The Effect of Preference Elicitation Methods on the User Experience of a Recommender System

Bart P. Knijnenburg
Eindhoven University of Technology
Human-Technology Interaction group
P.O. Box 513, 5600 MB Eindhoven
B.P.Knijnenburg@tue.nl

Martijn C. Willemsen
Eindhoven University of Technology
Human-Technology Interaction group
P.O. Box 513, 5600 MB Eindhoven
M.C.Willemsen@tue.nl

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.

ACM 978-1-60558-930-5/10/04.

Abstract

To increase the user experience, preference elicitation methods used by recommender systems can be adapted to individual differences such as the level of expertise. However, we will show that the satisfaction and perceived usefulness of a recommender system also depends strongly on subtle variations of the implementation of these methods.

Keywords

Preference elicitation, recommender systems, user experience, satisfaction, usefulness, understandability

ACM Classification Keywords

H.1.2. Models and principles: User/Machine Systems–software psychology; H.4.2. Information Systems Applications: Types of Systems–decision support; H.5.2 Information Interfaces and Presentation: User Interfaces–evaluation/methodology, interaction styles, user-centered design

General Terms

Measurement, Design, Experimentation, Human Factors

Introduction

Recommender systems are often evaluated in terms of recommendation accuracy [8]. The user experience of recommender system may however also significantly be influenced by its interface. Specifically, for these systems, the preference elicitation (PE) method is important [17]. Preference elicitation is the process in which the system discovers what kinds of items the user does and does not like. Since users with little domain knowledge (novices) have less stable preferences [6], it might be important to adapt the PE method to the expertise of the user.

The recommender system we study helps users to save energy by recommending energy-saving measures using Multi-Attribute Utility Theory (MAUT). Utilities are calculated for each item by multiplying the values of each of its attributes with the user's weight of that attribute [4]. The items with the highest utility are recommended. In this case, preference elicitation is the discovery of the user's attribute weights. A first study, reported in [7], using the first generation of the system (Gen1, see Figures 1a & 1b) compared two PE methods: an attribute-based PE method let users explicitly assign attribute weights [5,9], while a case-based PE method let the users evaluate entire choice options [10,11,15,16] Novices were more satisfied with the case-based PE method and also found this method more useful than the attribute-based PE method whereas the reverse was true for experts.

We realized that our PE methods had several potential usability problems and decided to conduct an additional experiment (Gen2) with improved PE methods. The original Gen1 attribute-based PE method let the user increase or decrease the *importance* of each attribute, which could have caused confusion for negatively phrased attributes. E.g. *increasing* the importance of "continuous effort" actually showed energy saving measures with *lower* effort levels. The Gen2 attribute-based PE method therefore explicitly showed the direction of the effect: "continuous effort" ("moeite continu") was replaced by "low continuous effort" ("weinig continue moeite"; Figure 1c). The Gen1 case-based PE method was very cluttered because it showed all the attribute values of the selected examples. Such information overload may have confused novice users [1]. The Gen2 case-based PE method therefore only showed the names of the examples (Figure 1d). Finally, both Gen2 PE methods included "double" increase and decrease buttons.

In this paper, we compare the effect of the four PE methods in Gen1 and Gen2 on the users' satisfaction and perceived usefulness.

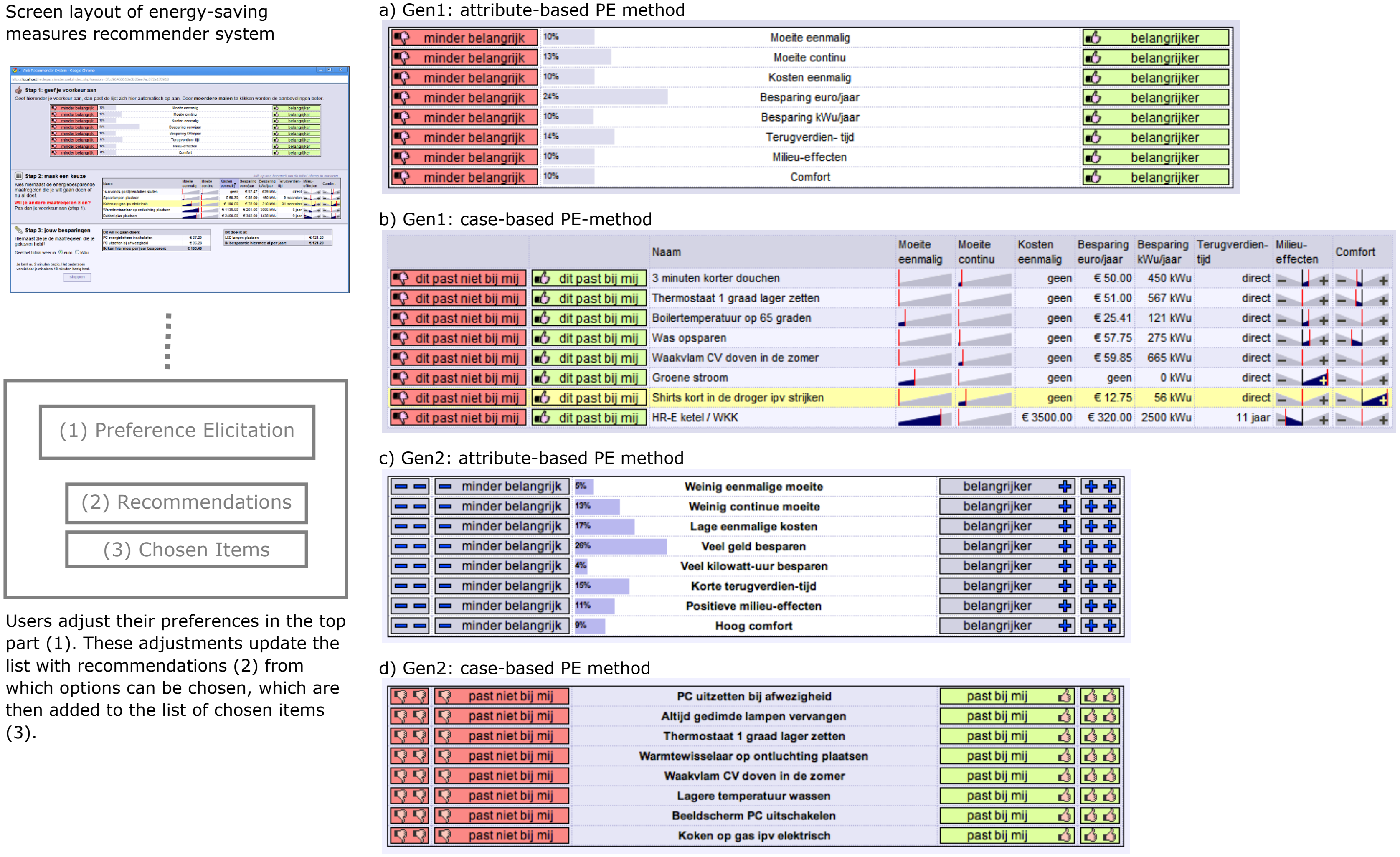
Method

System

The system used in both experiments included a wide variety of 80 energy-saving measures, each defined on 9 different attributes. Note that the Gen2 system had optimized server response times, and gave users a top-8 of recommendations instead of a top-5 in Gen1.

Participants

Participants were recruited via Internet forums and two local newspapers. They were asked to participate to "help make further improvements to the system". In Gen1 users could receive a financial reward, in Gen2 they were offered a printable version of their selected measures after the experiment. 219 participants finished the experiment (89 in Gen1, 130 in Gen2).



Screen layout of energy-saving measures recommender system

Users adjust their preferences in the top part (1). These adjustments update the list with recommendations (2) from which options can be chosen, which are then added to the list of chosen items (3).

a) Gen1: attribute-based PE method

minder belangrijk	10%	Moeite eenmalig	belangrijker
minder belangrijk	13%	Moeite continu	belangrijker
minder belangrijk	10%	Kosten eenmalig	belangrijker
minder belangrijk	24%	Besparing euro/jaar	belangrijker
minder belangrijk	10%	Besparing kWu/jaar	belangrijker
minder belangrijk	14%	Terugverdien- tijd	belangrijker
minder belangrijk	10%	Milieu-effecten	belangrijker
minder belangrijk	10%	Comfort	belangrijker

b) Gen1: case-based PE-method

		Naam	Moeite eenmalig	Moeite continu	Kosten eenmalig	Besparing euro/jaar	Besparing kWu/jaar	Terugverdien- tijd	Milieu- effecten	Comfort
dit past niet bij mij	dit past bij mij	3 minuten korter douchen			geen	€ 50.00	450 kWu	direct	– +	– +
dit past niet bij mij	dit past bij mij	Thermostaat 1 graad lager zetten			geen	€ 51.00	567 kWu	direct	– +	– +
dit past niet bij mij	dit past bij mij	Boilertemperatuur op 65 graden			geen	€ 25.41	121 kWu	direct	– +	– +
dit past niet bij mij	dit past bij mij	Was opsparen			geen	€ 57.75	275 kWu	direct	– +	– +
dit past niet bij mij	dit past bij mij	Waakvlam CV doven in de zomer			geen	€ 59.85	665 kWu	direct	– +	– +
dit past niet bij mij	dit past bij mij	Groene stroom			geen	geen	0 kWu	direct	– +	– +
dit past niet bij mij	dit past bij mij	Shirts kort in de droger ipv strijken			geen	€ 12.75	56 kWu	direct	– +	– +
dit past niet bij mij	dit past bij mij	HR-E ketel / WKK			€ 3500.00	€ 320.00	2500 kWu	11 jaar	– +	– +

c) Gen2: attribute-based PE method

– –	– minder belangrijk	5%	Weinig eenmalige moeite	belangrijker +	+ +
– –	– minder belangrijk	13%	Weinig continue moeite	belangrijker +	+ +
– –	– minder belangrijk	17%	Lage eenmalige kosten	belangrijker +	+ +
– –	– minder belangrijk	20%	Veel geld besparen	belangrijker +	+ +
– –	– minder belangrijk	4%	Veel kilowatt-uur besparen	belangrijker +	+ +
– –	– minder belangrijk	15%	Korte terugverdien-tijd	belangrijker +	+ +
– –	– minder belangrijk	11%	Positieve milieu-effecten	belangrijker +	+ +
– –	– minder belangrijk	9%	Hoog comfort	belangrijker +	+ +

d) Gen2: case-based PE method

past niet bij mij	PC uitzetten bij afwezigheid	past bij mij	
past niet bij mij	Altijd gedimde lampen vervangen	past bij mij	
past niet bij mij	Thermostaat 1 graad lager zetten	past bij mij	
past niet bij mij	Warmtewisselaar op ontluchting plaatsen	past bij mij	
past niet bij mij	Waakvlam CV doven in de zomer	past bij mij	
past niet bij mij	Lagere temperatuur wassen	past bij mij	
past niet bij mij	Beeldscherm PC uitschakelen	past bij mij	
past niet bij mij	Koken op gas ipv elektrisch	past bij mij	

Figure 1. The attribute-based and case-based preference elicitation methods used in the Gen1 and Gen2 versions of the interface. The callout shows where the PE method and other parts of the interface reside in the system.

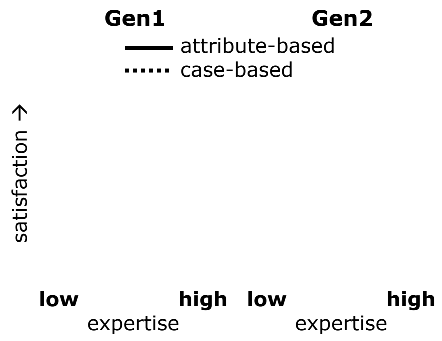


Figure 2. The satisfaction with the system, for both generations of each PE method.

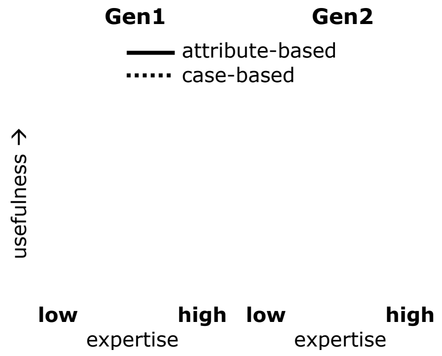


Figure 3. The perceived usefulness of the system, for both generations of each PE method.

Procedure

After several pre-experimental questionnaires and an explanation of the system, participants were instructed that the goal of the interaction was to "find new saving measures that match your preference and at the same time catalogue saving measures that you are doing already." They were then randomly assigned to one of the PE methods and routed to the actual system. In Gen1, participants were required to use the system for at least 10 minutes. In Gen2, interactions that lasted less than 3 minutes were excluded. Finally, participants were given several post-experimental questionnaires.

Questionnaires

Before interaction with the system, four five-point scale questions were asked to measure expertise. The items were summed and centered to obtain a single expertise measure (Chronbach's α = .813). Between generations there was no significant difference in expertise.

After interaction with the system, satisfaction with the system was measured using the five general items of the QUIS[1]. The nine-point scaled items were summed to obtain a single satisfaction score (Chronbach's α = .851). The post-experimental questionnaires also included nineteen five-point scale questions covering other aspects related to satisfaction. These questions were entered in an exploratory factor analysis, using Maximum Likelihood extraction and Oblimin rotation (δ = -.5). Three factors were extracted that together explained 50.4% of the variance: 'perceived usefulness of the system', 'understandability of the interaction' and 'satisfaction with the chosen measures'.

Results

We pooled the data of both experiments and conducted regression analyses to predict our dependent variables (satisfaction with the system, perceived usefulness of the system, understandability of the PE method, and satisfaction with the chosen energy-saving measures) using expertise, generation and PE method as independent variables.

Figure 2 presents the satisfaction of the users of both generations of each PE method. The Gen1 graph shows the effect reported earlier [6]: For the case-based system, satisfaction decreases with expertise (i.e. novices are more satisfied with the case-based PE method than experts), while for the attribute-based PE method satisfaction increases with expertise. Interestingly, this effect reverses in Gen2: the attribute-based PE method is more satisfactory for novices, while the case-based PE method is more satisfactory for experts. This reversal of the effect tested significantly as a three-way interaction between generation, PE method and expertise (B = .381, SE = .135, t(212) = 2.82, p < .01).

Figure 3 presents the perceived usefulness of both generations of each PE method. We observe a similar reversal of our previously found effect: In contrast to Gen1, the Gen2 attribute-based PE method is more useful for novices, while the Gen2 case-based PE method is more useful for experts. This three-way interaction is again significant (B = .038, SE = .016, t(212) = 2.36, p < .05). Furthermore, the usefulness is lower for users with high expertise (B = .138, SE = .065, t(212) = 2.14, p < .05). This is in line with earlier work [12,14].

[1] See http://hcibib.org/perlman/question.cgi?form=QUIS. We excluded item 4, because it raised questions during pretesting.

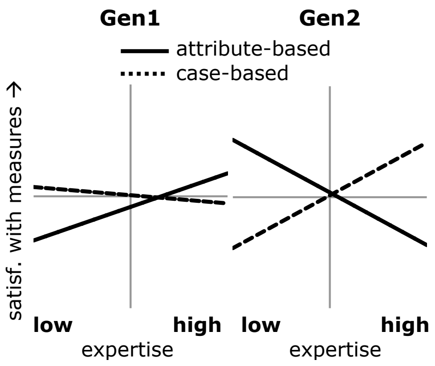


Figure 4. The satisfaction with the chosen measures, for each of the two PE methods in the two experiments.

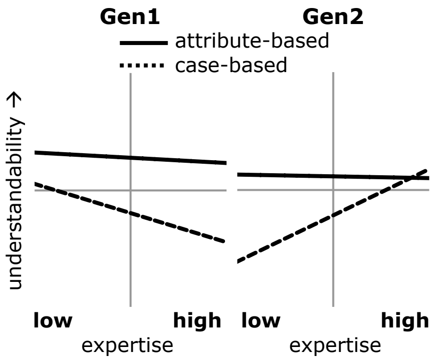


Figure 5. The understandability of the two PE methods in the two experiments.

Figure 4 presents the satisfaction of our participants with their chosen measures, for both generations of each PE method. We again find the same reversal in effects, by means of the three-way interaction (B = .041, SE = .015, t = 2.68, p < .01).

Figure 5 presents the understandability of both generations of each PE method. A main effect of PE method shows that the case-based PE method is significantly less understandable than the attribute-based PE method (B = .200, SE = .065, t = 3.09, p < .005). This effect seems to be driven mostly by the reduced understandability of the case-based PE method for experts in Gen1, and for novices in Gen2. This three-way interaction is however not significant.

Conclusion

Most importantly, it seems that conceptually different PE methods as well as subtle variations on a given method significantly influence the user experience of our system. Whereas important differences between experts and novices seem to exist, there is no general PE method that is better for experts or novices: the best method for each type of user strongly depends on the specific variation as well.

Removing the ambiguity from the Gen1 attribute-based PE method makes this method more satisfying and useful for novices: compared to experts, they may have been more prone to misinterpret the ambiguous attribute directions of this version.

The Gen2 case-based PE-method is less satisfying for novices than the Gen2 attribute-based PE method, which may be explained by the observed reduced understandability for novices between the two generations. Novice users probably not only used the names of the energy-saving measures to evaluate the exemplary cases, but also their attributes. In Gen2, without the attributes, it seems harder to make case-based trade-offs. This is an interesting finding in the light of decision-making research stating that novices are sensitive to information overload [1]: apparently too little information may have a detrimental effect on their understanding as well.

Future Work

As the subtle variations of our PE methods seem to have significant effects on the user experience, future work should primarily focus on investigating the effect of such variations on novice and expert user experience in appropriate detail.

For example, our Gen2 attribute-based PE method resembles a needs-based approach. Other researchers have used more sophisticated needs-based PE methods in which needs are a linear combination of the attributes, effectively reducing the number of preference dimensions. Such a needs based approach seems to be effective [2,3,13], but the question remains how many and what kind of dimensions would be most suitable for experts and novices.

The two case-based versions show a trade-off between screen clutter and understandability: the more information given, the more understandable the PE method, but the more cluttered the screen. The optimal balance in this trade-off could again be different for experts and novices. Several versions of the case-based system with different levels of information granularity could be tested against each other.

Acknowledgements

We thank Benedict Dellaert for feedback during our research, Steven Langerwerf for help implementing our system, and Evelien Matthijssen for help with the energy-saving use-case. We gratefully acknowledge the partial funding of our work through the EC FP7 project MyMedia under the grant agreement no. 215006.

Citations

[1] Alba, J.W. and Hutchinson, J.W. Dimensions of consumer expertise. *Journal of Consumer Research 13*, 4 (1987), 411-454.

[2] Chen, L. Survey of Preference Elicitation Methods. Unpublished (2004). http://hci.epfl.ch/members/lichen/IC_TECH_REPORT_200467.pdf.

[3] Felix, D., Niederberger, C., Steiger, P., and Stolze, M. Feature- oriented vs. needs-oriented product access for non-expert online shoppers. In Towards the E-Society: E-Commerce, E- Business, and E-Government, Springer (2001), 399-406.

[4] Guttman, R.H. and Maes, P. Agent-mediated integrative negotiation for retail electronic commerce. In *Proc. AMET-98*, Springer (1998), 70-90.

[5] Haubl, G. and Trifts, V. Consumer decision making in online shopping environments: the effects of interactive decision aids. *Marketing Science 19*, 1 (2000), 4-21.

[6] Hoeffler, S. and Ariely, D. Constructing stable preferences: a look into dimensions of experience and their impact on preference stability. *Journal of Consumer Psychology 8*, 2 (2002), 113-139.

[7] Knijnenburg, B.P. and Willemsen, M.C. Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. *Proc. RecSys 2009*, ACM Press (2009), 381-384.

[8] McNee, S., Riedl, J., and Konstan, J. Being accurate is not enough: how accuracy metrics have hurt recommender systems. *Ext. Abstracts CHI 2006*, ACM Press (2006), 1097-1101.

[9] Olson, E.L. and Widing II, R.E. Are interactive decision aids better than passive decision aids? *Journal of Interactive Marketing 16*, 2 (2002), 22-33.

[10] Pu, P. and Chen, L. Integrating tradeoff support in product search tools for e-commerce sites. In *Proc. EC'05*, ACM Press (2005), 269-278.

[11] Pu, P. and Kumar, P. Evaluating example-based search tools. In *Proc. EC'04*, ACM Press (2004), 208-217.

[12] Spiekermann, S. Online Information Search with Electronic Agents: Drivers, Impediments, and Privacy Issues. Unpublished (2001). http://dissertationen.hu-berlin.de/dissertationen/spiekermann-sarah-2001-11-22/PDF/Spiekermann.pdf.

[13] Stolze, M. and Nart, F. Well-integrated needs-oriented recommender components regarded as helpful. *Ext. Abstracts CHI 2004*, ACM Press (2004), 1571-1571.

[14] Urban, G.L., Sultan, F., and Qualls, W. Design and evaluation of a trust based advisor on the internet. Unpublished (1999). http://ebiz.mit.edu/research/papers/123 Urban, Trust Based Advisor.pdf.

[15] Viappiani, P., Faltings, B., and Pu, P. Preference-based search using example-critiquing with suggestions. *Journal of Artificial Intelligence Research 27*, (2006), 465-503.

[16] Viappiani, P, Pu, P., and Faltings, B. Conversational recommenders with adaptive suggestions. In *Proc. RecSys 2007*, ACM Press (2007), 89-96.

[17] Xiao, B. and Benbasat, I. E-commerce product recommendation agents: use, characteristics, and impact. *MIS Quarterly 31*, 1 (2007), 137-209.