User Model User-Adap Inter (2012) 22:357–397
DOI 10.1007/s11257-011-9116-6

ORIGINAL PAPER

Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process

Alina Pommeranz · Joost Broekens · Pascal Wiggers · Willem-Paul Brinkman · Catholijn M. Jonker

Received: 14 December 2010 / Accepted in revised form: 7 November 2011 /
Published online: 15 March 2012

Abstract Two problems may arise when an intelligent (recommender) system elicits users' preferences. First, there may be a mismatch between the quantitative preference representations in most preference models and the users' mental preference models. Giving exact numbers, e.g., such as "I like 30 days of vacation 2.5 times better than 28 days" is difficult for people. Second, the elicitation process can greatly influence the acquired model (e.g., people may prefer different options based on whether a choice is represented as a loss or gain). We explored these issues in three studies. In the first experiment we presented users with different preference elicitation methods and found that cognitively less demanding methods were perceived low in effort and high in liking. However, for methods enabling users to be more expressive, the perceived effort was not an indicator of how much the methods were liked. We thus hypothesized that users are willing to spend more effort if the feedback mechanism enables them to be more expressive. We examined this hypothesis in two follow-up studies. In the second experiment, we explored the trade-off between giving detailed preference feedback and effort. We found that familiarity with and opinion about an item are important factors mediating this trade-off. Additionally, affective feedback was preferred over a finer grained one-dimensional rating scale for giving additional detail. In the third study, we explored the influence of the interface on the elicitation process in a participatory set-up. People considered it helpful to be able to explore the link between their interests, preferences and the desirability of outcomes. We also confirmed that people do not want to spend additional effort in cases where it seemed unnecessary. Based on the findings, we propose four design guidelines to foster interface design of preference elicitation from a user view.

Keywords Preference elicitation · Constructive preferences · Interface design

A. Pommeranz (✉) · J. Broekens · P. Wiggers · W.-P. Brinkman · C. M. Jonker
Department of Mediamatics, MMI Group, Delft University of Technology,
Mekelweg 4, 2628 CD, Delft, The Netherlands
e-mail: a.pommeranz@tudelft.nl; alina.pommeranz@tudelft.nl

1 Introduction

Web technology and computational intelligence enable the development of systems that assist users in tasks that are cognitively demanding. These smart systems are becoming essential tools for people to deal with information overload, huge search spaces and complex choice sets in different domains, such as product or service recommendations (Adomavicius and Tuzhilin 2005) or decision support e.g. in health care, real estate, jobs or divorce negotiations (Johnson et al. 2005; Bellucini and Zeleznikow 2006). Whereas a substantial amount of research in the field of recommender and decision support systems focuses on recommendation algorithms, formal representations and reasoning mechanisms, little research addresses the design of the user interfaces of these systems. Recently, see e.g., the work of Knijnenburg et al. (2012) in this issue, the importance of the user interface design and its effects on the user experience of recommender systems has been emphasized. The interface between the user and the system plays a major role in acceptance of the systems as well as the user's trust and satisfaction (Pu and Chen 2007; Pu et al. 2012). In particular, the method and the interface designed to elicit user preferences influences decision accuracy and the intention to return (Chen and Pu 2009).

Smart systems need accurate preference models to be able to give useful advice to the user. A preference elicitation interface needs to extract information from the user's mental representation of that preference and translate it into a representation the system can reason with. Preference modeling thus always involves three components: mental representation, elicitation and the system's preference representation. In this article, we focus on the middle part: the preference elicitation interface. Two key issues in preference elicitation are (1) a potential mismatch between the user's mental model of his or her preferences and the system's preference representation and (2) the influence of the elicitation process on the created preference profile. The first issue is a result of the discrepancy between the rational, quantitative models used in systems and the constructive, qualitative mental models of people. Whereas rational models assume that people have stable and coherent preferences that are always known to them, people rather construct their preferences as they go along in the decision task. In addition, it is difficult for people to express their preferences in numerical attribute weights and values as needed by automated systems, particularly if they are not experts in the domain. Whereas people might easily state that they prefer, e.g., more holidays to less holidays specifying this relation in concrete numbers, such as "I like 30 holidays 2.5 times as much as 28 holidays" is not intuitive.

The nature of how humans construct their preferences leads to the second issue, namely that the method or process employed to extract preference information from the user influences the preferences the user constructs. Preference construction can be influenced by the decision context, the framing of the choice task and the way relevant information is presented. This has to do with psychological effects (Fischer et al. 1999; Johnson et al. 2005), e.g., loss aversion (the tendency of people to prefer avoiding losses to acquiring gains) or anchoring effects (relying too heavily on one piece of information), the emotions induced or earlier experiences retrieved from memory when the elicitation question is posed (Weber and Johnson 2006). The fact that the process of eliciting preferences (e.g., giving information about choices and asking

a number of valuation questions) influences the construction of preferences should be taken into account when designing a preference elicitation interface by actively supporting this process. Active involvement in the construction process and a user interface design leading to a positive user experience (Knijnenburg et al. 2012) during the elicitation as well as an understanding and trust in the system's output later on (Carenini and Poole 2002) is important for the success of recommender systems.

Our work focuses on informing the design of preference elicitation interfaces from a user-centered point of view. In this paper we present three studies that explore how we can bridge the gap between users' mental models and a system's representation of the preferences and how the constructive nature of human preferences can be supported in an interface. We combined experimental as well as qualitative research involving users in the design process to be able to create a number of design guidelines for such interfaces.

In the first experiment we investigated input methods and elicitation process in a structured way. We presented users with different ways of entering preferences, including ratings (Likert scale rating), affective feedback, and sorting, both on an item (i.e., a complete holiday) as well as attribute (i.e., beach, mountain, active, etc.) basis. Based on the results of this experiment we hypothesized that users are willing to spend more effort if the feedback mechanism (i.e., process and preference representation) enables them to be more expressive (e.g., by giving more dimensional feedback or navigating through the outcome space). We examined this hypothesis in two follow-up studies. In the second experiment we explored the trade-off between giving detailed preference feedback and effort. We investigated factors, such as content type, familiarity, ownership and directed opinion (positive or negative), that may influence this tradeoff in an experimental setup. In a third study we explored how people prefer the preference elicitation process to be structured using hi-fi interface prototypes and a participatory design method. We looked at four fundamentally different processes of eliciting preferences based on different ways to process information. We used the mind style theory by Gregorc (2006) which categorizes people based on perceptual and ordering preference. Perceiving information can be abstract (based reason and intuition) or concrete (using one's senses). The order of information processing can be sequential or random. Thus there are four types to process information: concrete sequential, concrete random, abstract sequential and abstract random. We built one interface prototype per style and evaluated the prototypes in individual user sessions followed by a creative design session with all users. Based on the results of all three experiments we constructed a number of design guidelines to support further development of preference elicitation interfaces.

The experiments will be discussed in sects. 3 to 5. Section 6 discusses the results and presents the guidelines and Sect. 7 concludes the article. But first, we will give an extensive background on how people construct their preference, how current systems elicit them, and how well the theory matches to the practice.

2 Background

People's preferences have been the interest of researchers in many different fields including psychology, (behavioral) decision making, consumer research, e-commerce,

intelligent systems as well as negotiation and decision support. We focus on topics relevant for designing user interfaces for preference elicitation for intelligent systems. In the following sections we give insights into (1) how people construct their preferences (the process we need to support with preference elicitation interfaces), (2) the state-of-the-art in preference elicitation interfaces and (3) how the latter take the human preference construction into account.

2.1 Constructive preferences

A dominant model in contemporary economy is that of the rational consumer trying to always maximize his outcome. Preferences are seen as primitive, consistent and stable (McFadden 1999). It also assumes that people know their preferences. Since the rational consumer tries to maximize the value outcome it is implied that he is able to compute the maximal outcome based on his preferences and make a rational choice. This computation can be represented in utility functions—a mathematical representation of a person's preferences.

Whereas these assumptions serve rational economic theories well, they are not always true for human behavior. More and more researchers gathered proof supporting a constructive view of human preferences. This view implies that people construct their expressions of preferences at the time the valuation question is asked. Furthermore, the decision process itself and the context play a major role in the construction process (Payne et al. 1999).

There are different views on how people construct their preferences. Simon et al. (2004) for instance found in their experiments that while people processed the decision task, their preferences of attributes in the option that was chosen increased and those for attributes of rejected options decreased. This is in line with achieving the meta goal of trying to maximize the ease of justifying a decision (Bettman et al. 1998). Similar effects have been found in negotiation settings reported by Curhan et al. (2004).

Fischer et al. (1999) focused on the goals of the decision task in relation to a prominence effect. This effect occurs when people prefer an alternative that is superior only on the most prominent, i.e., the most important, attribute. They confirmed in three studies that the prominent attribute will be more heavily weighted when the goal was making a choice between alternatives than when the goal was to arrive at a matching value.

Weber and Johnson (2006) state that people construct preferences from memory. The so-called PAM (preferences-as-memory) framework assumes that "decisions (or valuation judgments) are made by retrieving relevant knowledge (attitudes, attributes, previous preferences, episodes, or events) from memory in order to determine the best (or a good) action." Weber and Johnson emphasize that this is not an entirely cognitive view on preference construction since affect determines what the person recalls first. Information consistent with emotions is more available in memory. Johnson et al. (2005) found psychological effects, such as anchoring effects and effects occurring when complicated numbers or information are presented in the choice task. In their experiments different ways to measure preferences led to different results. To help people to construct their preferences in health care scenarios, Johnson et al.

suggest to present default choices that lead to the best outcome for most patients and present information in a way that helps the patient to understand the outcomes of each choice. Consumer research looked at the interplay between affect and cognition on decision making (Shiv and Fedorikhin 1999). In cases where people have only few cognitive resources available affective reactions tend to have a greater impact on choice, whereas with high availability of cognitive resources thoughts related to the consequences of the choice are more dominant. This finding can be influenced by personality and by the representation of the choice alternatives.

In summary, there is consensus in the literature discussed that people do not always have stable and consistent preferences but rather construct them when necessary. There are numerous views on how people might construct their preferences. An easy, ready-to-implement recipe for designing interfaces for this task has not yet been established. There have been few attempts to guide system developers in this difficult task. Carenini and Poole (2002) point to the problems of clustering and matching algorithms in relation to the constructive process humans go through. Since users may not have the chance to construct their preferences they might also not be able to understand the system's output. Kramer (2007) also found that consumers are more likely to choose a recommendation that matches their measured preferences when it is easy to see through the preference elicitation method and by that identify their expressed preferences. The gap between the user's mental model and the system's preference profile of the user can be bridged with explanations from the system (Carenini and Poole 2002). More research is needed to design preference elicitation interfaces that elicit correct preference information from the user. In the following sections we will give an overview over current preference elicitation methods used in state-of-the-art interfaces.

2.2 Preference elicitation methods and interfaces

Methods for acquiring user preferences range from implicit to explicit ones depending on the nature of the system. By implicit we refer to approaches in which the user is not "actively" involved in the elicitation task as it is the case in explicit. One example of implicit methods can be preference learning based on user behaviour (e.g., items the user looked at or bought). Users may still be aware of the workings of implicit methods and expect an interpretation of their actions. By explicit we, however, refer to methods that require specific preference input such as ratings. The range from implicit to explicit is continuous, meaning that various degrees of user involvement are used in the methods explained below. In the following sections we will describe the methods typically used in recommender systems and decision support together with examples representing typical systems in the area. For more exhaustive reviews in the area see (Chen and Pu 2004; Peintner et al. 2008).

2.2.1 Methods used in recommender systems

Recommender systems (Adomavicius and Tuzhilin 2005) are tools that provide personalized recommendations to people. They are integrated either in shopping websites, e.g., amazon.com or dedicated recommendation websites (Resnick et al. 1994;

Burke 2000; Miller et al. 2003; Stolze and Ströbel 2003). The interaction models employed to acquire preferences vary from implicit to explicit methods. Typically, preference elicitation is done through item-rating (and using filtering methods) or more conversational interaction using tweaks or critiques.

Rating-based recommender systems collect a number of initial ratings from a user and then try to estimate ratings for the yet unrated items. Based on a user's profile and estimations for unseen items they can recommend new products to the user that he or she could be interested in. Two methods are mainly used, collaborative filtering (Herlocker et al. 2004) based on similarities between users (as reflected by their ratings) and the content-based method (Pazzani and Billsus 2007) based on item attributes instead. Most content-based recommenders employ machine-learning techniques to create a user profile. To recommend items the attributes of the item are compared to the user's profile to see which items would be of interest to the user. Systems using collaborative filtering are, for instance, MovieLens (Miller et al. 2003), or GroupLens (Resnick et al. 1994) (www.grouplens.org), but also commercial systems like Amazon.com; an example of a system using a content-based method is the book recommender system developed by Mooney and Roy (2000).

Recommender systems using collaborative filtering or the content-based method mostly focus on getting a substantial number of ratings from their users when they sign up. During use the explicit interaction between system and user is limited to recommendations from the system and voluntary ratings from the user. Carenini et al. (2003) have instead proposed a more conversational and collaborative interaction. Key in the proposed interaction is that the system tries to elicit ratings or preferences when people are particularly motivated to give them, e.g., when the system cannot give a requested recommendation due to a lack of preference information or when the given rating puzzles the user. Methods developed based on this conversational model are 'Example Similarity and Tweaking' and 'Example-Critiquing Interaction'.

Most so-called FindMe systems [e.g. Car Navigator, PickAFlick, RentMe or Entrée (Burke et al. 1996; Burke 2000, 2002)] employ example similarity and tweaking techniques. In the first step the user selects an item from the system's catalogue and requests similar items. The system then retrieves a large number of alternative items from its database, sorts them according to similarity to the chosen item and returns a small number of alternatives with highest similarity to the user. In case the system offers tweaking the process is essentially the same with the only difference that the user gives a tweak in the first step, e.g., "show me similar, but cheaper items". The system only returns items to the user that satisfy the tweak, that are cheaper in this case. A similar technique is example-critiquing or the candidate/critique model (Pu and Chen 2008). In example-critiquing users are presented with a set of candidates they can critique. Candidates have to motivate users to state their preferences and the most preferred solution needs to be among the displayed candidates (Faltings et al. 2004). Several strategies have been proposed to select candidates, e.g., using extreme examples (Linden et al. 1997), diverse examples (Smyth and Mcginty 2003) or so called Pareto-strategies (Viappiani et al. 2005). In an iterative process the system learns the users' preferences from their critiques and updates the user models. Critiques can be system-suggested or user-initiated. Work to enhance the critiquing has been done in the area of dynamic critiquing, in which compound critiques

(critiques operating over multiple features) are generated on-the-fly (McCarthy et al. 2005). One of the first interfaces using example critiquing is the APT Decision Agent (Shearin and Lieberman 2001). For details of other systems and algorithms in this area and specific design guidelines on how to develop example critiquing interaction (see Chen and Pu (2009)).

2.2.2 Decision support systems

Decision support systems (DSS) are interactive systems that support users in taking decisions by eliciting preferences and offering analytical tools to scrutinize decisions. Unlike recommender systems that focus on finding the best outcome in a huge set of possible outcomes, decision support systems focus much more on the process of taking a decision and the role of preferences influencing that process.

Preferences are elicited explicitly because it is important that the user understands the relation between his or her preferences and the possible outcomes of the decision making process. Decisions that are supported by these systems are often of much higher importance than choosing to buy a book or see a movie, e.g., medical systems (Hunt et al. 1998; Johnson et al. 2005). It is, therefore, important to have a precise model of the users preferences.

The majority of decision support systems are based on multi-attribute utility theory (MAUT) (Keeney and Raiffa 1993) and, therefore, represent preferences in form of utility functions. In order to construct utility functions the system needs to elicit values and weights for the given attributes of an item. The two most popular preference elicitation techniques are *absolute measurement* and *pairwise comparison* (active elicitation) (Aloysius et al. 2006). Absolute measurement (e.g. salary scores 9 on a scale 1–10 of importance, whereas number of holidays score 5 out of 10) does not require the user to make explicit trade-off judgments. Pairwise comparison (e.g., salary is more important than number of holidays) explicitly asks for a trade-off between attributes. Weights in most existing systems are entered by users on discrete scales by selecting a rating from a drop down list or using horizontally aligned radiobuttons and on continuous scales by using a slider. Aloysius and colleagues (Aloysius et al. 2006) found an impact of the preference elicitation technique used on the user acceptance of DSS. Their study comparing absolute measurement and pairwise comparison showed that forcing the user to make explicit trade-off judgments has a negative effect on user acceptance of the system. Note, that this does not mean that the decision outcome will be worse. However, due to higher perceived effort and decisional conflict the user perceives the accuracy of the system to be lower.

2.2.3 Configuration systems

Similar to recommender or decision support systems are configuration systems, which support the configuration of complex products and services. A growing demand for customer individual, configurable products also asks for improvements of configuration systems that usually have to deal with a wide variety of users. In this area Ardissono et al. (2003) have developed the CAWICOMS workbench to develop configuration services. Interesting with respect to preference elicitation is the way this

workbench manages user models and personalizes the interaction between user and system by customizing the acquisition of requirements and information presentation. The system exploits user classes based on stereotypes that specify skills as well as interests. In the beginning of the interaction the system asks explicitly about background information of the user. This helps to define the user class and make estimates about the user's interests and skills based on the stereotypes. During the interaction the system observes the actions of the user to update the skills and interests continuously as they may evolve. The reasoning of the system is based on the rational assumption that the user always tries to maximize her own utility by setting item features to satisfy her needs. Depending on the system's assumptions it selects certain features as critical and others as less important which are presented as supplementary information.

2.3 Support of human preference construction in current methods

In order to build a system and in particular the interface of the system that is usable and supports the user in creating and entering his or her preferences system designers need to consider how human preferences develop. In detail, this means they have to support users in constructing their preferences in a cognitive as well as affective way and maybe look at underlying interests (or values) as a basis for selecting the right attributes. But in how far do the methods presented above actually take these aspects into consideration?

Generally, the implicit methods do not actively involve users in constructing their preferences. Collaborative filtering methods base their choices on the assumptions that similar people like similar things. However, there is always the danger that the system creates an erroneous user model and the user gets confused about seemingly unrelated recommendations.

Explicit methods focus more on the user. Conversational methods allow the users to construct and reflect upon preferences. Example critiquing has been explicitly developed based on the constructive view of human preferences. This is reflected in the attempts of improving the algorithms to pick examples that will help the users to uncover hidden preferences. Whereas tweaking and example-critiquing are widely used for recommenders, many decision support systems use active elicitation methods based on utility models. This requires users to enter values and weights in form of numbers. Other active elicitation methods like pairwise comparison do not require numbers but still assume that a user is able to compute which of the given options is better. With unknown items this is a difficult task.

A combination of explicit questions, in particular in the beginning of the interaction, to place the user in a certain class and continuous explicit updates of the preference model based on user behavior (Ardissono et al. 2003), can help to create an accurate preference model. Continuous updates of the model and adaptations of the interface support the constructive nature of human preferences as they may evolve during the interaction with the system.

Surprisingly few systems explicitly try to elicit underlying interests before deciding which attributes or items are worth looking at for preference elicitation, notable

exceptions are (Fano and Kurth 2003; Stolze and Ströbel 2003). Affect is also underexplored in current preference elicitation methods. Only the movie recommenders by Ono et al. (2007) and whattorent.com ask for input about emotions or moods. However, underlying interests and affect are important aspects of human preference construction that should be considered in a preference elicitation interface.

In summary, we can say that there are a few methods that are explicitly based on the constructive view of human preferences. There is much room for more explicit consideration of human preference construction also including values and affective aspects.

2.4 Related work

Whereas most work described in the previous subsections focuses on different techniques in which a system can either explicitly or implicitly arrive at a preference model, our work focuses strongly on the design of preference elicitation interfaces, in particular for explicit preference elicitation. Naturally, the interface design is also determined by the interaction style or technique that is chosen. However, besides that, we believe that there is much room for improvement and greater support of the way in which humans construct their preferences. This constructive view has been acknowledged by others who also established a number of guidelines for preference elicitation (Payne et al. 1999; Pu and Chen 2008; Pu et al. 2003, 2012). Our work is intended to build on this work by extending the number of guidelines in order to help other designers. However, our work differs as it focuses on a number of issues neglected in the current literature. These include: (1) an in-depth investigation of interface elements considered appropriate and preferred by people for entering preferences, (2) the match of the acquired input with input required by algorithms currently used to create a system representation of the preferences, (3) intrinsic motivational factors that lead people to spend more effort to give more detail about a preference and (4) ways to structure the process of preference construction by the design of an interface based on different information processing styles. Investigating these aspects required close interaction with target users. In our view, when designing new preference elicitation interfaces a participatory design process will lead to a greater understanding of interface aspects that would not be acquired with user evaluations of finished prototypes. The majority of the work presented above does not follow such an approach. Most closely related to our work is the work by Barneveld and Setten (2004), who also involved users actively in the design phase of a TV recommender system with the means of brainstorming and interactive design sessions. Furthermore, they also investigated which interface widgets would be preferred by users to give preference input. Our work differs from theirs as we looked more at different types of input (rating, ordering, affective, navigation in the first study and higher level elements, e.g., a chat, in the third). In addition, whereas they focused on a design for one particular domain (TV) our work aims at a general understanding of how to design preference elicitation for different types of systems and different user groups. In the following sections we will describe our studies in detail.

3 Study 1: Investigating different ways of entering preferences

In this experiment we compared different ways of giving preference input (ranking, ordering, navigational) regarding perceived liking and effort and how well the extracted information serves as input for an outcome ranking algorithm. We agree with conclusions of Knijnenburg et al. (2012), that the algorithms cannot be studied in isolation with end-users, but have to be investigated together with the preference input to fully understand the complete user experience. Therefore, these two aspects are combined in this study. Liking is only one aspect of the user experience. We decided to investigate effort as most preference elicitation tasks require some level of effort from the users even before they can actually judge the usefulness and efficiency of a system (e.g., the accuracy of recommendations provided after the initial preference elicitation), see also Pu et al. (2012).

In this study, we considered ordering and rating tasks on both a property and outcome level. To investigate the effect of affective input we compared standard Likert-scale rating to affective ratings using the AffectButton (Broekens and Brinkman 2009). This button (Fig. 1) enables users to enter dynamic (i.e., graded) emotions. It renders a face that changes directly according to the mouse position and scroll wheel. The mouse-coordinates inside the button and the scroll wheel together define the values on the affective dimensions Pleasure, Dominance and Arousal (PAD) (Mehrabian 1980) respectively. All three dimensions are represented by values on a scale from −1 to 1 (e.g., −1 displeasure to 1 pleasure and accordingly). The pleasure dimension indicates how pleasurable an emotion is, e.g., fear or anger are emotions that are not pleasant whereas joy or contentness are pleasant. Dominance indicates the nature of the emotion ranging from submissive (e.g., in fear) to dominant (e.g., in anger). The arousal dimension indicates the intensity of an emotion ranging from low to high. Whereas joy has a high intensity, contentness has a low intensity. By using the AffectButton the users select an affective triplet from the PAD space (as reflected by the emotional expression of the button itself; the PAD concept is not visible to the user).

Furthermore, we compared a *navigational input method*, inspired by guidelines proposed by Pu et al. (2003) (i.e., any preference in any order and immediate visual feedback) to traditional ordering of properties. In the navigational input method users navigate through the outcome space by changing any one property at a time and receiving visual feedback for the new choice.

To see how the interaction between the input method and the system's computation influences the end result (ranked list of items) we used the preferences over properties obtained from different methods as input for the lexicographic ordering. The lexicographic ordering was chosen as it does not require numerical input from the

Fig. 1 Example expressions: from left to right Happy (PAD = 1, 1, 1), Afraid (−1, 1, −1), Surprised (1, 1, −1), Sad (PAD = − 1, −1, −1), Angry (−1, 1, 1)

Table 1 Overview of 8 preference elicitation tasks

Task	Description
1A	Order 9 property values (given at the same time)
1B	Order 27 holidays
2A	Navigation through holidays
2B	Order 3x3 property values (given three at a time)
3A	Likert rating of holidays
3B	Affective rating of holidays
3C	Likert rating of properties
3D	Affective rating of properties

participants (properties need to be ordered, but are not associated with a numerical weight) and by that allowed us to use ordering tasks in the experiment. It has also been argued that it is a natural and intuitive way to derive preferences over objects from an importance ranking of properties (Liu 2008). This type of ordering compares two items according to the property that is rated most important. Other properties will only be considered if the value of the most important property is the same for both objects. So given a user prefers having a garage to a garden with his house, then an option A that has a garage is always better than an option B without a garage, even in cases where option B has many other attributes that the user also likes but finds less important than a garage. If option A and B contain a garage the algorithm will compare the options based on the next important attribute, e.g., whether they have a garden and so on.

3.1 Research questions

Overall, we addressed three topics: (a) different preference input methods (interface), (b) in specific the navigational input method and affective inputs and (c) the outcome ordering using a lexicographic algorithm with input from the property rating/ordering methods. In detail, we focused on the following research questions.

(1) *How do people perceive the different input methods in terms of liking and effort?*
(2) *Do users prefer the navigational input method to standard ordering and rating methods in terms of effort, intuitiveness, ease of use and liking? Can the navigational input method extract the same information as the property ordering method?*
(3) *Do users prefer to give affective feedback? How does the user perceive the quality of the resulting outcome orderings?*
(4) *How similar are outcome lists generated with the lexicographic ordering to a list created by the user (baseline)?*

3.2 Study setup

We ran an experiment consisting of 8 ordering/rating tasks (tasks will be numbered throughout the paper), 2 comparisons of results and a final questionnaire. An overview of the ordering/rating tasks is presented in Table 1 (each task will be discussed

Table 2 Properties of holidays and the alternative values for each property used in the experiments

Location	Accommodation	Type
Mediterranean	Apartment	Relaxation
Alps	Hotel	City trip
Scandinavia	Camping	Active

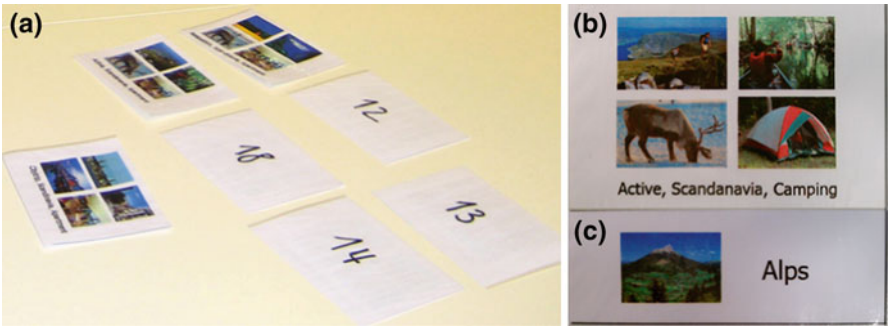

Fig. 2 **a** Navigational task: Card in the third row presents current holiday, the participant can look at two other holidays at a time. **b** Card representing a holiday. **c** Card representing one property

in more detail below). After execution of a task we asked participants to rate (on a 7-point Likert scale) how much effort the task cost and how much they liked the task. We chose holidays as our domain, since people can easily relate to holidays and have preferences about different aspects of holidays. Each holiday has the properties type, location and accommodation, with the respective alternative values relaxation, active and city trip, Mediterranean, Scandinavia and Alps, and hotel, camping and apartment (Table 2).

3.2.1 Material

The study material consisted of two sets of 9 cards, each showing one alternative value for a holiday property, one set with pictures (Fig. 2b) and the other without pictures. Further, there were two sets with 27 cards showing complete holidays; one set with 4 pictures to give an orientation about what the holiday could look like (Fig. 2c), and one set without pictures. Furthermore, we provided a computer interface for participants to rate either holidays or alternatives for properties of holidays one at a time. Rating was done using either a 9-point Likert scale from like to dislike or the AffectButton.

3.2.2 Participants

We tested 32 participants, 10 female and 22 male, which were mainly students and researchers within the field of information technology aged between 21 and 31. Each participant had to do all tasks the experiment consisted of. The order of the tasks was counterbalanced to avoid carry-on effects. However, as the property space is kept small we expect people to know their preferences for the holiday preferences from the start

or construct them easily. We do not see this as a problem as the focus of the study lies on different ways to enter a (possibly known) preference, not on constructing it.

3.2.3 Design

Effort and Liking of input methods After each input method we asked participants to fill in a short questionnaire rating how much they liked the method and how much effort it took them.

Standard input for lexicographic ordering and baseline In task 1A participants were asked to order all nine property values (see Table 2). This property ordering was later on used as input for the lexicographic ordering of holidays. Task 1B -ordering 27 cards showing complete holidays, each consisting of a combination of the three properties- was used as a baseline to compare holiday lists. Equally preferred holidays could be put on the same level. All cards had to be laid out on the table from most preferred to least preferred.

Navigational input method To test the effect of a navigation through the decision space, i.e., holidays to elicit preferences, two tasks were presented to the user. In the navigation task (2A), the participants were initially presented with a random card showing a complete holiday and asked to find their most preferred holiday by changing one property at a time to any of the two alternative values of that property, e.g., location could be changed from Scandinavia to either the Alps or the Mediterranean. As there were three properties (location, accommodation and type) that could be changed to two other values than the current, each holiday had six related holidays. The subjects could have a look at all six holidays (two at a time) related to the present one before deciding which one to navigate to (Fig. 2a). The task was presented as a paper prototype. Once the subjects found their most preferred holiday the procedure was repeated for the least preferred holiday starting with the most preferred one. The cards showed three property values of a holiday and four pictures, which were used to give the participant an idea about the kind of holiday.

In the second task (2B), the subject had to order the alternative values of each of the three holiday properties (see Table 2). Each property was presented on a card with a picture. Furthermore, the subject was asked to order the properties (type, location and accommodation) according to importance when searching for a holiday.

In addition to the effort and liking questionnaire, a questionnaire was presented to the user containing a number of questions about the intuitiveness and ease of use of the navigation (2A) and property ordering (2B) tasks.

Affective feedback To study the effect of affective rating methods we used a 2×2 experimental setup. We had four different conditions: (1) 9-point Likert rating of nine holidays (3A), (2) affective rating of the same holidays (3B), (3) 9-point Likert rating of all nine property values (3C) and (4) affective rating of all nine properties (3D). Holidays and properties were presented one by one and in random order. For each condition a simple algorithm generated an ordered list containing nine holidays based on the user input. In the first condition the list was ordered directly based on the user's holiday preference feedback. In the second condition feedback variables pleasure, arousal and dominance were summed and then used to order the list. In the third condition the weight of the property value entered by the user was used to calculate a sum

for each holiday. This sum was used to order the list of holidays. In the fourth condition the pleasure, arousal, and dominance feedback was summed and then used to order the property values; from this property ordering an ordering of the nine holidays was derived. These algorithms resulted in four differently sorted lists, each containing the same holidays. After the rating and ordering tasks, users were asked to compare the four lists to their own holiday ordering.

Preference ordering We used the information collected in tasks 1A, 2A, 3C and 3D (tasks based on holiday properties) as input for the lexicographic ordering to compute orderings of all 27 holidays [for details see (Pommeranz et al. 2008)]. Besides an objective comparison, we asked participants to judge which list better reflected their preferences; the one they specified themselves in task 1B or the list generated with the lexicographic ordering method from the input from task 1A.

3.2.4 Procedure

The study was conducted during two weeks. Each experiment took about 45 min and consisted of eight tasks considering preference input, two comparisons of resulting lists and a final questionnaire. The ordering tasks and the navigational task was carried out using cards whereas for the rating tasks we used a computer interface. To compute the resulting lists with ordered holidays using the lexicographic algorithm one of the two present researchers entered the data from the ordering tasks into a computer program. This included the ordering of 27 holidays which we used to compute objective measures of proximity between the different lists. Before the tasks were explained and executed a general introduction was given about the goal of the experiment and the holiday domain. Furthermore, subjects were told that each task stands for itself, which means there is no need to remember anything between the tasks. The presentation of tasks to users was counter-balanced to avoid order of presentation effects.

3.3 Results and discussion

3.3.1 Effort and liking of input methods

After each task participants were asked to rate how much they liked it and how much effort it took them. A multivariate analysis of variance (MANOVA) [1] with repeated measures was conducted to examine an effect for the ordering/rating style (independent within-subject variable) on the perceived effort and liking (dependent variables). We found a significant main effect for ordering/rating style ($F(14,18) = 10.71; p < 0.001$)[2], which was found again in the univariate analysis of the effort rating ($F(7, 217) = 27.91; p < 0.001$), and the liking rating ($F(7, 217) = 3.17; p = 0.003$). As expected, Fig. 3 shows that task 1B (ordering all 27 cards) clearly stands out as least preferred and highest in effort. Figure 3 also shows that more traditional individual property

[1] www.statsoft.com/textbook/anova-manova/.

[2] F stands for *F*-statistic, *p* indicates significance.

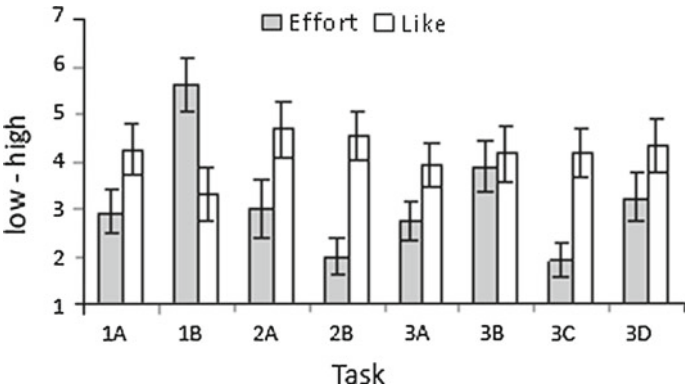

Fig. 3 The mean liking and effort rating of ordering/rating tasks, including a 95% confidence interval

ordering (2B) or rating (3C) tasks were rated low on effort and relatively high on liking. This suggests that people appreciate the relative cognitive simplicity of these tasks; dealing only with a small part of the outcome space complexity. From the tasks that involved evaluating the complete holidays (1B, 2A, 3A, and 3B) it seems that the navigational input method (2A) is most preferred. Considering rating tasks it is interesting to notice that both tasks involving affective feedback are scored equally high in liking as Likert-scale ratings, although affective input is considerably more effortful.

3.3.2 Navigational input method

Besides considering liking and effort we also compared the navigational input method to the ordering of alternatives of holiday properties in terms of intuitiveness and ease of use. With a MANOVA with repeated measures (various ratings as dependent measures, and the task as independent within-subject variable) we found a significant main effect ($F(4,28) = 3.14$; $p = 0.030$) for task, which was only found again in univariate analysis on effort ($F(1,31) = 9.02$; $p = 0.005$) and intuitiveness rating ($F(1,31) = 4.64$, $p = 0.039$). Examining the means shows that participants rated the navigational input method ($M = 0.3.0$, $SD = 1.65$) more effortful than ordering the property alternatives ($M = 2.0$, $SD = 1.16$) and less intuitive ($M = 4.9$, $SD = 1.48$) than the ordering ($M = 5.6$, $SD = 1.32$). This suggests that the more traditional ordering method is preferred.

Studying the tasks in more detail revealed the navigational input method was the only method that enabled participants to enter dependencies between the alternatives of the holiday properties. For a considerable group of the participants (34%) the most and least preferred holiday had at least one equal value. One participant even had two equal values. This means two things. First, a property independent approach is not suitable for all people to describe their preferences. Second, the navigational input method might be an effective approach to determine whether for a specific individual preferences over properties are dependent.

Table 3 Summary of mean liking and effort scores for the tasks 3A-3D and generated lists

Condition	Liking	Effort	Quality of outcome list
Likert and holiday	M = 3.938	M = 2.750	M = 6.188
	SD = 1.318	SD = 1.191	SD = 1.786
Affect and holiday	M = 4.188	M = 3.906	M = 5.500
	SD = 1.575	SD = 1.594	SD = 2.064
Likert and property	M = 4.188	M = 1.938	M = 6.031
	SD = 1.731	SD = 1.014	SD = 2.177

3.3.3 Affective rating

We analyzed the effect of affective rating using a MANOVA with repeated measures. It showed a main effect of affect versus Likert scale rating ($F(2,30) = 24.00$; $p < 0.001$) and property versus whole holiday rating ($F(2,30) = 6.73$; $p = 0.004$) with no significant interaction effect. These main effects were found again in the univariate analysis on effort for affect versus Likert scale rating ($F(1,31) = 46.32$; $p < 0.001$) as well as for property versus holiday rating ($F(1,31) = 13.90$; $p = 0.001$). This means that both affective-, as well as holiday-based feedback are associated with a higher perceived effort in preference elicitation (Table 3). With regards to the perceived quality of the resulting lists generated by the simple algorithms we found a significant main effect for affect versus Likert scale rating ($F(1,31) = 6.12$; $p = 0.019$). This suggests that the algorithmically-generated lists based on affective feedback matched the user's preferences less well than the lists that were generated based on Likert-scale feedback (see column 4 in Table 3). This can be due to two reasons, either the participants did not understand the semantics of the AffectButton well and by that could not express their preferences correctly or the algorithm used to calculate the outcome lists did not work well given the input variables (pleasure, arousal, dominance). We exclude the first reason as it seems the users understood how to give feedback with the AffectButton, as pleasure strongly correlated with the Likert-scale feedback ($r = 0$, $p < 0.001$). Also, previous research suggests that the AffectButton is a valid and reliable affective feedback device when used for rating the affective content of emotion words (Broekens and Brinkman 2009) as well as film music (Broekens et al. 2010b). A deeper analysis suggested that our way of mapping affective dimensions to algorithms that are intended for one dimensional preference values was too simplistic. To understand which factors are most important in predicting the holiday list created by the user (1B) we did a regression analysis given the Likert rating and pleasure, arousal, dominance ratings (stepwise) over all holidays. The same analysis was repeated for the property values (now predicting the property ranking of task 1A). The regression analysis predicting holiday-ranking resulted in a significant model ($r = 0.66$; $F(2,285) = 110$; $p < 0.001$). The model included the Likert rating ($\beta = -0.55$; $t = -9$; $p < 0.001$) and pleasure rating ($\beta = -0.15$; $t = -2.5$; $p = 0.012$) as significant items. The regression analysis predicting property-ranking showed similar results, but included Likert rating and dominance as significant items. In both cases at least one affective factor

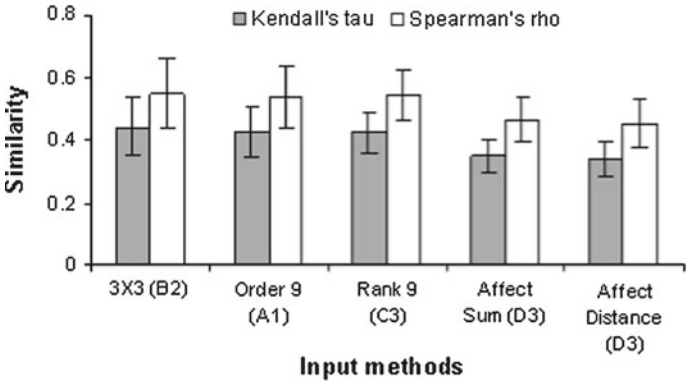

Fig. 4 Similarity of lists generated from different input methods to standard list

was included in the predictive model. This suggests that affective feedback helps the user to express preferences.

3.3.4 Preference ordering

We used the different methods of rating and ordering properties (see Sect. 3.2.3) as input for the lexicographic ordering algorithm to investigate how well this algorithm can perform given a variety of inputs. These methods include affective rating (D3), 9-points rating (C3), ordering 9 property values (A1), ordering the properties and then 3×3 values (B2). The algorithm generated ordered lists for each user, and these lists were compared with the lists that the users specified themselves in the 27-card ordering task (B1). This is essentially a comparison between two rank-ordered lists containing the same items. The similarity between these lists is computed in two ways. Kendall's τ can be seen as a distance measure; it is based on the minimal number of switches between two adjacent items in one list that is needed to attain the second list. Spearman's ρ is another well-known rank correlation method. Both measures are normalized and range from -1 to 1, where 1 indicates that the lists are identical, 0 no relation at all, and -1 indicates reverts ordering. Figure 4 shows the correlation coefficients averaged over participants between the standard list (specified by the participant in task B1) and the lists generated with the lexicographic ordering method with different types of user input. All correlations are significant ($p < 0.001$), which indicates that the generated lists are much more similar to the standard list than random lists. This suggests that different input method combined with lexicographic ordering can result in "true" preference orderings. For more details on the analysis, please see (Pommeranz et al. 2008). As we cannot guarantee that the user-specified list (ranking of 27 holidays) is ideal, given that the task was tedious and little appreciated by the users, it is hard to say how close each generated list came to an ideal list of a person's preferences. Interesting to note, however, is a clear difference between the lists generated from affective feedback and non-affective feedback, whereas the lists generated with affective feedback are less similar to the user-generated lists.

Fig. 5 AffectButton: the cross indicates the position of the mouse cursor inside the button, the face changes accordingly

We believe this is due to a difficulty of translating the 3-dimensional affect feedback into a one-dimensional ranking, as explained in the previous subsection.

3.3.5 Summary of results

The results showed three important aspects relevant for understanding the process of preference elicitation and the match between the user's mental representations and the system's model. First, the results confirmed that cognitively less demanding ordering or rating tasks were perceived as less effortful and liked most by users. So, liking and effort go hand in hand (see similar results of FT2 trial in Knijnenburg et al. (2012)). Second, navigation through the outcome space (moving from item to item by changing an attribute value at a time) enables users to express dependencies between attributes that were not revealed by other methods, and, affective feedback enables users to express preferences in other dimensions additional to liking. Third, effort is not an indicator of how much a method will be liked in these last two cases. Affective feedback and navigation were rated significantly higher in effort than other methods, but still high in liking. We hypothesize that this indicates that users are willing to spend more effort if the feedback mechanism (process and preference representation) enables them to be more expressive (or maybe more entertaining as mentioned by Pu et al. (2012), which is good because it enables the system to extract more preference information and by that build a more accurate user model. 3

4 Study 2: Testing user motivation to give preference detail

To test the above mentioned hypothesis we investigated the tradeoff between giving detailed preference feedback and effort. We examined factors (e.g., familiarity—also mentioned considering recommendations by (Sinha and Swearingen 2002; Pu et al. 2012)) that can influence this tradeoff in an experimental set-up. The focus of this study was on investigating the influential factors in a neutral set-up, i.e., without other motivational factors that could be present in a recommender set-up, such as giving detailed feedback to receive better recommendations or to serve the community of users. Therefore, we chose for a simple content rating task. In order to make sure we did not introduce a bias in ratings by telling people that we investigate the level of detail people give, we instructed participants that the experiment was about creating an alternative top-40 list of famous people and popular music. While this is an incentive to take part in the study in general, there was no incentive to give more detailed preference feedback (Fig. 5).

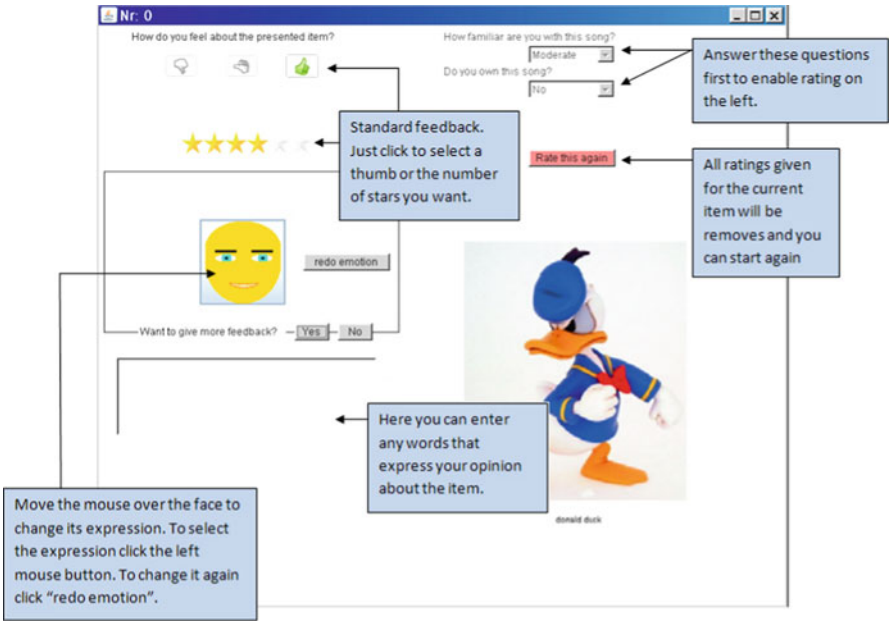

Fig. 6 Instructions for the interface used for testing the level of detail people are willing to give

The study consisted of two follow-up online experiments. Since the second experiment was an enhanced version of the first one, we will only elaborate on the second experiment here. To see the details and results of the first experiment please refer to (Broekens et al. 2010a). The main improvements we made in the study are changes to the interface including using more familiar rating mechanisms such as thumbs and stars, an option to replay a song and the omission of given emotional tags as an input level. Furthermore, we gave a more detailed explanation of the interface (screenshot seen in Fig. 6), let people try all levels before starting the experiment and always showed the following input level during the interaction. By this we reduced problems with the ratings that may have occurred in the first study due to misunderstanding the interface.

4.1 Research hypothesis

Our hypothesis for the following experiment was that *the level of detail persons are willing to give in their feedback depends on content type of the item, familiarity with the item, ownership of the item and opinion about the item.*

4.2 Study setup

We set up a content-rating experiment with content type, familiarity, ownership and opinion as factors and detail as dependent variable. The content was preselected by

the experimenters. We purposefully chose two types of content, music and pictures of famous people, each allowing for different ways to form a preference. We hypothesized that people would be able to form preferences for music spontaneously and give detailed feedback (also in form of affect feedback) while listening to the music even if the song was unknown to them. In the case of famous people we did not expect this behavior as a picture alone does not allow for getting to know the person better and by that would lead to a formation of a directed (positive/negative) opinion. The other experimental factors familiarity, ownership and opinion for each content item were indicated by the subjects during the study.

4.2.1 Material and procedure

Participants received an email with the invitation to participate including a link to the application needed for the study. The study was done online, so subjects did the experiment at a place and time of their own choice. The email contained detailed instructions about how to use the application including a screenshot of the interface (see Fig. 6). After the participants started the application, they were asked to fill in demographic information (age, gender, and education). Then they were presented with an example picture (Donald Duck) that had to be rated using all levels of detail to ensure that all participants were familiar with the interface. After that, the application presented 30 songs and 30 pictures of famous people (one at the time, at random). The pictures were labeled with the famous person's name. Songs were presented as audio samples without an indication of the title or artist name. They could be played as often as the participant wanted. For each picture/song they were asked to fill in their familiarity with the song or person (6-point scale: 0 and 1 was interpreted as not knowing the item and 2–5 was interpreted as knowing the item) and whether or not they owned the song or media concerning the person (yes/no) (see right side of the window in Fig. 6). Then they were asked to give their opinion about the picture/song. The four levels of feedback detail were:

(*1*) *Thumbs-down/neutral/thumbs-up.* All subjects had to rate their opinion about each item using this input level. This is the minimum level of detail that can be given on one dimension (liking) including a neutral position.
(*2*) *A 6-point scale (represented by 6 stars, one star being the minimum).* This is the usual form of giving more detailed feedback, as used on many websites. It introduces the possibility to give a higher resolution of detail but still on one dimension (liking).
(*3*) *Affective feedback* using the AffectButton (see Fig. 4) an interactive button that can be used to give affective (emotional) feedback based on three dimensions: pleasure, arousal and dominance. It is a dynamically changing selectable emotion expression. This introduces the possibility to give fine grained feedback on 2 extra dimensions (arousal and dominance) in addition to the liking dimension.
(*4*) *Free text input.* This option enables subjects to tag the item. We assume this to be the most fine-grained and high dimensional kind of feedback, as essentially users can use any tag they want. People were instructed to use any words that express their opinion about the item.

For each stimulus they had to give at least a thumbs-down/ neutral/ thumbs-up opinion (3-point scale) rating. Neutral was interpreted as no opinion, thumbs-down and thumbs-up were interpreted as having an opinion. After that they had the choice to enter more detail to their opinion or go to the next picture/song. There were 4 levels of detail and each level had to be filled in before the participant could go to the next to make sure the user takes an active decision in whether to give more feedback or not. The user could always see the following level of detail. At every level, subjects could stop giving feedback and go to the next stimulus, except at the obligatory first level.

4.2.2 Participants

A broad range of people, in total 41, participated in the online experiment of which 13 female and 28 male, aged between 11 and 58 (M = 31, SD = 10). Participants have different cultural backgrounds as well as nationalities (including Dutch, German, Swedish, and Chinese) and education level (education level ranged between high school (with an exception of children aged 11 and 13) and post master level, Median = Bachelor).

4.3 Results

Before analyzing the data in detail we checked for any effects of the experimental setup. First, we found that items rated in the last half had an average level of detail equal to 1.8, while in the first half this was equal to 1.9. This indicates that participants gave less feedback later in the experiment, which can be attributed to the time it took (30 min) to rate the 60 stimuli. However, as the effect is rather small, this poses no problems for interpreting those items rated later. Second, we found a healthy distribution of thumbs-based feedback about items (26, 40 and 34% of the cases were rated as bad, neutral or good respectively). The fact that 40% were rated as neutral and 60% with a positive or negative opinion allowed us to use opinion as factor in the analysis. Third, we found positive correlations (all correlations significant and $r > 0.7$) between the ratings entered in levels 1–3 indicating that users were consistent when rating an item with different input methods (thumbs, stars or AffectButton).

In the further analysis we focused on main effects, as familiarity, ownership and opinion are not experimental controlled variables. Ratings were aggregated per subjectXfactor, averaging over the rated levels of detail, resulting in 41 paired measurements per main effect analysis. We interpreted familiarity ratings < 2 as unfamiliar and ≥ 2 as familiar. For the factor opinion we differentiated between directed opinions (thumbs-up and thumbs-down ratings) and neutral.

Our hypotheses were confirmed with respect to the influence of ownership, opinion and familiarity, and to a lesser extend the influence of content.

Subjects rated familiar items (M = 2.12, SD = 0.87) with more detail (paired $t(40) = -5.19$, $p < 0.001$) than unfamiliar items (M = 1.72, SD = 0.75). Items that are owned are rated (M = 2.19, SD = 0.88) with more detail (paired $t(37) = -4.12$, $p < 0.001$) than items that are not owned (M = 1.83, SD = 0.78). These two effects might influence each other, owned items have a much higher chance of also being

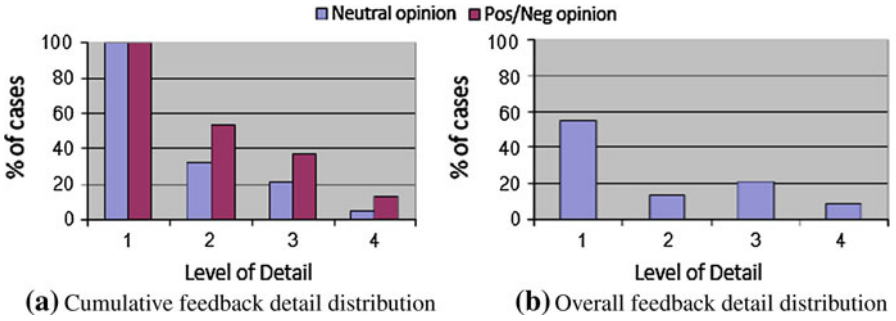

Fig. 7 Distribution of four levels of feedback detail

familiar. For these two factors we checked the interdepency using a 2×2 repeated measured ANOVA, and this showed indeed that when taking both factors into account, only ownership remained a significant factor ($F(1, 37) = 25.2$, $p < 0.001$), and familiarity did not ($F(1, 37) = 2.78$, $p = 0.104$). The interaction effect was not significant ($F(1, 37) = 0.20$, $p =$ ns). When subjects had a positive or negative opinion (M = 1.98, SD = 0.79), they rated with more detail (paired $t(40) = -5.77$, $p < 0.001$) than when they had no opinion (neutral) (M = 1.60, SD = 0.75)). Further analysis revealed that a positive opinion was related to rating with the highest amount of detail (M = 2.10, SD = 0.81), followed by negative opinion (M = 1.85, SD = 0.77), and no opinion having the lowest detail (M = 1.57, SD = 0.73). All differences were significant in paired t-tests at the level of $p < 0.01$.

We did find a significant effect of content type (paired $t(40) = -2.58$, $p = 0.014$). However, the difference was small. Music (M = 1.91, SD = 0.73) was scored with only a little bit more detail than images (M = 1.78, SD = 0.80). This means that, although the effect of type of content was significant, the effect was relatively small compared to the effects of the other three factors (a difference in means of about 0.13 compared to around 0.40 for the other factors). The tendency to give detail seems to be a factor that should be explained from within the subject, an important finding in light of preference elicitation.

Finally, we show the distribution of the level of detail used to rate cases in Fig. 7a and Fig. 7b. Each bar in Fig. 7a represents the number of cases rated with a level of feedback (so, if a user stopped at level 3, he/she rated the item with level 1, 2 and 3; explaining why 100% of the cases was scored with at least level 1, as this was obligatory) split between having an opinion (positive or negative) or not having an opinion. Fig. 7a shows an overall trend for using more feedback when a positive or negative opinion is present. Most notably, in about 40% of the cases where a positive or negative opinion is present, affective feedback (level 3) was used to express more detail. Figure 7b shows the distribution of highest level of detail used. Each bar represents the number of cases at which a user stopped giving feedback (so, if a user stopped at level 3, it is counted under level 3 only). In general, Fig. 7b shows that the majority of cases was scored using only thumbs-based feedback. Interestingly, more cases ended with affective feedback then with stars-feedback, indicating that when more feedback is given, a preference exists for giving multidimensional affective feedback (although this difference was not statistically significant in a paired T-test comparing

the number of times raters stopped at level 2 versus level 3, (paired $t(40) = -1.15$, $p < 0.256$)). Finally, our results show that text input was used least often by the participants. As described by others, e.g. Ames and Naaman (2007), incentives for tagging often have a social basis, e.g., to help others in a community to find content. The lack of community-based motivators in our study may be one of the reasons for low tagging responses. Therefore, the outcomes with regard to tagging as a feedback level may be less representative for preference feedback in general.

4.3.1 Summary of results

The data analysis showed that familiarity, ownership and having an opinion about that item are the main factors in influencing the preference detail people are willing to give, and thus the amount of effort they are willing to put into giving feedback. As we found only a small difference in detail for pictures versus music, we can at least tentatively conclude that the willingness to give feedback is not so much triggered by content types but more so by what one thinks or knows about content. Although both content types in our study could be owned in form of media (music, books, film) we hypothesized that people would generally be able to give more detailed feedback for music items as they could form a preference by listening to the music during the study. This hypothesis was not confirmed and follow up studies should be done to investigate different content types.

Our results also show that multidimensional affective feedback is used when people have the choice to do so. Moreover, people in general prefer to give more feedback in the form of multidimensional affective feedback (at least when they can use the AffectButton) than to give more feedback using a finer grained one dimensional method (stars). This suggests that a preference elicitation interface—trying to adapt the amount of feedback detail it extracts from a user—should either give thumbs or stars as first level, after which the next level of detail should be affective or at least add a new feedback dimension. As the study set-up (no post-questionnaire or interview) did not allow a deeper analysis of people's reasons for giving a certain level of feedback, we do not know exactly why people who gave detailed feedback stopped more often at the AffectButton level than the stars level. One explanation would be that they just liked the AffectButton. However, we can exclude this reason because the effort that people spent to go all the way to the affective feedback level varied under the experimental conditions. Especially in the case where people had a positive or negative opinion they went to this level. If they just liked the AffectButton or wanted to try it due to its novelty they would have used it equally across all conditions. Our interpretation of why participants used the AffectButton is that due to its multiple dimensions (pleasure, dominance, arousal) the expressive power is enhanced. Whereas the stars only offer a finer grained scale on the liking dimension compared to the thumbs, the AffectButton allows people to express their attitude towards an item in two additional dimensions dominance and arousal. These dimension could be more applicable if people have a strong opinion about an item and feel the need to express this opinion. This has to be confirmed in future studies.

The obtained results give interesting insights for the design of preference elicitation interfaces used in different systems including recommenders, especially with regard

to adaptive preference elicitation (also suggested by Pu et al. (2012)). Although users of recommenders may have extrinsic motivations to give detailed preference feedback in general, it is important to know in which cases they are able or willing to give more details and in which form. Knowing that a user has a directed opinion (obtained by simple thumbs input) or is familiar with an item can be used to ask the user for more detailed and multi-dimensional input. In the case of a neutral opinion a system asking the user to spend effort of giving more details that she may be able to provide can be perceived as annoying and should be avoided. Note also the difference between knowing an item or having a directed opinion. Often in recommender systems people do not know the items, however, by providing samples (e.g., music, book excerpts) the user can still form an opinion and by that be motivated to give more detailed feedback.

5 Study 3: Exploring the preference elicitation process with interface prototypes

Until now we have looked at different ways to enter a preference including rating, ranking and navigating in the first experiment, and different detail levels of rating in the second experiment. Besides motivation to spend effort, the design of the elicitation process is a second factor we would like to investigate. As Pu et al. (2003) pointed out "stating preferences is a process rather than a one time enumeration of preferences that do not change over time". Therefore, it is important to explore how to facilitate the human preference construction by the means of preference elicitation interfaces that are intuitive for users and allow as well as motivate them to be expressive. This can only be done by involving the user in the design process. We addressed the interface design in our third study. In specific, we explored different ways of structuring the process of preference construction in an interface. Next, we elaborate on an exploratory study in which we investigated four fundamentally different processes of eliciting preferences represented in four hi-fi preference elicitation interface prototypes. Similar to the suggestion of Pu et al. (2012) of comparing systems or interfaces side-by-side in user experiments we presented the four prototypes to each participant. In addition, we also allowed participants in a creative participatory design session to construct new ways of eliciting preferences based on (elements of) the four interfaces.

5.1 User-centered prototype design

We created four interface prototypes for eliciting preferences. To be able to include decision context into the interface we chose to elicit preferences for a certain domain. The domain in this case was jobs, which allowed us to show example job offers as decision context. Different from the holiday domain we used in the first study where it was important to arrive at a ranking of outcomes, we wanted to support people in this study in (1) constructing their job preferences and (2) getting an idea of the resulting preference profile. Choosing to negotiate for a new job is different from picking the next holiday destination as it has a bigger impact on people's lives. This is also why we focused more on underlying interests which are stable over a longer time period and influence one's preferences. The navigational input we used in the first study was not applied in the following prototypes because it focuses more on finding the best outcome than giving people an understanding of their preferences.

To design the prototypes we first compiled a set of design guidelines from the relevant literature. Please refer to our previously published work (Pommeranz et al. 2010) for the detailed guidelines.

Given the set of design guidelines we selected appropriate existing interface elements (e.g., *ValueCharts* (Carenini and Loyd 2004), a virtual job agent) and created new ones (e.g., job offer clusters, post-it notes with preference information). Next, we combined these elements into the four interfaces. There are, of course, many combinations of elements possible, which would lead to an exponential number of prototypes. Instead of creating this high number of prototypes we combined the elements in a way that each prototype differs in how it structures the elicitation process and how it interacts with its users. Structuring the process in different ways can be linked to how people process information. Therefore, we created different ways of user-system interaction, each supporting one thinking style based on the theory by Gregorc (2006). The mind styles theory categorizes people based on perceptual and ordering preference. Perceiving information can be abstract (based reason and intuition) and concrete (using one's senses). The order of information processing can be sequential or random. This leaves us with four types: concrete sequential, concrete random, abstract sequential and abstract random. Concrete sequential thinkers like order and logical sequence and learn best in a structured environment. Concrete random thinkers like experimenting to find answers, using intuition and therefore, learn best when they are able to use trial-and-error approaches. Abstract sequential thinkers like analyzing situations before making a decision or acting and applying logic in solving or finding solutions to problems. Abstract random thinkers like to listen to others and establishing healthy relationships with others. They focus on the issues at hand and learn best best in a personalized environment. Based on these different characteristics we, first, chose an overall way of interaction, that would fit a mind style, e.g., a structured, step-wise approach for the concrete sequential thinker. Second, we identified which elements could be combined to achieve such an interaction, e.g., in the step-wise approach first a simple selection of values, then *ValueCharts* (Carenini and Loyd 2004) showing links between values and fit of job offers, then tables with details for one offer and last an overview/summary showing the elicited preferences.

Following this approach we could create meaningful combinations of the elements. However, people do not perfectly fit into one style but have a unique combination of characteristics. In the evaluations we did not try to find the best prototype to choose and develop further, but rather evaluate the different design elements used. In the following creative session we then gave the participants the chance to combine them in different ways that they preferred and found more usable. We implemented the designs as hi-fi prototypes because this was the best way to ensure that the users get a feeling for the interaction with the system. In the following sections we describe the four prototypical interfaces highlighting the interface elements used (*italic font*).

5.2 Conversation: abstract-random style

This prototype (Fig. 8) focuses mainly on a collaborative interaction style, in particular the natural interaction, between the user and the system employing mixed-initiative.

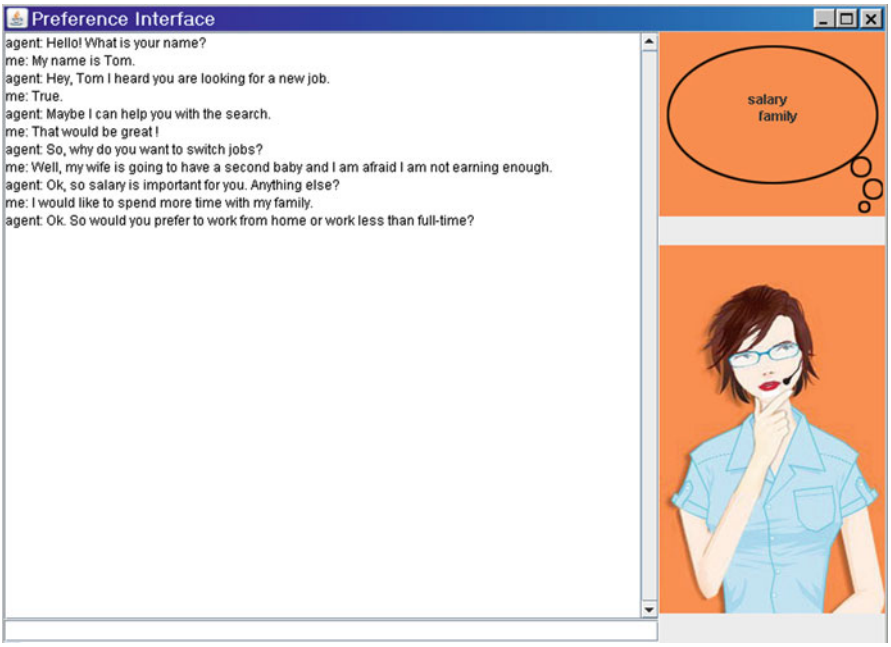

Fig. 8 User interface for conversation with intelligent agent

A natural way of building a preference model is being questioned by an expert, who can understand what you want by asking the right questions. In real life this could be a job agent. Since this is a known and intuitive way for people to express their preferences we designed a very simple interface based on a conversation with a *virtual agent*. Another design criterion used in this prototype is system transparency. We tried to reach transparency by two means: the affective state of the agent and the "thoughts" of the agent regarding the user's preferences. In the first simple version there are three states of the agent implemented, speaking with positive expression, thinking and confused. The second feature is a *thought bubble* above the agent's head. In the beginning of the conversation it is empty. It gets filled with tags (forming a *tag cloud*) whenever the agent could retrieve an interest or issue from the chat that seems to be important to the user. To ensure natural interaction during the evaluation sessions the prototype was implemented as a client-server application for a Wizard-of-Oz testing, i.e., the role of the agent was taken by a real person who was invisible to our participants.

5.3 Post-its: concrete-random style

This prototype focuses on supporting the constructive nature of human preferences. Two things inspired the interface shown in Fig. 9. First, preferences are rather unstructured to begin with. They are not necessarily linked to each other. Second, preferences change dependent on the context.

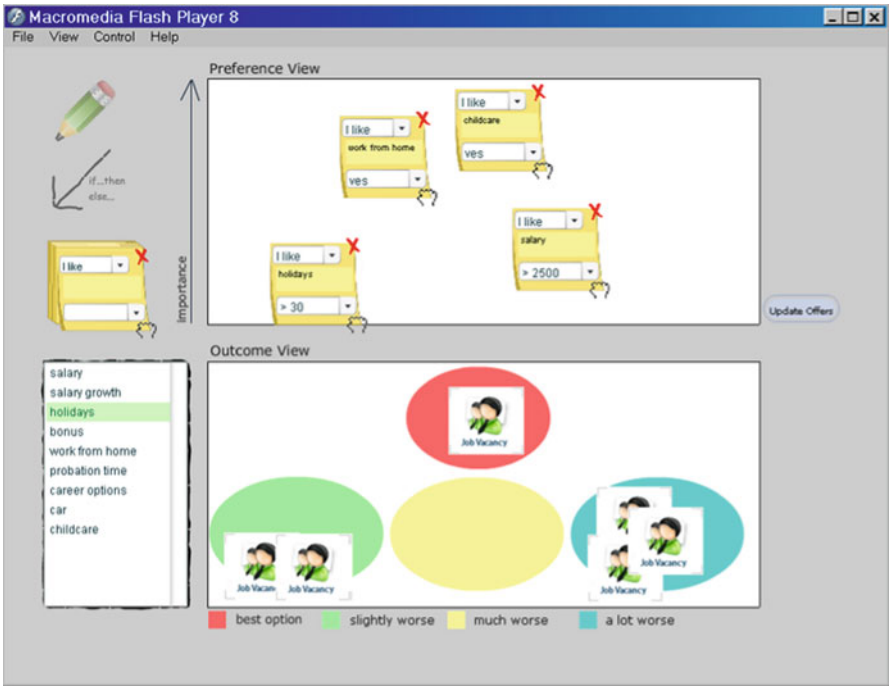

Fig. 9 Visual construction of preference profile

We used *post-it notes* as a real-world metaphor for organizing thoughts. The interface allows dragging as many post-it notes onto the so-called preference view as the users want. They can then write the important issues on the notes, add a value and specify whether they like, want, dislike or do not want these issues. At any time they can remove, add or drag around the post-its to structure their profile. More important issues can be dragged further up and less important ones down.

At the same time we provide the users with the needed context to make their choices of how to structure the notes. The context is a number of job offers in the *outcome view* that get arranged into *clusters* according to good fit to the current preference profile. This could be done in real-time while the user is interacting with the notes to give immediate visual feedback. For simplicity reasons the arrangement takes place after pressing the "update offers" button. In the evaluation we discussed both options.

5.4 Comparison: abstract-sequential style

In this prototype (Fig. 10), based on the value-focused thinking approach the user chooses from a list of *interest profiles*: family-oriented, money-oriented, career-oriented, or self-fulfillment. We chose these profiles because they represent life goals that are linked closely to jobs. In a real system this needs to be scientifically proven. In order to help people choose a profile we added a visual stimulus to each profile. We chose a moodboard-like collection of images as often used in advertising to

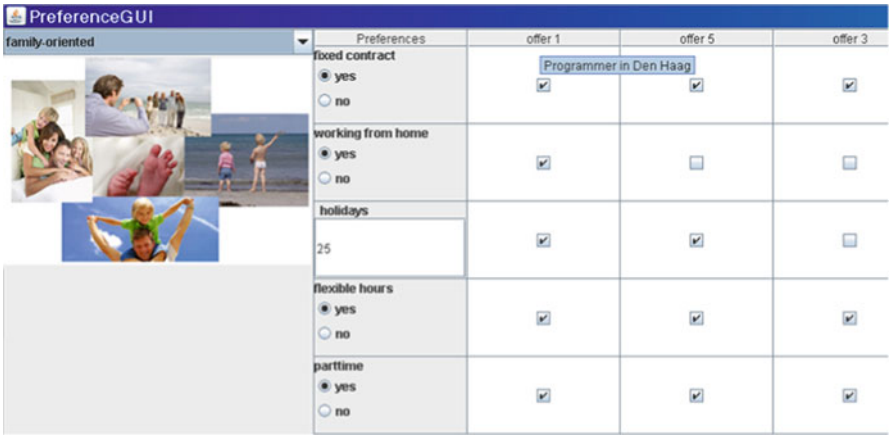

Fig. 10 Choosing and adjusting a default profile

convey a certain feeling or style. Each moodboard consists of a collection of images that represent the particular profile at a glance. The selection of images aimed at giving a diverse view of the profile (e.g., career profile: doctor, model, business man etc.) in order to avoid that users focus too much on a particular image. In the second step, the user received a filled-in list of preferences that fit the chosen profile. To give the user decision context to understand their preferences and refine the preselected ones we present a list of job offers.

The data is presented in form of a *decision matrix* similar to the ones often used on product comparison websites. Both the preferences and the offers are ordered by importance, from top to bottom and left to right respectively. By hovering over the job offer with the mouse the user gets a description of the jobs. Since we are not expecting that people fit perfectly into a profile the users have the chance to adjust the preference values as well as the ordering. As soon as they enter a new value or drag and drop the rows around the job offers get ordered based on the new input to give visual feedback of the consequences. We use a lexicographic ordering, since it delivered good results in our first study. During the evaluations we also discussed the possibility for the user to drag the job offers, which will result in adapted preferences.

5.5 Stepwise: concrete-sequential style

In the fourth prototype (Fig. 11) the interaction is similar to the APT Decision agent (Shearin and Lieberman 2001) following three steps: (a) letting the user give only a small number of preferences, (b) then receiving a list of offers to compare and (c) giving feedback to attributes that appear in the offers. We adapted this approach and ask the users in the first stage about their three most important interests (e.g., work-life balance or professional development) instead of negotiable issues. By that we follow the value-focused thinking approach (Keeney 1992). After choosing the interests the user enters the interface depicted in Fig. 11. The interface aims at helping

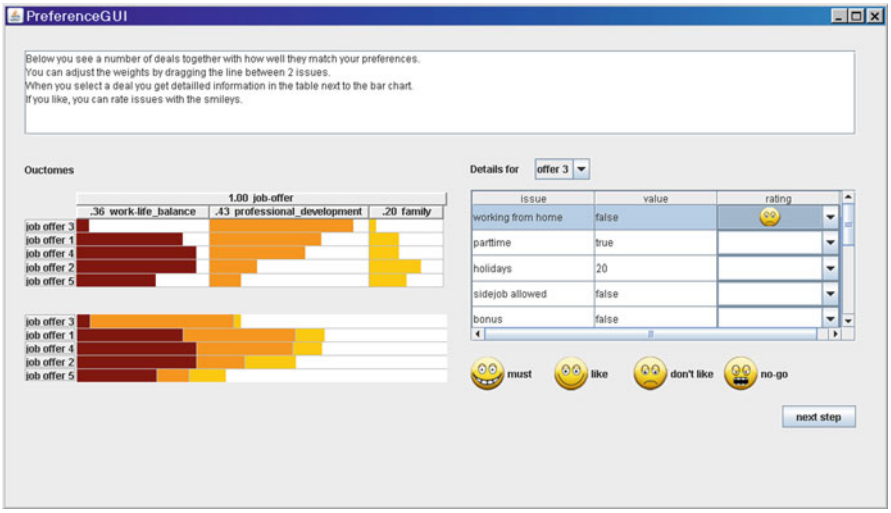

Fig. 11 Preference elicitation using ValueCharts and affective feedback

the user explore several job offers (decision context) with regard to the user's interests and by that construct his preference profile. To compare the offers we used *ValueCharts* (Carenini and Loyd 2004). The user can adjust the (initially equal) importance of the interests. He receives immediate visual feedback on how well the job offers match his interests, while adjusting the importance by growing or shrinking of the job offer bars. By double clicking on an interest the job offers get ordered according to good fit. The interface also offers the possibility to critique any attribute of a job offer. Once the user chooses to look at a job offer in more detail the table on the right gets filled with all values for existing attributes in the job offer. The users are free to give affective feedback on any issue-value pair they want, but are not forced to rate all of them. We included "musts" and "no-goes" as hard constraints in the system, i.e., a job that does not comply with either will not be an option to the users. When the user is done exploring his options, the interface reveals an overview over elicited preference profile, which supports the transparency of the system.

5.6 Exploratory user study

In order to understand in depth how we can support the human process of preference construction with an adequate interface we did an exploratory user study. By collecting large amounts of qualitative data we aimed at informing the design process of preference elicitation interfaces. Our prototypes served as a means to discuss relevant issues to the participants and foster a creative process rather than finding usability problems of the prototypes. We specifically aimed at receiving feedback on the different interface elements used and how they can be combined in an optimal way to support the process of constructing one's preferences. In the following sections we elaborate on the set-up of the study and its outcomes.

Fig. 12 Interface elements for creative session

5.6.1 Material

We used the four hi-fi prototypes elaborated above in this user study. Furthermore, we created paper versions of all interface elements we had used in the four hi-fi prototypes (Fig. 12), e.g., the virtual agent, the post-its, the value charts or the tag cloud, as well as standard interface elements such as text fields, check boxes, sliders, comboboxes, etc. Additionally, we had a number of blank papers, pens and scissors to give the participants the chance to create their own interface elements. These materials were used by the participants in the second part of the session to design their own preference elicitation interfaces.

5.6.2 Participants

We included 5 male and 3 female participants. The participants were people with different backgrounds, i.e., artificial intelligence, affective computing, design, linguistic and visual perception. We intended to have a mixture of people with diverse backgrounds in order to get different views on the interfaces.

5.6.3 Procedure

The study was divided into two parts: eight individual sessions with one participant at a time and a collaborative creative session with all eight participants.

The sessions were carried out in a lab setting. Participants were first briefed about the background of the study and the intention. We emphasized that we would like to receive constructive feedback on the different elements of the prototypes to inform future designs of preference elicitation interfaces. After the briefing we provided the

participants with a scenario describing a 35 year-old family father who would like to switch jobs. We chose using a scenario rather than the participants' real job preferences for two reasons. The first is of practical nature: Since our interfaces were limited regarding their domain knowledge, we wanted to make sure that the issues and interests people want to express preferences over were available in the system. The second reason was trying to get participants to use the interfaces in a similar way to be able to compare the feedback. The participants then interacted with each prototype for about 10 min on average. The order of prototypes was changed per participant to avoid ordering effects. Their task was to fill in job preferences that would fit the person described in the scenario. During the interaction the participants were asked to think aloud. All actions and voices of the participants were recorded by the help of the Camtasia Studio software (http://www.techsmith.com/camtasia.asp). Each prototype saved the preferences to a log file. The person leading the evaluation intervened whenever participants seemed to be lost, asked for help or forgot to think aloud. Often the evaluator and the participant already got into discussions about new ideas and problems with the interfaces during the interaction. After interacting with the prototypes we interviewed the participants informally to get a grasp of their experiences, constructive critique and new ideas. We used printed screenshots of the interfaces as reminders. Together with the evaluator new ideas were developed and discussed and drawn onto the printed screenshots.

The individual sessions were followed by a creative session with all eight participants. Goal of this session was to explore new ways to structure the elicitation process in the interface from the users' point of view. The session consisted of two parts, a group discussion and participatory design session aimed at creating new paper prototypes. After a short introduction to the meeting including a reminder of all four interfaces and the agenda, we started a general discussion about the interface elements. The discussion took part with the whole group for about 20 min. After that we split the participants into two groups of four participants each. Each group was provided with the same set of materials described above and instructed to use the material to create their own version of a preference elicitation interface. They were encouraged not only to combine the elements existing in the four presented prototypes but also create new ones. This part of the creative session was planned for about 30 minutes. However, since both groups were not done within that time frame, the session took about 1 h. The creative session was concluded with a presentation of the two groups' results to each other. During the presentation new discussions arose about design decisions.

5.7 Results

During the individual sessions, the informal interviews and the group discussions we gained detailed feedback on the four prototypical interfaces as well as new ideas, including tips and new combinations of the interface elements. In order to extract the feedback from the collected data we annotated the recordings from the individual sessions using NVivo (www.qsrinternational.com). Based on the annotations we created a table with feedback on each prototype per participant. In addition, we made a list of observations of how users used the prototypes and a list of new ideas that

Table 4 Feedback per interface element

Element	Positive	Negative
virtual agent	Engaging, straightforward, natural way to enter preferences, easy to use, no constraints	Low feasibility, too slow, vague, profile not clear, no comparison of jobs, depends on how good the agent is
tag cloud/thought bubble	Gave users a hint of what the system is "thinking"	-
post-its	Liked by most participants	Yoo difficult to operate, too many hidden things
Outcome view/ clusters of offers	Participants liked seeing and exploring job offers, tie between preferences and consequences	offers were not draggable
Interest profiling	Most users liked it as a starting point, efficient, less effort, use of pictures	Trouble deciding on one fitting profile (preferences should already be visible when choosing)
Decision matrix	Similar to product comparison websites	Problems with visualization: difficult to understand that offers are ordered and draggable
ValueChart	Gives an overview of how the job offers fit the profile but without losing the detailed information of how well each interest/issue scores in an offer, immediate visual feedback	No link between the ValueChart and the table with issue ratings
Affective feedback	Natural	Must-have smiley was not interpreted as a hard constraint
Preference summary	Was liked, gives overview, clarifies preference profile	Should appear while you are adjusting your preferences, missing interactivity

were discussed in the individual and the collaborative session. Next, we will elaborate on the main findings that are relevant for designing preference elicitation interfaces. Table 4 shows the positive and negative comments per element. For a more detailed description of the feedback per interface element, we refer to our previously published work (Pommeranz et al. 2010).

Some of the interface elements had obvious usability issues, e.g., the checkboxes in the decision matrix which were not interactive. These were due to programming difficulties or time constraints during the creation of the prototypes. As we already anticipated some of these issues before conducting the study, a researcher was present

Fig. 13 Design proposal group 1 (*left*) and group 2 (*right*)

during the study to clarify such issues whenever a participant seemed to have a problem. We asked the participants to focus on the fit of the different elements for entering preferences. We also asked for constructive feedback on improving and combining the elements once the users understood how the elements worked. Regardless of the interface elements used, an important aspect for our participants was the ability to explore the link between their preference input and the desirability of outcomes (in this case job offers). An element that the participants found highly useful for this exploration were the ValueCharts, because they give immediate visual feedback while keeping details about the selected interests/issues. Another well-liked element supporting the construction of preferences was the post-it note. Furthermore, using default profiles was anticipated since it gets the elicitation process started more easily than starting from scratch. Based on a given profile a number of common preferences can already be displayed. Carefulness needs to applied with designing the interface in this case. Most people had trouble fitting themselves into one of the four given profiles. Therefore, a more flexible input of the separate interests should be possible. During the collaborative session several ideas were mentioned to create a more flexible input of interests, e.g., using a questionnaire, pictures combined with sliders for importance or the virtual agent.

This feedback was also reflected in the new preference elicitation interfaces that the two groups designed in the second part of the collaborative session. The results are depicted in Fig. 13.

Both groups were in favor of having three views on their preferences, i.e., the underlying interest profile, the issue preferences including an importance ranking and values for each issue, and a number of job offers representing the decision context. Whereas group 1 left it all up to the user where to start in the interface and which views to maximize/minimize, group 2 focused on (stable) underlying interests in the first step before giving a number of preferences in the context of example job offers.

Regarding our hypothesis (users are willing to spent more effort if the feedback mechanism (process and preference representation) enables them to be more expressive) we can conclude that people are indeed willing to spend more time on investigating the links between their interests, issue preferences and outcomes (jobs). This was mentioned by the participants and observed by the researcher during the study.

The participants emphasized that it allows them to be more in control of creating their own preference profile, which will then be used by the system. Having that level of control and understanding of the system's model was anticipated by the participants (see similar results on user control in Pu et al. (2012)). We believe, this shows the importance of supporting this constructive process in order to make the outcome of the system comprehensible and trustworthy. However, participants also expect the system to support them during this exploration of links where possible, e.g., by offering default preferences based on profiles and by giving immediate visual feedback while adjusting the different elements. This shows that they are not willing to spend much time on cognitively demanding tasks that do not seem necessary (e.g., creating every single post-it).

6 Discussion and design guidelines

In the three studies we presented, we tackled the problem of designing user interfaces for explicit preference elicitation. Two important aspects to consider when designing such interfaces are: matching the mental models of users' preferences to the representations of the system and supporting the process of human preference construction so that "true" preferences can be elicited by the system. In our first study we investigated both aspects by studying (a) different ways of giving preference feedback (process) and (b) what kind of information the methods deliver and how the outcomes (ranked holiday lists) compare to a baseline created by the participant. We learned that effort generally goes hand in hand with liking when comparing tasks that are similar with regard to the process and type of input (e.g., rating with Likert scale or ordering attributes). However, in cases where the process (navigation) and the type of feedback (affective) was more sophisticated in terms of expressive power and understanding of one's own preferences, participants rated the methods high in liking even though the results show a substantial increase in perceived effort or are less easy to use. Therefore, we hypothesized that people are willing to spend more effort if the feedback mechanism enables them to be more expressive. In the two following studies we tested this hypothesis.

The following online rating experiment focused on the motivation people have to give feedback in a neutral setting (by that we mean that they are not motivated, e.g., by social aspects as it is often the case in recommender systems) and which factors influence that motivation. The main factors we found were familiarity of an item (also predicted by ownership) and whether people already have a formed opinion about the content. Furthermore, we could conclude that once people decided to give more levels of feedback they went more often all the way to the affective feedback level than just the 6-point star rating. While we can conclude safely that an interface should offer motivated users the possibility to enter more detailed feedback (guideline 1), we do not know the exact reasons for people to enter affective feedback (with the AffectButton). Given the fact that the star based rating offers only a finer grained one-dimensional feedback (liking) compared to the thumbs, whereas the AffectButton offers two additional dimensions (dominance, arousal), we believe it offers more expressive power. In the case of people having a defined (positive or negative) opinion on an item they

might feel the need to express this opinion with more detail and on more dimensions. Based on the fact that people liked giving affective feedback despite increased effort (compared to traditional methods) in the first study and more participants in the second study stopped at this level than at the stars level we can say that affective feedback should be considered when detailed preference feedback is needed (guideline 2).

After studying motivation in this structured way, we took a more explorative approach in the third study to understand how to design the process of preference elicitation interfaces from a users point of view. By actively involving the participants in the design process we were able to understand how they prefer an interface to be designed. We learned that an important aspect of the process is that it allows people to understand their own preferences and that people feel in charge of creating their profile as opposed to just answering questions that are used by the system to build the profile. In particular, being able to explore their preferences from different angles including underlying interests and consequences (in form of rankings of decision outcomes) within the same interface supported people's process of constructing their preferences. Participants liked design elements that supported this exploration in a natural way that allowed immediate visual feedback. Whereas design guidelines established earlier (Pu and Chen 2008) already point to giving decision context and immediate visual feedback, we would like to add the importance of exploring *interests, preferences* and *outcomes in the same physical space*. This enables the user to receive feedback on three related concepts at the same time while adjusting one of the views, which is not the case in interfaces proposed by Pu and her colleagues. As participants were in favor of this kind of interaction and view of their preferences we believe there is a basis for a new guideline (see guideline 3).

Furthermore, the study supported results from our first study regarding the effort people would like to spend. People preferred using interest profiles as a first step and getting preference suggestions from the system (on an attribute basis). The comments of the participants indicated that they considered starting *from scratch* (i.e., filling in values for every attribute themselves) as an effortful task that seems to be redundant if the system is able to give suggestions based on the interest profile (guideline 4).

Given the results from the three studies we established the four following design guidelines for preference elicitation interfaces:

(*1*) *As motivated users are willing to spend more effort, users should be given the option to express more detail if they feel the need to do so.*
(*2*) *Affective feedback should be considered as a way for specifying detailed preference feedback with multiple dimensions.*
(*3*) *The user must be able to explore his/her interests, preferences and outcomes in the same physical space in a way that gives immediate feedback on the links between the three concepts.*
(*4*) *Profile/interest selection serves as an easy (i.e. reduced effort) starting point for showing default preferences that can subsequently be adapted by the users.*

These specific guidelines are meant to extend the more general existing guidelines from the literature (e.g., giving immediate visual feedback, context in form of example outcomes, focusing on values, any preference in any order etc., see Pu et al. (2003) and Pommeranz et al. (2010)), instead of being an exhaustive list by themselves.

6.1 Limitations and further investigations

Our goal was to inform the design of preference elicitation interfaces in general. The results should therefore not be restricted to specific tasks or systems. We believe that they are generally valid for preference elicitation done for recommender systems as well as decision support systems. However, the research questions we investigated had an influence on the choice of domain and type of tasks for each experiment. We chose holidays in the first study with the assumption that most people either have holiday preferences or are able to construct them easily. The focus of the study was on preference input mechanisms in connection to their use in an algorithm that computes a preference ranking over outcomes. In order to compare the different outcoming lists to a baseline we asked people to give their own ordering of the items in the outcome space. This limited the size of our domain to a great extend, as with nine (3 properties times 3 alternative values) property values the number of holidays that could be created was already 27. We thought that sorting an even higher number of holidays would be an overwhelming task for the participants, and the effort ratings confirm this. The limitations to the value space of the properties, however, poses difficulties to transfer the results of the study to other domains as most real-world applications of recommender systems deal with a high number of values, properties and outcomes. Especially, the navigational task would not be feasible in the same way as in study 1 if the number of values and properties was higher than three. It would have to be adapted by using an intelligent algorithm showing only a small portion of the outcome space at a time. If people still liked the task in other scenarios would have to be retested. Another aspect of study 1 that leads to a limitation of the results is the fact that we tested only the lexicographic algorithm to generate outcome rankings. To generalize the results connected to the liking and similarity of outcome lists other algorithms should be employed and a detailed investigation needed to be done of how to map the 3-dimensional feedback obtained by the AffectButton into a 1-dimensional outcome ranking.

Considering the second study two things need further investigation. One is the relation of the strength of an opinion to the need to give feedback. In the current set-up this was not possible. Second, we need to further investigate why people preferred to give more detailed feedback in form of affective feedback with the AffectButton and how to use this multi-dimensional feedback as additional information on the user's preferences.

Based on the results from the third study questions about the design of the interest profiling arose. Interesting work that we will consider for this aspect has been done by Kay (2000), who focused on the scrutable student models in learning environments. Scrutable stereotypes are used to support learners in tuning their student models. By scrutinizing the models the user can also understand what the system believes about them and what these beliefs are based on.

With regard to the guidelines, it has to be noted that whereas the first guideline is applicable to any preference elicitation task, guideline 2 is more suited for domains in which the user is either familiar with the items or can easily form an opinion about an item (e.g., music or book recommenders). Guidelines 3 and 4 are focused more on domains in which users have to construct preferences (due to being a novice or

changing preferences). Guideline 3 is especially helpful for negotiation/decision support systems or recommenders that advice users in important decision-making tasks (real estate, financial advice, job negotiations etc.). Guideline 4 is useful in domains where the number of properties are very high (e.g., cameras or other electronic devices with many features).

7 Conclusion

The importance of preference models for intelligent systems of different sorts (e.g., recommender systems, decision support systems) has long been acknowledged by researchers. However, focus within the area of preference modeling has been mainly on algorithms for computing preferences (elicited in form of numbers and weights) and system representations. A group of researchers has lately focused on designing methods for preference elicitation from a user's point of view, that are in accordance with behavioural decision making theories (constructive preferences). More research in this direction is needed to give researchers and practitioners a good understanding of how to design trustworthy preference elicitation interfaces, that involve users in the process of constructing their own profile that reflects true preferences. We have pointed to two main difficulties that still exist, namely matching people's mental models and the influence of the elicitation process on the elicitation outcome. Furthermore, our studies showed that affective factors are important to consider in preference elicitation. The results suggest that more research in this direction seems worthwhile. With the studies presented in this paper we have only done a first step towards an optimal design of preference elicitation interfaces. However, we believe that the results we obtained from tackling the problem in different ways (with structured experiments and explorative, participatory research) help in advancing the research on interface design for preference elicitation and encourage others in the field to follow that route. Our own research agenda includes usability testing of the interface design obtained from the third study as well as investigating the links between underlying values and attribute preferences.

Acknowledgements We would like to thank the participants of all three studies. This research is supported by the Dutch Technology Foundation STW, the Applied Science Division of NWO and the Technology Program of the Ministry of Economic Affairs. It is part of the Pocket Negotiator project with grant number VICI-project 08075.

References

Adomavicius, G., Tuzhilin, A.: Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng **17**(6), 734–749 (2005)

Aloysius, J.A., Davis, F.D., Wilson, D.D., Taylor, A.R., Kottemann, J.E.: User acceptance of multi-criteria decision support systems: The impact of preference elicitation techniques. Eur J Oper Res **169**, 273–285 (2006)

Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI conference on Human factors in computing systems, Gaithersburg (2007)
Ardissono, L., Felfernig, A., Friedrich, G., Goy, A., Jannach, D., Petrone, G., Schäfer, R., Zanker, M.: A framework for the development of personalized, distributed web-based configuration systems. AI Mag **24**, 93–108 (2003)
Barneveld, J., Setten, M.: Designing usable interfaces for tv recommender systems. In: Personalized digital television, human-computer interaction series, vol 6, pp. 259–285. Springer, Amsterdam (2004)
Bellucini, E., Zeleznikow, J.: Developing negotiation decision support systems that support mediators: a case study of the familywinner system. J Artif Intell Law **13**(2), 233–271 (2006)
Bettman, J.R., Luce, M.F., Payne, J.W.: Constructive consumer choice processes. J Consum Res **25**(3), 187–217 (1998)
Broekens, J., Brinkman, W.P. : Affectbutton: towards a standard for dynamic affective user feedback. In: Mühl, C., Heylen, D., Nijholt, A. (eds.) Proceedings of Affective Computing and Intelligent Interaction (ACII), pp. 1–8. IEEE Computer Society Press, Amsterdam (2009)
Broekens, J., Pommeranz, A., Wiggers, P., Jonker, C.M.: Factors influencing user motivation for giving preference feedback. In: Fifth Multidisciplinary Workshop on Advances in Preference Handling in Conjunction with ECAI 2010, Ulrich Junker, Jérôme Lang and Patrice Perny, pp. 19–24. Lisbon, Portugal, (2010a)
Broekens, J., Pronker, A., Neuteboom, M.: Real time labeling of affect in music using the affectbutton. In: Proceedings of the 3rd international workshop on Affective interaction in natural environments, Ginevra Castellano, Kostas Karpouzis, Jean-Claude Martin, Louis-Philippe Morency, Laurel Riek and Christopher Peters, pp. 21–26. ACM, New York, AFFINE '10 (2010b)
Burke, R: Knowledge-based recommender systems. In: Kent, A. (ed.) Encyclopedia of Library and Information Systems, vol 69, Marcel Dekker, New York (2000)
Burke, R.: Hybrid recommender systems: Survey and experiments. User Model User-Adapt Interact **12**(4), 331–370 (2002)
Burke, R.D., Hammond, K.J., Young, B.C.: Knowledge-based navigation of complex information spaces. In: 13th National Conference on Artificial Intelligence, pp. 462–468. AAAI Press, Portland, (1996)
Carenini, G., Loyd, J.: Valuecharts: analyzing linear models expressing preferences and evaluations. In: AVI '04: Proceedings of the working conference on Advanced visual interfaces, pp. 150–157. ACM, New York (2004)
Carenini, G., Poole, D.: Constructed preferences and value-focused thinking: Implications for ai research on preference elicitation. In: AAAI-02 Workshop on Preferences in AI and CP: symbolic approaches, pp. 1–10. AAAI, Edmonton, Canada (2002)
Carenini, G., Smith, J., Poole, D.: Towards more conversational and collaborative recommender systems. In: 8th international conference on intelligent user interfaces, pp. 12–18. ACM, Miami (2003)
Chen, L., Pu, P.: Survey of preference elicitation methods. Tech. rep., Swiss Federal Institute of Technology In Lausanne (EPFL), Lausanne (2004)
Chen, L., Pu, P.: Interaction design guidelines on critiquing-based recommender systems. User Model User-Adapt Interact J (UMUAI) **19**(3), 167–206 (2009)
Curhan, J.R., Neale, M.A., Ross, L.: Dynamic valuation: preference changes in the context of face-to-face negotiation. J Exp Soc Psychol **40**(2), 142–151 (2004)
Faltings, B., Pu, P., Torrens, M., Viappiani, P.: Designing example-critiquing interaction. In: IUI '04: Proceedings of the ninth international conference on intelligent user interfaces, pp. 22–29. ACM, New York (2004)
Fano, A., Kurth, S.W.: Personal choice point: helping users visualize what it means to buy a bmw. In: IUI '03: Proceedings of the eighth international conference on intelligent user interfaces, pp. 46–52. ACM, New York (2003)
Fischer, G.W., Carmon, Z., Ariely, D., Zauberman, G.: Goal-based construction of preferences: task goals and the prominence effect. Manag Sci **45**(8), 1057–1075 (1999)
Gregorc, A.: The Mind Styles Model: Theory, Principles, and Practice. AFG, Columbia (2006)
Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst **22**(1), 5–53 (2004)
Hunt, D., Haynes, R., Hanna, S., Smith, K.: Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. J Am Med Assoc **280**(15), 1339–1346 (1998)

Johnson, E., Steffel, M., Goldstein, D.: Making better decisions: from measuring to constructing preferences. Health Psychol **24**(8), 17–22 (2005)
Kay, J: Stereotypes, student models and scrutability. In: Gauthier, G., Frasson, C., VanLehn, K. (eds.) Intelligent Tutoring Systems, Lecture Notes in Computer Science, vol 1839, pp. 19–30. Springer, Berlin (2000)
Keeney, R.: Value-Focused Thinking: A Path to Creative Decision Making. Harvard University Press, Cambridge (1992)
Keeney, R., Raiffa, H.: Decisions with Multiple Objectives: Preferences and Value Trade-Offs. Cambridge University Press, Cambridge (1993)
Knijnenburg, B., Willemsen, M., Ganter, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. User Model User-Adapt Interact 22. (2012). doi:10.1007/s11257-011-9118-4
Kramer, T.: The effect of measurement task transparency on preference construction and evaluations of personalized recommendations. J Market Res **44**(2), 224–233 (2007)
Linden, G., Hanks, S., Lesh, N.: Interactive assessment of user preference models: The automated travel assistant. In: User Modeling: Proceedings of the Sixth International Conference, pp. 67–78. Vienna, Austria (1997)
Liu, F.: Changing for the better: preference dynamics and agent diversity. PhD thesis, University of Amsterdam, Amsterdam (2008)
McCarthy, K.J., Reilly, K., McGinty, L., Smyth, B.: Experiments in dynamic critiquing. In: Proceedings of the 10th International Conference on Intelligent User Interfaces, pp. 175–182. ACM, San Diego (2005)
McFadden, D.: Rationality for economists?. J Risk Uncert **19**(1), 73–105 (1999)
Mehrabian, A.: Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies. Cambridge (1980)
Miller, B.N., Albert, I., Lam, S.K., Konstan, J.A., Riedl, J.: Movielens unplugged: experiences with an occasionally connected recommender system. In: Proceedings of the eighth international conference on intelligent user interfaces, pp. 263–266. ACM, New York, IUI '03 (2003)
Mooney, R., Roy, L.: Content-based book recommending using learning for text categorizations. In: Proceedings of the Fifth ACM Conference on Digital Libraries, pp. 195–204. ACM, San Antonio (2000)
Ono, C., Kurokawa, M., Motomura, Y., Asoh, H.: A context-aware movie preference model using a bayesian network for recommendation and promotion. In: Conati, C., McCoy, K., Paliouras, G. (eds.) User Modeling 2007, Lecture Notes in Computer Science, vol 4511, pp. 247–257. Springer, Berlin (2007)
Payne, J.W., Bettman, J.R., Schkade, D.A.: Measuring constructed preferences: towards a building code. J Risk Uncert **19**(1–3), 243–270 (1999)
Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The adaptive web: methods and strategies of web personalization, pp. 325–341. Springer-Verlag, Berlin (2007)
Peintner, B., Viappiani, P., Yorke-Smith, N.: Preferences in interactive systems: technical challenges and case studies. AI Mag **29**(4), 13–24 (2008)
Pommeranz, A., Broekens, J., Visser, W., Brinkman, W.P., Wiggers, P., Jonker, C.: Multi-angle view on preference elicitation for negotiation support systems. In: Brinkman, W.P., Hindriks, K. (eds.) Proceedings of First International Working Conference on Human Factors and Computational Models in Negotiation (HuCom08), pp. 19–26. Delft University of Technology, Mediamatica, Delft (2008)
Pommeranz, A., Wiggers, P., Jonker, C.: User-centered design of preference elicitation interfaces for decision support. In: Leitner G, Hitz M, Holzinger A (eds) USAB2010: HCI in Work & Learning, Life & Leisure, Lecture Notes in Computer Science, vol Vol. 6389, pp. 14–33. Springer, Klagenfurt (2010)
Pu, P., Chen, L.: Trust-inspiring explanation interfaces for recommender systems. Knowledge-Based Systems **20**(6), 542–556 (2007)
Pu, P., Chen, L.: User-involved preference elicitation for product search and recommender systems. AI Mag **29**(4), 93–103 (2008)
Pu, P., Faltings, B., Torrens, M.: User-involved preference elicitation. In: Workshop notes of the Workshop on Configuration, the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03), pp. 56–63. Menlo Park (2003)
Pu, P., Chen, L., Hu, R.: Evaluating recommender systems from the user's perspective: Survey of the state of the art. User Modeling and User-Adap Inter 22. (2012). doi:10.1007/s11257-011-9115-7
Resnick, P., Iacovou, N., Sushak, M., Bergstrom, P., Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. In: Smith, J.B., Smith, F.D., Malone, T.W. (eds.) ACM Conference

on Computer Supported Collaborative Work Conference, Association of Computing Machinery, pp. 175–186. ACM Press, Chapel Hill (1994)
Shearin, S., Lieberman, H.: Intelligent profiling by example. In: Sidner, C., Moore, J. (eds.) IUI '01: Proceedings of the Sixth International Conference on Intelligent User Interfaces, pp. 145–151. ACM, New York (2001)
Shiv, B., Fedorikhin, A.: Heart and mind in conflict: the interplay of affect and cognition in consumer decision making. J Consum Res **26**(3), 278–292 (1999)
Simon, D., Krawczyk, D.C., Holyoak, K.J.: Construction of preferences by constraint satisfaction. Psychol Sci **15**(5), 331–336 (2004)
Sinha, R., Swearingen, K.: The role of transparency in recommender systems. In: Wixon, D. (ed.) CHI '02 Extended Abstracts on Human Factors in Computing Systems, pp. 830–831. ACM, New York (2002)
Smyth, B., McGinty, L.: The power of suggestion. In: Gottlob, G., Walsh, T. (eds.) IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pp. 127–132. Morgan Kauffman, Acapulco (2003)
Stolze, M., Ströbel, M.: Dealing with learning in ecommerce product navigation and decision support: the teaching salesman problem. In: Piller, F.T., Reichwald, R., Tseng, M. (eds) In: Proceedings of the Second Interdisciplinary World Congress on Mass Customization and Personalization, TUM, Munchen, Working Paper Series of the Department for General and Industrial Management, TUM Business School, Technische Universität München (2003)
Viappiani, P., Faltings, B., Zuber, V.S., Pu, P.: Stimulating preference expression using suggestions. In: Aha, D.W., Tecuci, G. (eds.) Proceedings of the AAAI Fall Symposium on Mixed-Initiative Problem Solving Assistants (FS-05-07), pp. 128–133. AAAI, Washington, Arlington (2005)
Weber, E.U., Johnson, E.J.: Constructing Preferences from Memory. Cambridge University Press, Cambridge (2006)

Author Biographies

Alina Pommeranz is a Ph.D. candidate in Human Computer Interaction at the Man-Machine Interaction Section at Delft University of Technology. She received her M.Sc. degree in Interactive Systems Engineering from the Royal Institute of Technology, Stockholm, Sweden, in 2008. Her research interests are primarily in the areas of human computer interaction, decision and negotiation support systems, preference elicitation and intelligent user interfaces. Recent interests include value eliciation and reflective decision support. She is currently a visiting researcher at the Value Sensitive Design Lab at the University of Washington.

Joost Broekens is a post-doc researcher at the Man-Machine Interaction Section at Delft University of Technology. He received a M.Sc. degree in Computer Science at the University of Delft, The Netherlands, in 2001 and a Ph.D. in Artificial Intelligence at the University of Leiden, The Netherlands, in 2007. He has published in the area of computational models of emotion (ranging from theoretical to applied approaches), developed master-level courses and course material on the topic, and has given several invited lectures on Affective Computing. His recent interests include reinforcement learning, affective computing, human-robot and human-computer interaction and gaming research.

Pascal Wiggers is currently an assistant professor at the Man-Machine Interaction Section at Delft University of Technology. He received his M.Sc. degree in Computer Science in 2001 and his Ph.D. degree in Context Modeling for Automated Speech Recognition in 2008 from Delft University of Technology. His research interests are in the areas of natural language processing, machine learning techniques, human-computer interaction and multimodal interaction. His research centers around modeling context to improve the automatic recognition and understanding of human language and emotions. His teaching interests include computational intelligence and probabilistic reasoning.

Willem-Paul Brinkman is an assistant professor at the Man-Machine Interaction Section at Delft University of Technology. His main research interest is Mental Health Computing, which focuses on computer support systems for preventing, diagnosing and relieving psychologically based distress or dysfunction and for mental coaching to promote physical health, well-being and the prevention of illness. His work places a strong emphasis on the empirical evaluation of technology usage. He has published work on usability

evaluation methods, adoption of broadband internet, negotiation support systems, affective computing, social interaction and e-health.

Catholijn M. Jonker is full professor and head of the Man-Machine Interaction Section at Delft University of Technology. She received her Ph.D. in Computer Science at Utrecht University, The Netherlands. After a post-doc in Bern, Switzerland, she became assistant (later associate) professor at the Department of Artificial Intelligence of the Vrije Universiteit Amsterdam. Between 2004 and 2006 she was a full professor of Artificial Intelligence/Cognitive Science at the Nijmegen Institute of Cognition and Information of the Radboud University Nijmegen. She chaired the Young Academy of The Royal Netherlands Society of Arts and Sciences in 2005 and 2006. Her research interests include decision support systems, analysis and modeling of cognitive processes, dynamics of behavior of agents and organization dynamics. In 2008 she received a prestigious VICI grant of which the presented work is part of.