Understanding the Utility of Rationale in a Mixed-Initiative System for GUI Customization

Andrea Bunt, Joanna McGrenere, Cristina Conati

Computer Science Department, University of British Columbia
2366 Main Mall, Vancouver, BC, V6T 1Z4
{bunt, joanna, conati}@cs.ubc.ca

Abstract. In this paper, we investigate the utility of providing users with the system's rationale in a mixed-initiative system for GUI customization. An evaluation comparing a version of the system with and without the rationale suggested that rationale is wanted by many users, leading to increased trust, understandability and predictability, but that not all users want or need the information.

1 Introduction

In recent years, substantial research efforts have been dedicated to finding ways to provide users with customized graphical user interfaces (GUIs) as a means of coping with the problem of increasing GUI complexity (e.g., [5, 14]). Solutions can be divided into three categories: i) *adaptive*: the system customizes the interface (e.g., [6]), ii) *adaptable*: the user customizes the interface (e.g. [14]), or iii) *mixed-initiative [10]*: the system and user cooperate to customize the interface through a combination of automation and direct manipulation (e.g., [2], [5, 15]). In combining aspects of adaptive and adaptable interfaces, mixed-initiative approaches address a number of their common disadvantages. In particular, by automatically generating customization recommendations, a mixed-initiative approach addresses concerns with adaptable interfaces related to the fact that they require additional user effort [13] and that not all users make good customization decisions [1]. By letting users make the final decision on when and how to customize, the mixed-initiative approach addresses one of the main drawbacks in purely adaptive approaches – lack of user control [9].

With a mixed-initiative approach, however, if users don't understand why and how the customization suggestions are generated, two potential disadvantages of adaptive interfaces remain: 1) lack of transparency, and 2) lack of predictability [9]. In this paper we explore whether both issues can be partially addressed by providing the user with access to the rationale underlying the customization suggestions. We investigate this concept within the MICA (Mixed-Initiative Customization Assistance) system, which provides support for GUI customization in Microsoft Word (MSWord) [2]. One of MICA's distinguishing traits is that its customization recommendations rely on a formal assessment of the performance savings, based on information on user expertise, task, and interface layout. MICA also includes an interface mechanism to explain its decision-making process to the user. A previous evaluation provided evidence that MICA's suggestions have a positive impact on task performance and that its mixed-initiative support is preferred to the purely-adaptable alternative [2]. The main contribution of this paper is a formal evaluation of MICA's rationale, which provides insight into the qualitative impact of including rationale within a mixed-initiative system for GUI customization.

There are numerous examples of adaptive or mixed-initiative systems that provide access to all or part of their rationale (e.g., [4, 16]), but none that do so in the context of GUI customization. For example, inspectable student models allow users to view and sometimes edit their student model, which in turn gives them a sense of what causes the particular adaptive behaviour to occur (e.g., [4, 16]). Provision of rationale has also been explored in recommender systems (e.g., [8]) and in expert systems (e.g., see [11]). Evaluations provide encouraging evidence that the rationale can increase transparency [16], promote reflection [16], and improve users' reactions to system recommendations [8]. If not properly designed, however, rationale can be difficult to use [4], and can even lead to less favourable responses towards the system [8].

Since, to our knowledge, there has been no work investigating rationale utility within mixed-initiative GUI customization systems, little is known about what information to include in the rationale, whether users want access to it, or how it will affect users' impressions of the system. We show that providing access to system rationale in this context has the potential to be beneficial for many users, but that impressions of its utility vary widely from user to user.

2 The MICA System

We begin by outlining MICA's mixed-initiative customization support. A more complete description can be found in [2].

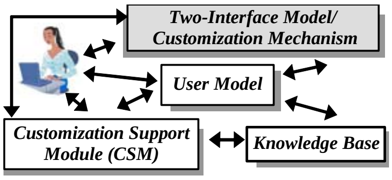

Two-Interface Model/ Customization Mechanism
User Model
Customization Support Module (CSM)
Knowledge Base

Fig. 1. MICA's architecture

MICA, whose architecture is depicted in Fig. 1, helps users customize within a two-interface version for Microsoft Word (MSWord) [14]. The two interfaces are: 1) the Full Interface (FI), which is the default full MSWord interface (Fig. 2, right), and 2) the Personal Interface (PI), a feature-reduced version, containing only features that the user has chosen to add (Fig. 2, left). A toggle button (circled in Fig. 2) allows the user to switch between interfaces.

MICA tries to identify the user's optimal PI by evaluating which features should be included in the PI and which should reside solely in the FI. The Customization Support Module (CSM) is responsible for determining this optimal PI and generating

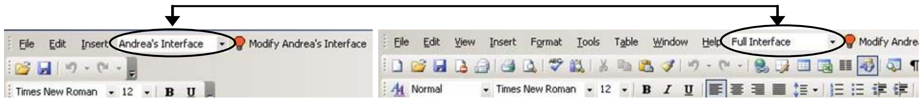
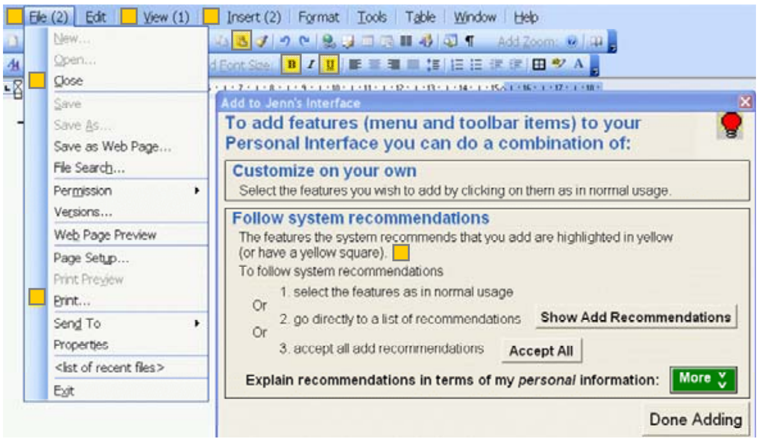


Fig. 2. The two-interface model. The PI is only the left; the FI on the right.



Add to Jenn's Interface

To add features (menu and toolbar items) to your Personal Interface you can do a combination of:

Customize on your own

Select the features you wish to add by clicking on them as in normal usage.

Follow system recommendations

The features the system recommends that you add are highlighted in yellow (or have a yellow square).

To follow system recommendations

1. select the features as in normal usage

Or

2. go directly to a list of recommendations Show Add Recommendations

Or

3. accept all add recommendations Accept All

Explain recommendations in terms of my *personal* information: More

Done Adding

Fig. 3. MICA's customization interface.

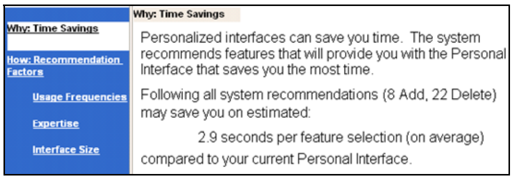
corresponding customization suggestions. To do so, the CSM relies on the User Model, which assesses the user's time performance given a particular PI. This assessment is done in cooperation with the Knowledge Base using a novel extension of a cognitive modelling technique known as GOMS analysis [3]. The performance assessment relies primarily on three factors. 1) *Expected Usages:* how often the user is expected to access each feature. 2) *Expertise:* the amount of time the user takes to locate each feature in the interface (users with lower expertise are likely to be more negatively impacted by excess functionality [1]). 3) *Interface Characteristics:* detailed layout information on the FI and the PI currently under consideration, including the number of features present and where they are located.

It should be noted that Expected Usages and Expertise are not yet assessed on-line. Although there are techniques that could guide both types of assessments (e.g., [7],[12]), we felt that giving priority to investigating rationale utility would provide the most benefit to GUI customization research, despite a user model with some "black box" components.

Fig. 3 shows MICA's mixed-initiative customization interface for adding features to the PI (a direct extension of the adaptable mechanism proposed in [14]). The central part of Fig. 3 shows the dialogue box that pops up when the user initiates customization. MICA's recommended additions are made visually distinct within the menus and toolbars (by yellow highlighting or a yellow square). Users can accept the recommendations using any combination of three methods: 1) selecting features as in normal usage, 2) selecting from a list accessible through the "Show Add Recommendations" button, and 3) using the "Accept All" button.

3 MICA's Rationale

MICA's rationale component describes why the system is making recommendations and the relevant user- and interface-specific factors impacting its decision-making process. Presenting this rationale has the potential to provide valuable insight into



Why: Time Savings

Personalized interfaces can save you time. The system recommends features that will provide you with the Personal Interface that saves you the most time.

Following all system recommendations (8 Add, 22 Delete) may save you on estimated:

2.9 seconds per feature selection (on average)

compared to your current Personal Interface.

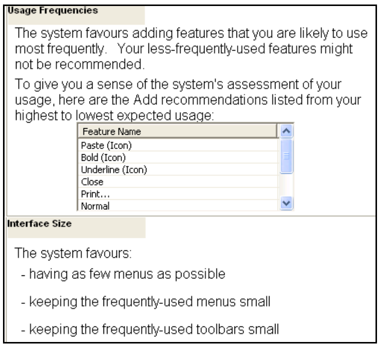
Fig. 4. The "Why" component of the rationale.

how the system works; however, effectively communicating the information to the average user is a challenging design task, particularly since MICA's algorithm is relatively complex.

Since this is the first attempt at providing rationale in GUI customization research, we undertook an iterative design and evaluation process. During the evaluation of a previous version of MICA [2], none of the study participants accessed the rationale spontaneously because of both usability and study methodology issues. We were, however, able to gather usability feedback by asking participants to view the rationale during post-session interviews, which we used to redesign the interface. Next, we pilot tested the new design with eight computer science graduate students. The pilot evaluation consisted of 30-minute interviews targeting issues such as: 1) wording clarity, 2) missing/unnecessary information, and 3) whether it was clear where to access, and how to navigate through the rationale. The pilot testing led to a number of improvements to the interface. One worth mentioning here is stressing that the rationale contains personalized information as opposed to canned text, because the pilot participants found this information most compelling.

In the final interface resulting from the aforementioned iterative design process, users can access the rationale by clicking on the "More" button next to the line "Explain recommendations in terms of my personal information" (Fig. 3, bottom). Once clicked, the dialogue box in Fig. 3 expands to include information on why and how the system makes recommendations. The "Why" component, displayed in Fig. 4, indicates that the recommendations are based on time savings and provides an estimated savings per feature invocation (based on the User Model's performance assessment) should the user choose to accept all recommendations.

The "How" component is a simplified explanation of MICA's decision-making. The first screen, "How: Recommendations Factors," explains that MICA balances the three factors described in Section 2, with their names altered based on pilot feedback: 1) Usage Frequencies (i.e., Expected Usages), 2) Expertise, and 3) Interface Size (i.e., Interface Characteristics). Next, three screens describe each factor in greater detail (two are in Fig. 5). The Usage Frequencies and Expertise screens also display the recommendations ranked according to the User Model assessments for that factor (e.g., Fig. 5, top).



Usage Frequencies

The system favours adding features that you are likely to use most frequently. Your less-frequently-used features might not be recommended.

To give you a sense of the system's assessment of your usage, here are the Add recommendations listed from your highest to lowest expected usage:

Feature Name
Paste (Icon)
Bold (Icon)
Underline (Icon)
Close
Print...
Normal

Interface Size

The system favours:
- having as few menus as possible
- keeping the frequently-used menus small
- keeping the frequently-used toolbars small

Fig. 5. Usage Frequencies and Interface Size factors, excluding the navigation bar (see Fig. 4).

4 Evaluation

To understand the utility of the rationale, we compared two versions of the MICA system: one with and one without the rationale. The goal was to better understand the qualitative impact of the rationale on users' attitudes toward the system. Based on user feedback from the previous informal studies, we anticipated large individual differences along this dimension. We did not, however, expect rationale to significantly impact customization decisions, since in our earlier study, participants already followed most of the system's recommendations (96%) for additions to the PI without accessing the rationale [2]. While there would be room for improvement in terms of following recommendations for deletion, previous work has shown users to be very reluctant to do so [2], [14]. Therefore, we anticipated the most interesting findings to come from the qualitative data on user attitudes and preferences. Based on previous feedback, we expected that some users would appreciate the rationale and find it useful, while others would find it unnecessary. We wanted to better understand the reasons underlying different reactions, and the qualitative advantages and disadvantages of providing access to rationale information.

4.1 Method

Sixteen participants, recruited throughout the University of British Columbia campus, completed the experiment. The experiment was within subjects with two conditions: 1) *Rationale*, the MICA system with the rationale accessible (see Fig. 3) and 2) *No-Rationale*, the system without the rationale. A within-subjects design was chosen to elicit direct comparative statements. To account for carry-over effects, version order (*Rationale* vs. *No-Rationale*) and task order (described below) were counterbalanced.

In this section we briefly describe the experiment methodology, a direct extension of our previous methodology [2]. With this methodology, interacting with the rationale is not an explicit experimental task. Instead, the majority of the session is spent performing pre-assigned word-processing tasks with the target application, MSWord. Alternatively, we could have required users to interact with the rationale for a period of time, for example, by having them complete a worksheet or questionnaire based on information in the rationale (e.g., [4, 16]). We chose to build on our previous methodology, as opposed to designing tasks specific to the rationale, because we felt that it would generate more realistic feedback about when and why users may access the system's rationale.

The experimental procedure was as follows. First, participants completed a detailed questionnaire designed to assess their expertise for each interface feature used in the experiment. The questionnaire results were used to initialize the "Expertise" portion of the User Model since, as discussed earlier, it cannot yet assess expertise on line. Participants then performed two tasks, one with each version of the system (*Rationale* and *No-Rationale*). Prior to each task, the appropriate system version was briefly demonstrated. After finishing the tasks, participants completed a post-questionnaire and were interviewed by the first author using a semi-structured interview format. A session typically lasted 3 hours, but ranged from 2 hours 45 minutes to 4 hours.

We used a *guided task* structure [2], where users were provided with a list of step-by-step instructions and a target final document. The guided tasks served two purposes: 1) they required a large number of menu selections (necessary for customization to be beneficial) while still being of reasonable length, and 2) they provided "Expected Usage" information for the User Model. We further motivated customization through task repetition and a small amount of deception. Each task was actually repeated three times, however, participants were told that the tasks would be repeated up to five times. If at the end of the second task repetition the participant had yet to customize, the experimenter asked her do so at a point of her own choosing during the third repetition. We did so in hopes of achieving a higher customization rate than in our previous experiment, which was 66% without prompting.

Our goal was to give participants as much autonomy as possible with respect to rationale usage; however, we did want participants to look at it. To balance these two objectives, we showed participants where to access the rationale during the initial interface demonstration and requested that the participants "look through the information at some point." Apart from this request, no prompting to look at the rationale was done during the experiment.

4.2 Main Measures

Our emphasis in the evaluation was on qualitative measures. The questionnaire gathered preference information. In particular, participants who viewed the rationale during the study were asked which version of the system they would choose to install (Overall Preference). The post questionnaire also asked participants to state which version they preferred, or whether they found the two equal, for the following five criteria: 1) agreeing with the system recommendations (Agreement); 2) trusting the system to make good recommendations (Trust); 3) understanding why the system was making *specific* recommendations (Specific Understandability); 4) understanding why the system was making recommendations in *general* (General Understandability), and 5) ability to predict future recommendations (Predictability). The interview gathered more detailed qualitative data on topics such as: 1) influence of study methodology on rationale viewing, 2) additional reasons for viewing (or not viewing) the rationale, 3) the impact of the "Why" component on motivation to accept recommendations, and 4) impressions of the utility of the "How" component.

In addition to qualitative measures, we report the time spent viewing the rationale and the percentage of add and delete recommendations followed in both conditions.

4.3 Results

Similar to our last experiment, 69% of participants (11/16) customized in both conditions without any prompting. Once prompted, the remaining 5 participants customized. Since separate analysis of those who were prompted versus those who were not failed to reveal any substantial differences, the remainder of the analysis includes data from all participants.

In the *Rationale* condition, 94% (15/16) of participants accessed the rationale. Of these participants, 47% (7/15) accessed the "Why" component only, with an average viewing time of 15.1 seconds (sd: 9.6 seconds). The remaining 53% (8/15) accessed all of the rationale, with an average viewing time of 63.4 seconds (sd: 30.4 seconds).

To analyze the qualitative data, the interviews were first transcribed. Next, detailed coding was done by the first author, based on thorough analysis of the interviews and questionnaires. We report themes and trends that emerged from this analysis, along with the number of participants whose statements matched the given theme or trend. Our intention was not to prove or disprove hypotheses through statistical analysis, which, given the anticipated diversity of opinions, would have required a much larger number of participants.

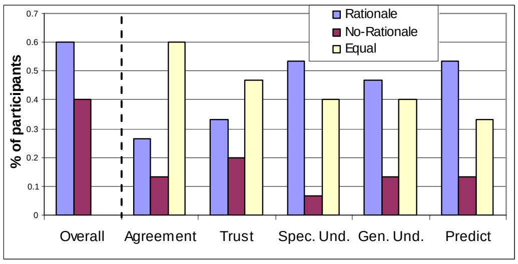


Fig. 6. Preference overall and on individual criteria.

Preference. Fig. 6 depicts the preference data both overall and for each of the individual criteria, and indicates that, in general, the preference data was mixed. When forced to choose, the majority of participants indicated that they would prefer to install the *Rationale* version (60%); however, the *No-Rationale* also had reasonable support (40%). For the individual criteria, participants were given the option of rating the two conditions "Equal." Having the rationale appeared to have the largest impact on both Specific and General Understandability, as well as the Predictability of the recommendations. While the *Rationale* version was preferred by some users for Agreement and Trust, the most popular response for these criteria was "Equal."

Influence of Study Methodology on Rationale Viewing. To understand whether users looked at the rationale solely because of the request during interface demonstration, users were asked i) why they looked at the rationale, and ii) to comment on the role of the request. Out of the 15 users who viewed the rationale, 33% (5/15) said they were not influenced by the request. Another 20% (3/15) indicated that they were partially influenced by the request, but had additional reasons for accessing the rationale. The remaining 47% (7/15) said the request during interface demonstration was their sole reason for accessing the rationale. Just over half of these users (4/7) said that there would be circumstances where they would want the information, but that our particular study methodology did not provide the right motivating conditions. Finally, three users indicated that they had no interest in the rationale. Therefore, 80% (12/15) of the participants either i) viewed the rationale for reasons other than our particular study methodology or ii) could see circumstances outside of the study where they would want to view the rationale.

Additional Reasons for Viewing or Not Viewing the Rationale. Three reasons were given for viewing the rationale by the 53% (8/15) that accessed it for reasons other than the request during interface demonstration. The first was general curiosity (3/8). The second was to have the recommendations explained (3/8), e.g., "*if something is customizing it for you […] I want to have an understanding of why it is doing things.*" The third reason was to have an aspect of the interface explained, such as an explanation of how the PI works (2/8).

The three users who were not interested in the rationale gave unique reasons for why not. One felt that the rationale is unnecessary in a mixed-initiative system, since she could follow the recommendations if she found them useful or customize on her own if she didn't. Another pointed to the fact that the rationale is embedded within a productivity application: "*…when it comes to a program like Microsoft Word most of*

Table 1. Reasons for finding the "How" component useful or not useful.

"How" Useful (10/16)	"How" Not Useful (6/16)
• Gained a better understanding (or confirmed) (5/10) • Recommendations more trustworthy/believable (3/10) • Simple explanation (1/10) • Could use knowledge to become more efficient (1/10)	• Unnecessary or common sense (4/6) • Too technical (1/6) • Didn't influence customization decisions (1/6)

the time you only care about getting the job done. You don't really care about why." The final participant expressed trust in the system: *"I just assume recommendations are because they are useful for you. That's all really I need to know."*

Effectiveness of Rationale: Impact of "Why" on Recommendation Acceptance. Out of those who accessed the rationale, 93% (14/15) indicated that they actually read the "Why" component. Since its purpose is to illustrate the potential time savings that could result from accepting recommendations, we asked users to discuss whether or not this information was, in fact, motivating. Of these users, 43% (6/14) felt that the "Why" component motivated them to accept recommendations. Another 43% (6/14) were generally interested in having a PI that would save time, but were not motivated by the particular amount of time savings listed. They either felt that the amount of time savings was too small, or that its expression was unintuitive: *"I couldn't relate it to the real world. It was like saying how fast you are driving in meters per second..."*

In this study, three users did delete features and did so after having viewed the rationale. For two of the three users, the time savings was a motivating factor.

Effectiveness of Rationale: Usefulness of the "How" Information. Only 47% (7/15) of those who accessed the rationale indicated that they read the "How" component. To obtain as much feedback as possible, during the interview we asked all 16 users to read through the information and comment on its usefulness. After reading the information 62% (10/16) found it useful, including six of the seven users who read the information while customizing, however, 38% did not (6/16). Table 1 summarizes their reasons. The most popular reason for finding the information useful was gaining a better understanding of how the system makes recommendations or confirming their existing understanding. For those who didn't find the information useful, the majority indicated that it was unnecessary or "just common sense."

We also asked users to indicate which pieces of information, if any, were most or least useful. While participants responded favourably to the Expertise and Usage Frequencies factors, 50% (8/16) disliked the Interface Size factor. Many commented correctly that this factor wasn't as personalized, or that having a small interface was the point of customization. Participants didn't seem to understand that MICA balances this factor with both usage frequencies and expertise. This result is consistent with our pilot feedback, which indicated that users respond most favourably to information that is personalized.

Impact on Recommendations Followed. To analyze the impact of the rationale on the percentage of recommendations followed, we ran repeated-measures ANOVA with Version (*Rationale* or *No-Rationale*) as the within-subjects factor. Version Order and Task Order were included as between-subjects controls.

As anticipated, the percentage of Add recommendations followed was similar in both conditions, with participants following 94.2% (sd: 21.6%) of the Add Recommendations in the *Rationale* condition, compared to 93.5% (sd: 9.0%) in the *No-Rationale* condition (F(1, 11) = 0.001, p = 0.982). In terms of Delete recommendations, three users did delete, leading to an average of 14.2% (sd: 34.3%) Delete recommendations followed in the *Rationale* condition compared to 7.2% (sd: 24.8%) in the *No-Rationale*, a difference which was also not statistically significant (F(1, 11) = 0.978, p = 0.334).

Another result of interest was a marginally-significant between-subjects order effect for the percentage of Add recommendations followed (F(1,11) = 3.990, p = 0.071). Participants who completed the *Rationale* condition first followed more Add recommendations overall (average: 99.1%, sd: 2.5%) than those who completed the *Rationale* condition second (average: 87.9%, sd: 14.3%). This order effect was anticipated; we expected knowledge that the system would make principled recommendations in the first condition to transfer to the second. This result suggests that the rationale may be most effective when viewed earlier rather than later and that frequent viewing isn't necessary.

4.3 Discussion

Our findings indicate that the majority of users prefer to have the rationale present, but that a non-insignificant group of users do not need or want the information. For some users, the rationale led to increased trust, understanding, predictability, and motivation to accept recommendations. Some users, however, felt that the rationale was just common sense, or was unnecessary in a mixed-initiative system or productivity application. Others expressed an inherent trust in the system. These findings may suggest that, contrary to previously stated guidelines [9], transparency and predictability may not, in fact, be important to all users in all contexts. However, since some users found the rationale to be just common sense, it may be that our particular design did not always succeed in improving transparency and predictability.

In terms of rationale design, feedback from our iterative design and evaluation process suggests that the personalized aspects of the rationale should be emphasized when possible. In addition, since reactions to the rationale are mixed, the information should clearly visible for those who want it without disrupting those who don't, which was the approach taken here. Finally, we might see different reactions to more lightweight graphical representations of the rationale.

While the rationale was a motivating factor for two of the three users who deleted, and participants who viewed the information in the first condition tended to accept more add recommendations overall, the rationale had limited quantitative impact. Understanding whether the rationale could have a larger quantitative impact may require finding a target application where users are less likely to accept recommendations without the rationale, for reasons such as recommendations being contrary to expectations or a higher cost associated with accepting recommendations. Alternatively, it may be the laboratory environment that led to such high acceptance of recommendations. The rationale may have a larger impact in the field, where users might have lower levels of trust in the system.

5 Summary and Future Work

This paper described the iterative design and the formal evaluation of rationale provision within a mixed-initiative system for GUI customization. Qualitative reactions to having this information varied across individuals. While the evaluation revealed aspects of our rationale that could be improved, the most promising avenue of future research would be to gain a more global understanding of when and why rationale is useful. In particular, we are interested in evaluating how user variability, the target application's complexity, and the division of control between the system and the user affect the qualitative and quantitative utility of a system's rationale.

References

1. Bunt, A., Conati, C., McGrenere, J.: What Role Can Adaptive Support Play in an Adaptable System? In: Proc. of IUI (2004) 117-124
2. Bunt, A., Conati, C., McGrenere, J.: Supporting Interface Customization Using a Mixed-Initiative Approach. In: Proc. of IUI (2007) 92-101
3. Card, S. K., Newell, A., Moran, T. P.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Inc., Mahwah, NJ (1983)
4. Czarkowski, M., Kay, J.: How to Give the User a Sense of Control over the Personalization of Adaptive Hypertext? In: Proc. of Adaptive Hypermedia and Adaptive Web-Based Systems (in conjunction with UM'03) (2003) 121-131
5. Debevc, M., Meyer, B., Donlagic, D., Svecko, R.: Design and Evaluation of an Adaptive Icon Toolbar. User Modeling and User-Adapted Interaction 6(1) (1996) 1-21
6. Gajos, K., Weld, D. S.: Supple: Automatically Generating User Interfaces. In: Proc. of IUI (2004) 93-100
7. Greenberg, S., Witten, I. H.: How Users Repeat Their Actions on Computers: Principles for Design of History Mechanisms. In: Proc. of CHI (1988) 171-178
8. Herlocker, J., Konstan, J. A., Riedl, J.: Explaining Collaborative Filtering Recommendations. In: Proc. of CSCW (2000) 241-250
9. Hook, K.: Steps to Take before Intelligent User Interfaces Become Real. Interacting with Computers 12 (2000) 409-426
10. Horvitz, E.: Principles of Mixed-Initiative User Interfaces. In: Proc. of CHI (1999) 159-166
11. Horvitz, E., Breese, J., Henrion, M.: Decision Theory in Expert Systems and Artificial Intelligence. Journal of Approximate Reasoning 2 (1988) 247-302
12. Horvitz, E., Herckerman, D., Hovel, D., Rommelse, R.: The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In: Proc. of UAI (1998) 256-265
13. Mackay, W. E.: Triggers and Barriers to Customizing Software. In: Proc. of CHI (1991) 153-160
14. McGrenere, J., Baecker, R. M., Booth, K. S.: An Evaluation of a Multiple Interface Design Solution for Bloated Software. In: Proc. of CHI (2002) 163-170
15. Oppermann, R.: Adaptively Supported Adaptability. International Journal of Human-Computer Studies 40 (1994) 455-472
16. Zapata-Rivera, J. D., Greer, J. E.: Interacting with Inspectable Bayesian Student Models. International Journal of AI in Education 14 (2003) 127-163