeling and handling, and e-commerce technologies.

Dr. Pearl Pu (pearl.pu@epfl.ch) is currently a research scientist and director of the HCI Group in the School of Computer and Communication Sciences at the Swiss Federal Institute of Technology in Lausanne (EPFL). She obtained her master's and PhD degrees from the University of Pennsylvania in artificial intelligence and computer graphics. She was a visiting scholar at Stanford University in 2001, both in the database and HCI groups. She was also co-founder of Iconomic Systems (1997-2001), and invented the any-criteria search method for finding configurable and multi-attribute products in heterogeneous electronic catalogs. Her recent research activities are in the areas of decision support systems, information visualization, query optimization for digital libraries, scalable user experience, social navigation, and advanced display techniques.