UMAP 2009

Workshop on

User-Centred Design and Evaluation of Adaptive Systems

June 26th 2009, Trento

Organizers

Stephan Weibelzahl

Judith Masthoff

Alexandros Paramythis

Lex van Velsen

Sixth Workshop on User-Centred Design and Evaluation of Adaptive Systems

Stephan Weibelzahl[1], Judith Masthoff[2], Alexandros Paramythis[3] and
Lex van Velsen[4]

[1]National College of Ireland, sweibelzahl@ncirl.ie
[2]University of Aberdeen, j.masthoff@abdn.ac.uk
[3]Johannes Kepler University Linz, alpar@fim.uni-linz.ac.at
[4]University of Twente, l.s.vanvelsen@utwente.nl

Introduction

The sixth workshop on User-Centred Design and Evaluation of Adaptive systems followed in the tracks of five successful workshops held in conjunction with UM2001, UM2003, AH2004, UM2005 and AH2006. The workshop's guiding perspective is that novel design approaches, adequate evaluation methods, and reliable assessment criteria and metrics are prerequisites for improving the quality and usability of the next generations of adaptive systems. This installment of the workshop had a special focus on the user-centred design of adaptive systems, and early formative evaluation studies that inform and guide the development process. This includes the re-use or tailoring of usability- and requirements- engineering methods to facilitate the design and assessment of concepts and prototypes in all phases of system development.

The workshop was divided into two parts. One part included an introduction delivered by the workshop organizers on the state of the art in formative evaluation methods for adaptive systems, serving both as a mini-tutorial as well as a discussion starter. The second part was devoted to paper presentations and the discussions that result from them.

Thematic Areas

The workshop, in line with the steps of its predecessors, focused on the following general themes:

Design. There is a wide array of user-centred methods that can be used to inform different development stages of adaptive systems. The workshop addressed the question of which ones can be applied best in this context and how. Moreover, it explored extending the value of these methods by looking at ways to account for typical user problems with adaptive systems (e.g., privacy, reduced levels of predictability) in early phases of system development.

Evaluation. With regard to evaluation, one of the workshop's continuing aims is to uncover suitable evaluation methods and approaches for adaptive systems. At a more specific level of interest are the evaluation criteria that can be applied during the evaluation of sub-classes of adaptive systems and their underlying user models.

Experiences, problems and plans. Among the workshop's major goals has been to initiate a discussion among participants about user-centred design and evaluation practices. Towards this end, participants were encouraged to bring in the problems they encountered while employing user-centred activities, or to present the open issues in a design or evaluation approach that has yet to be carried out. There was also ample room for participants to share their insights regarding user-centred design or evaluation.

Presented Papers

The following papers, divided over the general themes, have been presented during the workshop:

Design. Vernero et al. described a study aimed at evaluating different ways to represent and visualize user models, with three different representations and nine visualizations tested over two experiments. Gabrielli and Jameson compiled an overview of the factors that can lead to differences and changes in user's preferences concerning adaptive systems and discuss how these can be accounted for when employing user-centred design activities.

Evaluation. Tarpin-Bernard et al. introduced AnAmeter, an open-source system that enables evaluators to characterize and determine the degree of adaptation in personalized systems. A longitudinal user evaluation of an adaptive meta-search engine was presented by Van Velsen et al., who combined a system- and user-centred approach and demonstrated its merits. Finally, Yudelson and Sosnovsky showed how to utilize previously collected interaction logs in a post-hoc approach to the layered evaluation of alternative user models; this approach was applied to the evaluation of blended modeling of heterogeneous learning activities.

Experiences, problems and plans. In the final category, Tintarev and Masthoff discussed the problems they faced when evaluating the effectiveness of recommender explanations and presented the lessons they learned from overcoming them. To conclude, Santos et al., introduced a plan for using user-centred design methods to enrich a recommendation model to be used in a learning management system.

We would like to thank the members of the workshop's Programme Committee and the presenters for making this another successful installment of the UCDEAS workshop series. Special thanks go to David Chin, Betsy van Dijk, Christina Gena, Eelco Herder, Anthony Jameson, and Gerhard Weber.

Table of Contents

Differences and Changes in Preferences Regarding Personalized Systems: A User-Centred Design Perspective

Silvia Gabrielli and Anthony Jameson[1]

Fondazione Bruno Kessler
Trento, Italy
{sgabrielli, jameson}@fbk.eu

Abstract. We introduce a tentative overview of the factors that can lead to differences and changes in users' preferences concerning user-adaptive systems. We then discuss how current methods in user-centred design could be deployed or adapted to develop personalized systems that better take into account the different preferences of their users, especially concerning the adaptive aspects of the systems, both in the short and long term.

1 Introduction

Personalized Systems (PS) can be defined as interactive artifacts that present a combination of adaptive and adaptable techniques [9][17]. They are systems that "… can alter aspects of their structure, functionality or interface on the basis of a user model generated from implicit and / or explicit user input, in order to accommodate the differing needs of individuals or groups of users and the changing needs of users over time" [1]. The design and evaluation of this type of system can be particularly challenging due to the complexity of generating user models that are rich and flexible enough to do justice to the various needs and preferences of different users, as well as their changes over time. Also, defining adequate methods to evaluate PSs has been recognized in previous studies [8][14][19] to be no trivial task.

In this paper we will propose a theoretical framework which provides a new perspective for the User-Centered Design (UCD) of interactive systems (and PSs in particular), by putting users' preferences in the spotlight, bringing to light some important aspects concerning their differences and change over time that have traditionally received little attention. Our focus will not be primarily on how an adaptive system can recognize and adapt to differences and changes in users' preferences, although this is a valid and commonly studied issue in the area of PS design. Instead, it will be on the analysis of how the new perspective provided can improve the way in which UCD of PS is carried out.

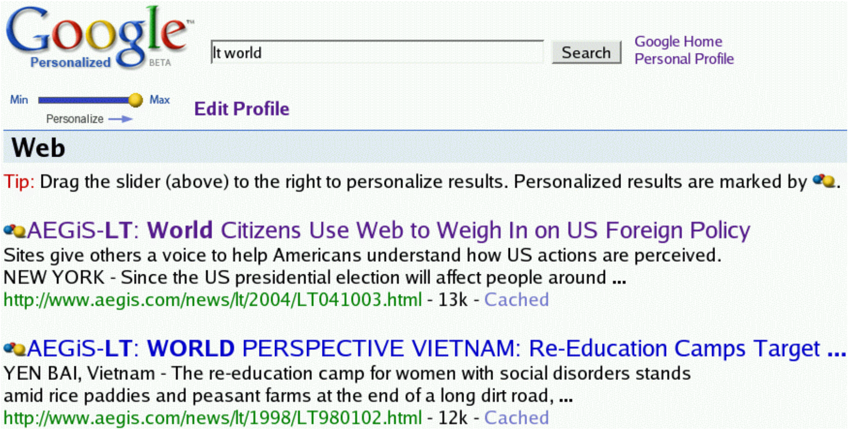
[1] The research described in this paper is being conducted in the targeted research unit Prevolution (code PsychMM), which is funded by the Autonomous Province of Trento.

In particular, the framework provides a tentative overview of factors that help to explain differences in users' preferences and offer some hypotheses about how each factor changes over time, including the rationale underlying each hypothesis. After a brief presentation of the framework in Section 2, we will analyse in Section 3 how designers could apply it to improve the design and evaluation of PS, and which adaptations of the typical UCD methodologies should be made to fully exploit the power of the new perspective presented.

2. A Framework for Understanding and Predicting Preferences

When looking at any particular system or study, we may seem to be able to explain the observed preference differences and changes in terms of one or two fairly obvious factors, such as the users' level of experience or their personality traits. But when we look at a broader range of experience with user-adaptive systems, we see that there is a great variety of factors that can influence user's preferences; a focus on any subset is likely to lead to inaccurate conclusions.

In the framework presented in Table 1 two categories of factors are distinguished that can help to explain differences in users' preferences: *Users' needs and the System's properties*. For concreteness, we will explain the table with reference to two related examples of user-adaptivity: (a) the early profile-based personalized search introduced for a few months in 2005 in Google's "Lab" (see Figure 1); and (b) the now familiar history-based personalized search that is turned on by default in Google's search engine (with which the reader is assumed to be familiar).



lt world

Tip: Drag the slider (above) to the right to personalize results. Personalized results are marked by ●.

AEGiS-LT: World Citizens Use Web to Weigh In on US Foreign Policy
Sites give others a voice to help Americans understand how US actions are perceived.
NEW YORK - Since the US presidential election will affect people around ...
http://www.aegis.com/news/lt/2004/LT041003.html - 13k - Cached

AEGiS-LT: WORLD PERSPECTIVE VIETNAM: Re-Education Camps Target ...
YEN BAI, Vietnam - The re-education camp for women with social disorders stands
amid rice paddies and peasant farms at the end of a long dirt road, ...
http://www.aegis.com/news/lt/1998/LT980102.html - 12k - Cached

Fig. 1. Screenshot of the experimental, profile-based version of Google's personalized search that was used in Google's "Lab" in 2005. (By moving the slider in the upper left, the user could change the degree to which the search results were reranked on the basis of a previously specified interest profile. The reordering was visualized with an animation.)

Table 1. Overview of the factors that can lead to differences and changes in users' preferences concerning user-adaptive systems. (Symbols like "+++" indicate roughly, on a scale from "()" to "+++", the amount of attention that the factor has attracted so far in the literature.)

Factor	Systematic evolution	Reason(s) for evolution
Users' needs		
Needs relevant to success at task performance:		
+++ Tasks to be performed	++ May expand in scope and increase in complexity	Experience with system makes additional tasks manageable
+++ Skills	+++ Will usually increase	Use of system provides relevant practice
+++ Typical usage contexts	++ May expand	Experience with system makes additional contexts manageable
++ Usability priorities	++ May shift away from learnability and become more realistic	Experience diminishes importance of learnability and reveals which priorities are most important
Other needs:		
++ Aesthetic preferences, values, and attitudes (e.g., culturally determined)	+ May shift in favor of system	Habituation to system
++ Relevant personality traits	+ No systematic tendency	Traits are generally stable; any change can be favorable or unfavorable
+ Habits formed with other systems	+ May decline in influence	Experience leads to formation of new habits more favorable to system
+ Desire for novelty	+ Should favor system less, if system was initially novel	What was initially novel does not remain novel
The system's properties		
Variable aspects of the system:		
+++ Aspects dependent on adaptation to and customization by user	+++ May (but may not) make system more suitable for each user	Customization and adaptation have this purpose, but success depends on various factors
+++ Aspects dependent on particular versions of the system	+++ Can become more or less favorable	The system itself tends to improve, but the match with user's knowledge and habits may decline, at least temporarily
Incidental aspects of the system:		
+ Features (e.g., functional limitations) incidentally associated with the adaptive system that evoke different responses in users	+ No systematic tendency	These features can take many different forms, making systematic prediction difficult

2.1 Users' needs

Generally speaking, the most obvious explanation for differences in users' preferences consists in differences in the users' relevant needs. Some of these are directly related to the users' success at performing tasks with the system.

For example, the profile-based search might be expected to work relatively well for, and therefore to be relatively preferred by:

– people with no specific information need who are just looking for information that corresponds to the general interests expressed in their profile (*Tasks to be performed*);

– people who are not very good at expressing their interests with query keywords (*Skills*);

– people who attach high value to being able to understand and control the system's adaptation (*Usability priorities*).

By contrast, the history-based search would be expected to be relatively preferred by users who often search for the same specific pages, who do not want to devote any attention to the matter of personalization, and who do not mind having their search histories stored on Google's computers.

When considering such factors, we can see some predictable changes: A novice web searcher who initially likes the profile-based search may acquire skill in formulating more precise queries and begin performing more ambitious search tasks— changes which, according to our table, may diminish their preference for the profile-based search.

Other types of users' needs are less closely related to performance of specific tasks, that is why they have been generally given less attention in previous design and evaluation studies. For example, the profile-based search, with its colorful icons and appealing animation, should appeal to people who like this type of interface or who more generally enjoy interacting with novel interfaces. But people who are accustomed to leaving search results in the order in which they are received from the search engine may never get used to the sort of active manipulation that is presupposed by the profile-based search (even though this manipulation does not require any particular skill). Here again, some changes may be expected: The unfamiliarity and novelty of the profile-based interface can wear off, for better or for worse; and over time the user may become so accustomed to moving the slider on the profile-based search that they would be frustrated if deprived of the opportunity to do so.

2.2 The system's properties

At first, it might be thought that the properties of the system itself cannot lead to differences or changes in preferences, since the system stays the same for all users. But especially user-adaptive systems are designed to adapt to—and/or be customized by— the user. So at any given moment, different users are likely to be working with variants of the system that are differently well-adapted to them, which in turn can lead to different preferences. For example, User A may be happier with the history-based search then User B simply because User A's web history currently captures A's interests better than User B's history captures B's interests.

Since adaptation and customization are supposed to make the system better for each user over time, the default expectation would be increased acceptance by users over time, providing that they use the system long enough to benefit from the adaptation or customization. But the time course of such changes can be very different for different adaptation techniques: With the history-based personalized search, there tends to be a gradual improvement over time, in addition to possible short-term benefits; with the profile-based search, the benefits may actually be greatest early on and decline over time as the user's initially specified interest profile falls out of date.

There are also a number of interesting aspects of the *process* of preference formation and evolution affecting differences and changes in preferences (they are not addressed here due to lack of space). As an example, even if adaptation is in principle able to bring great benefits for a given user, that user may never notice the fact for different reasons or events related to initial interaction with the system [19].

3 Designing Personalized Systems From the New Perspective

We turn now to consider how the framework presented above could be applied during the design and evaluation of PS, how it could help designers to choose, and adapt when necessary, methods from the vast repertoire of UCD [7][13] to take into account differences and changes in user preferences for these systems. We will address first the research questions raised by the framework with respect to system design (in the phases of user requirements analysis and early design) followed by those more concerned with PS assessment (in the phases of iterative testing and summative evaluation).

3.1 Requirements analysis

How can we help designers to acquire information about users' needs that are not commonly considered in requirements analysis?

If the to-be-designed PS is likely to include some highly novel elements and adaptation mechanisms, there should be a way of determining whether novelty is considered desirable or undesirable by the various subgroups of potential users. Similarly, designers should not overlook the investigation of users' aesthetic preferences, values and attitudes, as well as the relevance of personality traits and previously formed habits when conducting a PS requirements analysis.

Interviews and questionnaires are commonly used methods to investigate user preferences and needs when interacting with PS [18], as well as scenarios, storyboards and focus groups, which are well suited to shed light on this type of needs. Preference elicitation materials used in the requirements analysis should be designed in such a way to allow users to express their own meanings and interpretations about the type of personalization proposed or looked for, so that designers can achieve a more reliable and comprehensive understanding of the personalization value space [10] for a system, which is not just based on needs or preferences regarding tasks to be performed.

How can we add a temporal dimension to the characterization of user needs?
A first method would consist in directly questioning users about changes in the factors that can lead to preference evolution (e.g., to what extent do the habits formed with other systems are persisting once they have started using the new PS?).
Repetitive measurements based on questioning and user observations could also be useful to detect change when it occurs, as well as to design an adaptation mechanism that can sense and properly respond to it.
As an example, in mobile or ubiquitous systems design, experience sampling techniques are often deployed for questioning users even before having developed a prototype of the new system. Experience sampling is a method used in flow research [12] which consists in signaling participants during their daily lives, to collect assessments of their experiences in natural settings, in real-time (or close to the occurrence of the experience being reported), and on repeated time occasions. Reports can be made in response to a random signal (e.g., emitted by a pager or PDA), at predetermined times during the day (e.g., daily diary) or following particular events (e.g., interaction with another person or a digital system). This method would be particularly suitable to apply when the type of questions to pose are multiple-choice questions, true–false questions, text auto-completion, numerical ratings (e.g., Likert scales) and it is important to ask them when the user is immersed in the relevant usage context (e.g., [11] found that paper-based surveys are not reliable methods for measuring privacy concerns regarding context-aware services, while experience sampling questionnaires are more effective). Also, in [4] mobile questionnaires were augmented with Web-based diaries to gather qualitative accounts of user preferences and reasons for users' previous in situ choices and ratings. Real time data collection of user preferences would allow researchers to conduct preliminary data analyses almost immediately, send new questions or adapt questionnaires dynamically, and to prepare targeted ex situ inquiries (e.g., individualized interviews). Effective summary reports and visualizations of users' answers stored in the database can then be generated to better appreciate the relevant evolution of their preferences and needs that have been collected during the requirements analysis.

3.2 Early design

How can designers be helped to take into account the less well-understood types of user needs in the conceptual design?
The results of requirements analyses inspired by our framework are likely to yield additional requirements that should be borne in mind during the following stages of design and testing, which should also be planned in such a way to assess the extent to which these requirements are fulfilled.
A complementary strategy is to approach the conceptual design phase by keeping in mind how unlikely it is to understand completely what will happen in real use, no matter how good the previous phase has been conducted. That is why a PS conceptual design should be based on the principles of openness and flexibility, allowing the new system to be used also in unexpected ways, as a consequence of users' appropriation [5]. A relevant guideline to follow would be to make key aspects of the

personalization features developed and their effects as visible as possible to the user, to support an easier appreciation of the added value they provide if compared to nonpersonalized versions of the system. Designers should try to make as explicit and clear as possible the intentions behind the personalization features implemented (more than trying to explain to the user the complexity of the adaptation mechanism developed), so as to enable users to provide more informed feedback about their usefulness in practice; this could form also a more reliable base for making future decisions regarding refinements or, if required, re-examination of the adaptation provided.

How can the initial design take into account the expected evolution of needs?
Since both the environments in which a PS is used and the preferences of its users are likely to change over time, design should take into account the expected evolution of these requirements and find out adequate solutions to accommodate them.
Contextual design inquiries [2] have often been used to understand and follow how different situational factors and evolving user needs affect interaction with a system, by directly observing users in their natural settings and activities. A main issue relevant to PS design is that often if you optimise for one task you typically make others more difficult to perform [5]. Also, it is very difficult to define precisely tasks or activities that are performed in natural settings so that their description during initial design is often incomplete and approximate, and it typically ignores exceptions. A relevant heuristic to follow in most cases would be to design PSs that provide the necessary functions so that the user can perform a certain activity in different ways (according to their different preferences over time). Of course a PS should try to provide fast paths to optimize the performance of repetitive tasks, but this objective should not be the designers' only concern. Also, a PS focusing mainly on supporting task execution efficiency might encourage passive user behavior, thus exposing users to the risk of getting stuck in a local optimum, and preventing them from a more active exploration of the potential benefits that the adaptation features could provide in different situations.

How can the initial design take into account the expected characteristics of the process of preference formation?
During the early design phases of a PS, designers should try to involve users who have different overall approaches to preference formation. There are factors related to motivation, attractiveness, emotions and cultural background that need to be taken into account to understand the process of preference formation (e.g., some users may need more pleasurable experiences of interaction with the system to develop fidelization and start to appreciate the added value provided by personalization). Since in PS design affective and cultural dimensions of preferences have received less attention than cognitive factors, methods such as participatory design and cultural probes could be effectively applied to collect more detailed accounts on these dimensions (e.g., users' aesthetic preferences and cultural concerns [6])
A main issue to consider would also be how to best support user transition from novice to more expert interaction with the PS. Since online support or documentation is often disregarded by both users and interface designers, a better strategy might be to encourage end users to contribute material for these sections in the form of sharing

favourite tips and tricks about interesting experiences of PS usage, that could be useful to other (e.g., less experienced) users as well. These forms of mutual support or word of mouth could foster the formation of more accurate and realistic judgements about the system, as a result of knowledge sharing within a community of practice or informational cascades [3].

3.3 Iterative testing

How can we make it possible to take into account the full range of users' preferences in iterative testing?

The focus on predicting preference evolution is expected to provide a number of hypotheses that can be tested to some extent during formative evaluation. The different personalization techniques designed can be presented to study participants (e.g., by means of wizard-of-oz or prototyping methods [8][20][15]) asking for their comments, ratings or rankings to define which are the least or most favored ones according to the type of interaction experienced. Designers should define precise ways of measuring the direct experienced value for users when interacting with the variable components of the system developed (e.g., the alternatives offered in option setting). This could enable them to better understand and take into account which are different situated preferences of the target users and to refine the design and presentation of the system's options in such a way to get each user into the best possible system configuration, according to specific preference and usage conditions.

How can we test hypotheses about longer-term preference evolution within the type of short-term study that is typical of iterative testing?

For the study of changes in user preferences over time, wizard-of-oz methods should be adapted by taking care of providing participants with longitudinal scenarios regarding fictional characters' tasks and preferences for the PS, so as to provoke more reasoning and comments from users about the dynamics of preference formation and change. Another suggestion would be to combine these methods with a temporal compression strategy (comparable to a technique used to test the stability of an object by repeated placement of a heavy weight on it every few seconds for several hours, so as to simulate the type of stress which would occur in several years of normal use). This would allow designers to get comments and feedback from users that have already acquired high levels of familiarity with the adaptation mechanism under study, without having to wait until this level of familiarity has been spontaneously acquired.

Other possible methods would be adapted versions of expert-based assessments (also called discount techniques). Once an adequate corpus of empirical knowledge on user preferences for PS is available, new heuristics and guidelines representing this knowledge should be developed to support experts' assessments (e.g., by using a preference-oriented version of heuristics for Heuristic Evaluation [16]).

Another form of innovation would consist in adapting the type of questions used by experts when performing Cognitive or Heuristics Walkthrough of PS, to better focus their inspections on anticipating differences and changes in user preferences. There are important aspects of preference formation and expression (e.g., deciding which option to choose among the ones available for customizing an interface) that cannot be assimilated to the mere acquisition of procedures for executing tasks (which are typically assessed by employing usability criteria). More relevant questions to ask in assessing user preferences for a PS would be '*can the user understand the customization options available, their consequences, the intentions behind the adaptation features implemented?*', instead of '*can the user easily notice and perform the correct action available?*'. Although addressing preference evolution could raise the complexity of conducting expert-based evaluations, tuning the inspections on the special topic of user preferences and on the most critical components of an adaptive system should make this task rather manageable by experts.

3.4 Summative evaluation

Where longer-term summative studies are feasible, how can we support their design?
When feasible, large-scale, longitudinal studies of user preferences combining the collection of qualitative data (e.g., based on ethnographic, contextual or participatory approaches) as well as quantitative data (including logs) should be carried out. We expect that the framework provided will offer conceptual support in the design of these studies and will get the opportunity of being further refined and extended by the empirical evidence gradually accumulated by researchers.

A major outcome of the framework would be that of raising designers' awareness about the value of understanding user preferences for PS and their change, as they constitute important aspects for explaining both the adoption and appropriation processes regarding a PS. The guidance provided by the framework is also in terms of facilitating a more systematic analysis of previous studies on user preferences and an easier definition, as well as investigation, of future research questions. By adopting a preference-oriented perspective throughout the whole UCD of PS, the set up of longitudinal evaluations should become for designers more essential and (hopefully) less demanding to carry out.

4 Concluding Remarks

With this brief presentation of a new theoretical perspective, we hope to make readers aware of the many interrelated factors that can determine the ultimate acceptance of PS, just as strongly as the inherent intelligence of the systems themselves.

The application of our framework to the UCD of PS is likely to shed light on how to remove current obstacles in the adoption of these systems by their users, and to inspire new design solutions supporting a better evolution of user-system interaction in the long term.

References

1. Benyon, D.R., Innocent, P.R. & Murray, D.M.(1987). *System Adaptivity and the Modeling of Stereotypes.* Paper Presented at INTERACT '87, Second IFIP Conference on Human-Computer Interaction, the Netherlands.

2. Beyer, H. and Holtzblatt, K. (1997) *Contextual design: Defining customer-centred systems*. Morgan Kaufmann: San Francisco.
3. Bikhchandani, S., Hirshleifer, D., and Welch, I. (1998), "Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades," *Journal of Economic Perspectives*, Volume 12, Issue 3, pp. 151-170.
4. Consolvo, S., Harrison, B., Smith, I., Chen, M., Everitt, K., Froehlich, J., Landay, J. (2006). Conducting In Situ Evaluations for and with Ubiquitous Technologies. *International Journal of Human-Computer Interaction 2007*, Vol. 22, No. 1-2, pp. 103-118.
5. Dix, A. (2007). Designing for Appropriation. In *Procedings of BCS HCI 2007, People and Computers XXI,* Volume 2, BCS eWiC.
6. Gaver B., Dunne T., and Pacenti E. (1999). Design: Cultural probes. *Interactions 6*, 1, 21-29
7. Gena, C. (2006). *A user-centred approach for adaptive systems evaluation*. In S Weibelzahl, A Paramythis & J Masthoff (ed), Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, associated with AH'06 (Dublin, Ireland), 2006
8. Hook K., (2000). Steps to take before IUIs become real. *Journal of Interacting with Computers*, vol. 12, no. 4, 409-426.
9. Jameson, A.(2008). *Adaptive interfaces and agents*. In Sears, A., Jacko, J.A., eds.: The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications. 2nd edn. Erlbaum, Mahwah, NJ.
10. Karat, C., Brodie, C., Karat, J., Vergo, J., and Alpert, S. (2003) *Personalizing the User Experience on ibm.com*. In Vredenburg, K. (Ed.), *IBM Systems Journal*, 42, 2, pp. 686-701.
11. Khalil A.and Connelly K. (2007). Do I Do What I Say?: Observed Versus Stated Privacy Preferences. *Proceedings of 11th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT)*, September, 2007
12. Larson, R., & Csikszentmihalyi, M. (1983). *The experience sampling method*. New Directions for Methodology of Social and Behavioral Science, 15, 41-56.
13. Maguire, M. (2001). Methods to support human-centred design. *International Journal Human-Computer Studies, 55*, 587- 634.
14. Masthoff, J. (2002). *The evaluation of adaptive systems*. In N. V. Patel (Ed.), Adaptive evolutionary information systems. Idea Group publishing, 329-347.
15. Masthoff, J. (2006) *The user as wizard: A method for early involvement in the design and evaluation of adaptive systems.* In S Weibelzahl, A Paramythis & J Masthoff (ed), Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, associated with AH'06 (Dublin, Ireland).
16. Nielsen J., (1993). Usability Engineering. Boston, MA, Academic Press.
17. Van Velsen, L.S., Van der Geest, T.M. & Klaassen, R.F. (2006). *User-Centered Evaluation of Adaptive and Adaptable Systems.* Paper presented at the fifth workshop on User-Centred Design and Evaluation of Adaptive Systems. June 20th, Dublin, Ireland.
18. Van Velsen, L.S., Van der Geest, T.M. & Klaassen, R.F. (2007). *Testing the usability of a personalized system: comparing the use of interviews, questionnaires and thinking-aloud*. Paper presented at the IEEE Professional Communication Conference, Seattle, USA.
19. Weibelzahl S., (2005). *Problems and pitfalls in the evaluation of adaptive systems*. In S. Chen & G. Magoulas (Eds.). Adaptable and Adaptive Hypermedia Systems (pp. 285-299). Hershey, PA: IRM Press
20. Wilson, J. and Rosenberg, D. (1988). Rapid prototyping for user interface design. In M. Helander (Ed.), *Handbook of Human-Computer Interaction*, New York, North-Holland. pp. 859-875.

AnAmeter: The First Steps to Evaluating Adaptation

F. Tarpin-Bernard[1] , I. Marfisi-Schottman[1], H. Habieb-Mammar[2]

[1] LIESP
Université de Lyon, LIESP,
INSA-Lyon, F-69621, Villeurbanne, France
Tel: +33 (0) 4 72 43 70 99
Fax: +33 (0) 4 72 43 79 92
franck.tarpin-bernard@insa-lyon.fr
iza.marfisi@insa-lyon.fr

[2] Human-Centered Software Engineering Group
Concordia University, 1455 Maisonneuve West
Montreal Quebec Canada H3G 1M8
mammar@encs.concordia.ca

Abstract. This paper presents the online AnAmeter framework that helps characterize the different types of adaptations a system features by helping the evaluator fill in a simple form. The provided information is then processed to obtain a quantitative evaluation of three parameters called global, semi-global and local adaptation degrees. By characterizing and quantifying adaptation, AnAmeter provides the first steps towards the evaluation of the quality of a system's adaptation. AnAmeter is an open tool available as freeware on the web and has been applied to a selection of well known systems. To build this evaluation grid we also collected a number of systems that cover the full range of adaptation types.

Keywords: adaptation degree, evaluating adaptation, adaptivity, adaptability, characterization, quantification.

1 Introduction

People using computer systems are of various ages and have all different kinds of interests and background knowledge. In addition to traditional desktops, the variety of computing platforms includes mobile telephones, personal digital assistants (PDAs), pocket PCs, wearable and immersive environments and many more. In this context, novel adaptive and adaptable systems are emerging. Faced with this huge set of propositions, it is very difficult to characterize to what extent a specific application is adaptive or adaptable. Likewise, it is difficult to identify the new adaptation features that should be implemented in a system in order to increase its adaptation degree. For these reasons, it is necessary to *characterize* all the different kinds of adaptations that can possibly exist and define a proper way of *quantifying* the degrees of these adaptations. In order to accomplish a good evaluation framework, a measure of usability should be added (see Fig. 1). These indicators could be used either to improve a system by identifying its strengths and weaknesses, or for objectively comparing several systems of the same family.

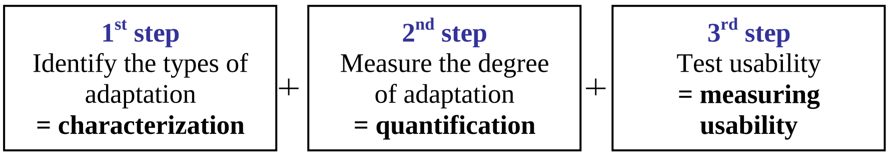


Fig. 1: The three steps to evaluating a system's adaptation

In this paper, we present a first proposal, AnAmeter, for *characterizing* and *quantifying* the adaptation of a system. This tool is largely based on our analysis of the multiple facets of adaptation we will develop in the first section. Then, we present the core of AnAmeter: a grid that helps characterize the adaptations by crossing the adaptation factors and the aspects of adaptation. Based on this grid, we build a quantification technique that provides a measure of the adaptation degree. Finally, the interest, limits and potential extensions of AnAmeter are discussed.

2 The multiple Facets of adaptation

Many studies have tried to build a taxonomy of adaptive interfaces [1, 2, 3]. Based on these works and on a first design space provided by Vanderdonckt & al. [4], we consider that the most important questions to characterize adaptation are:

- ➢ Who initiates the adaptation?
- ➢ What are the factors to which the system can adapt?
- ➢ What aspects of the system are adapted?

In the next part, we look at each of these questions and elaborate a list of answers.

2.1 Who initiates the adaptation?

To analyze systems adaptations in a functional way we chose to use the two possible sources responsible for initiating the adaptation as identified by Kobsa & al. [5]:

- ▪ The system itself. In this case, the adaptation is automatically initiated and the system is called *adaptive*.
- ▪ The user. In this case, the adaptation is requested by the user and the system is called *adaptable*.

2.2 What are the factors to which the system can adapt?

Usually, the need for adaptation is associated to the notion of context. In the human computer interaction (HCI) field, a context is generally described according to three dimensions [3]: the user, the platform and the environment. However, a specific user placed in the same environment using the same interaction platform could require some adaptation relevant to his/her activity. That is why we define four sets of factors of adaptation: user, interaction platform, environment and the activity.

In the next part, we itemize these four factors of adaptation into sub-factors. The lists of sub-factors are not meant to be fully exhaustive, but as we will see later, they will help evaluate the levels of adaptation provided by a system. This classification work was done thanks to a detailed review of the state of the art and "context of use" definitions provide by standards like IEC CDV TR 61997 [6] and ISO 9241-11 [7]. We also analyzed more than 50 systems found in articles or available in public distribution (complete list on http://liesp.insa-lyon.fr/AnAmeter/References.php).

2.2.1 User

A User model usually refers to various user characteristics [8, 9]. We can group these characteristics into four sub-factors:

- Knowledge and level of experience: The knowledge refers to the user's theoretical understanding of the subjects treated in the system. The level of experience refers to the necessary skills to use the system itself. For example, the systems can adapt the content of the lessons or give more helping tips to users that are not familiar with the system.
- Socio-demographic characteristics and user role: Characteristics such as age, gender, weight, height, wage, profession, hobbies, cultural preferences and user role… are useful factors of adaptation for all kinds of systems. When looking for tourist activities, a GPS system will filter the information showing only the attractions compatible with the user’s interests and his propensity to spend.
- Cognitive abilities and emotional state: The different modes of perceiving, memorizing, learning, judging etc. and the emotional state [10] (happy, sad, worried, frustrated, panicked, confident…) may also be considered for adapting.
- Perceptual and motor abilities: These characteristics are useful to enable the systems to be used with disabilities (vision, manipulation, etc.). These disabilities can range from slight myopia or color blindness to total deafness and paralysis.

2.2.2 Interaction platform

The interaction platform describes the physical characteristics of the devices. The major characteristics that a system may take into account are the following:

- Computing power and autonomy: Systems often need to be adapted to the platform’s processing power and the memory capacity. For some portable devices it is also worthwhile to adapt to the battery level by shutting of certain services for example.
- Input/Output device: Some systems are available on a wide variety of platforms. Certain web browsers adapt to the different screen sizes and input devices such as mice, keyboards and pens when used on desktops, laptops, or telephones.
- Software environment: Computer systems are almost always used alongside other systems on the same platform. These systems can adapt to cohabitate, synchronize and even cooperate with each other.
- Connectivity: More and more systems are now using network connections. The connectivity factor is therefore very important. Systems can adapt to cope with the lack of connection or slow connectivity or even type of network.

2.2.3 Environment

The third factor is the environment, a term used to cover the physical, social and organizational elements that are outside of the interactive system (platform & user).

- Human environment: In some cases, systems can adapt to the other people who are interacting with the user (directly or through the system). This kind of adaptation can be used for multi-user applications (games, forums, chats), for applications that adapt to the other users' actions (commercial websites such as Amazon that propose frequently bought products or Google that considers the popularity of websites) or even for applications that detect humans present in the same physical area.
- Machine environment: This type of environment is defined by any reachable material such as external servers, extra output and input devices (video projector, motion detector…) which are not part of the Interaction Platform but could be connected on the fly to the main system.
- Ambient characteristics: Systems can also adapt to the luminance, temperature, the level of noise and the movements of the device.
- Spacio-temporal characteristics: Many GPS navigation systems propose potential interesting tourist areas by using geographic latitude and longitude measures. Localisation can also be expressed semantically if the system identifies a specific area such as a room or on a larger scale, a country or time zone.

2.2.4 Activity

The fourth factor is the activity itself. At a micro level, it includes task characteristics and at a macro level it includes the general activity and the user's goal.

- Task characteristics: The frequency, complexity, dangerousness and confidentiality character of the task can be taken into account to adapt the system by using icons and fast links for frequent tasks (favorite links) or extra warning messages and backup copies for dangerous or confidential tasks.
- Task flow: Here, the task is considered as a part of a tasks flow. For example, if the user usually does task B after task A, the system might set a quick or automatic launch to task B each time task A is done. The system can also keep a historic of the different tasks done so that the user can access them faster.
- User's goal: For each activity (combination of tasks), the user can have a different set of goals. For example, Photoshop® could adapt according to whether the user is editing photos, looking at a slide show or sorting photos…
- General activity: The general nature of the activity weighs heavily in a successful adaptation. Someone wanting to have fun, for example, will not have the same way of using a system as someone who wants to learn or work.

2.3 What aspects of the system are Adaptable?

Many aspects of applications can be adapted. We ordered these aspects by using a common approach of HCI engineering: the PAC (presentation, abstraction, control) model, developed by Coutaz [11]. This model has the advantage of clearly separating the functional aspect of the system called "abstraction" from the interface components called "presentation". The "control" is in charge of linking these two worlds and thus externalizing the means and rules of communication. In the next section, we clarify these aspects by using an example of a GPS system.

2.3.1 Abstraction

In this part we will be talking about the adaptation of the information and the data proposed by the system and the way the different services behave.

- Data & information: A GPS system in a car will give different information when asked for the hotels in the surrounding area. The hotels proposed will depend on the localization of the car.
- Service behavior: A company time-table planner for example will authorize the boss to take holidays whenever he wants but will send an approval email and mark the holidays as “to be confirmed” for any other employee.

2.3.2 Control

The control module is in charge of giving access to the services and data available in the system by interacting in different ways with the user.

- Filtering services and data: For various reasons, adaptation might mean limiting the number of services offered or providing only a partial access to a complex service. On our GPS system, for instance, the services to find a tourist attraction are only available when the system is set on “vacation” mode.
- Interaction mode: Systems can choose to accept input and deliver output via many devices. For example, when the car is running, the information on the screen of the GPS system is read out loud by a voice synthesizer.

2.3.3 Presentation

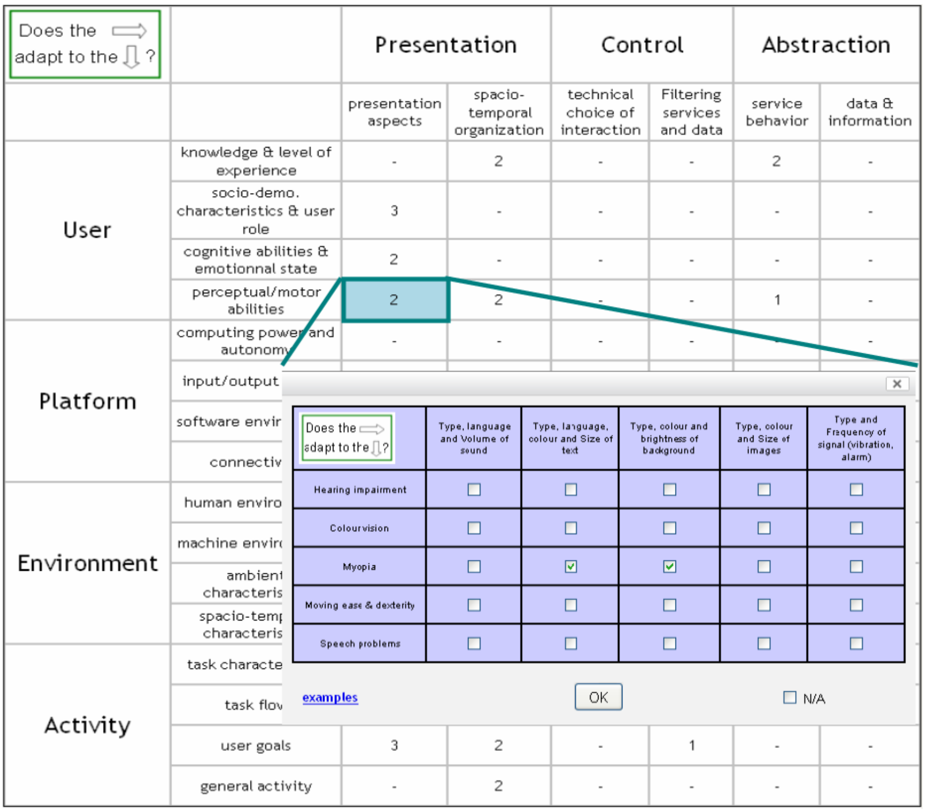
- Spacio-temporal organization: The elements of information can be arranged in a variety of ways. For example, the GPS system will present the descriptions of the hotels in a specific order by calculating the distance to the hotel.
- Presentation aspects: Finally, we get to the outermost layer of the surface, which includes elements such as colors, shapes, buttons, boxes, menus, volume, sound effects... For example, our GPS system will change the colors and the brightness of the screen when night falls.

3. Characterizing and Quantifying Tool

As mentioned in the introduction, the AnAmeter tool characterizes the adaptation and measures the quantity of this adaptation. It is important to keep in mind that it does not yet measure the usability of the adaptation. For a complete evaluation, it would be necessary to at least measure the utility of the adaptation to make sure the users really need it, the quality of the implementation to make sure the adaptation is easy to use but also the efficiency of the adaptation to make sure it actually helps the users work faster or better. This third step (see Fig. 1) is not in the scope of this paper but will benefit from our work. We present AnAmeter, an open system to support an iterative and participative building approach to develop a standard evaluating tool.

3.1 Characterizing adaptation

Using the classification presented in the previous section, we can build a first characterizing grid by crossing the adaptation aspects versus the factors of adaptation. This grid can be used to break down types of adaptability as well as the types of adaptivity (Fig. 2). Each factor (respectively aspect) is divided into sub-factors (respectively sub-aspects). The sub-factors and sub-aspects are also broken down into elements. For ease of presentation, we have not drawn these last subdivisions but each cell of the main grid contains a smaller grid composed of these elements which refer to the finest grain of description. Each lower level cell corresponds to the question "Does this aspect adapt to this factor?" If this is the case then the cell should be checked. For example, the system tested in Fig. 2 adapts the "size of the text" and the "type and color of the background" to the users "myopia".



Does the ⇨ adapt to the ⇩ ?		Presentation		Control		Abstraction	
		presentation aspects	spacio-temporal organization	technical choice of interaction	Filtering services and data	service behavior	data & information
User	knowledge & level of experience	-	2	-	-	2	-
	socio-demo. characteristics & user role	3	-	-	-	-	-
	cognitive abilities & emotionnal state	2	-	-	-	-	-
	perceptual/motor abilities	2	2	-	-	1	-
Platform	computing power and autonomy	-	-	-	-	-	-
	input/output						
	software envir						
	connectiv						
Environment	human enviro						
	machine envir						
	ambient characteris						
	spacio-temp characteris						
Activity	task characte						
	task flov						
	user goals	3	2	-	1	-	-
	general activity	-	2	-	-	-	-

Does the ⇨ adapt to the ⇩ ?	Type, language and Volume of sound	Type, language, colour and Size of text	Type, colour and brightness of background	Type, colour and Size of images	Type and Frequency of signal (vibration, alarm)
Hearing impairment	□	□	□	□	□
Colourvision	□	□	□	□	□
Myopia	□	☑	☑	□	□
Moving ease & dexterity	□	□	□	□	□
Speech problems	□	□	□	□	□

Fig. **2 : V1.0 of the main grid and a smaller grid containing aspect and factor elements.**

Some of these questions might not make very much sense in certain situations or for a specific type of system. This is why we add a N/A (non-applicable) option. The complete evaluation requires filling out two grids (one for the system's *adaptability degree* and one for the system's *adaptivity degree)* and therefore answering a long list of questions. In order to ease the work of the evaluator and speed up the process, we have built an online tool for handling the grid that only requires the evaluator to check boxes. The tool is available online at http://liesp.insa-lyon.fr/AnAmeter. To make sure the system was applicable to all adaptive or adaptable systems, we trialled it on a web browser (Google), a writing and calculating system (Office 2007), an online bookstore (Amazon) and a communication system (smart phone XDA O_2). For the first evaluations we carried out, filling out one grid took about 60 minutes.

3.2 Quantifying adaptation

Now that we have built a grid to *characterize* the adaptability and the adaptivity of a system, we want to *quantify* these adaptations. Once each cell of the smaller grid relevant to sub-aspect B and sub-factor C is filled in, an adaptation degree $A_{B/C}$ ranging from 0 to 3 is automatically calculated according to the number and distribution of the boxes checked using the rules detailed in Table 1. For example, Fig. 2 shows the small grid of the sub-aspect “presentation aspects” and the sub-factor “perceptual/motor abilities”. Once the evaluator clicks on the OK button, the adaptation degree $A_{\text{presentation aspects / perceptual, motor abilities}}$ will be automatically calculated according to the number and position of the ticks entered in the grid.

Table 1 : Scoring process for the adaptation degree.

Degree	Meaning	Reading in the grid	Example
$A_{B/C}$= 0	The system does not have this type of adaptation.	No checked boxes.	□□□□ □□□□ □□□□
$A_{B/C}$= 1	One aspect is adapted to one factor.	One checked box.	□□□□ □☑□□ □□□□
$A_{B/C}$= 2	One aspect is adapted to several factors or several aspects are adapted to one factor.	Checked boxes only on one row or only on one column.	□□□□ ☑☑□☑ □□□□
$A_{B/C}$= 3	More than two aspects adapt to more than two factors.	Checked boxes on at least two rows and two columns.	☑□□□ ☑□☑☑ ☑□□□

When all the cells in the main grid relevant to the aspect B and the factor C are filled in with a score of 0, 1, 2 or 3, a local adaptation degree, $LA_{B/C}$ is determined by calculating the average of these scores. The N/A cells will not be considered in the calculations. The results is then converted into a percentage as shown in equation n°1 (100% corresponds to a score of 3 in all the cells).

$$LA_{B/C} = \frac{\sum A_{B.item/C.item} \times 100}{(n-m)\times 3} \qquad \begin{array}{c}\text{n cells relevant to B and C}\\ \text{m N/A cells relevant to B and C}\end{array} \qquad (1)$$

Once all the local adaptation degrees $A_{B/j}$ relevant to an aspect B are calculated, the semi-global *aspect* adaptation degree AA_B can be found with equation n°2. In the same way, equation n°3 is used to determine the semi-global *factor* adaptation degree

FA$_c$ relevant to the factor C.

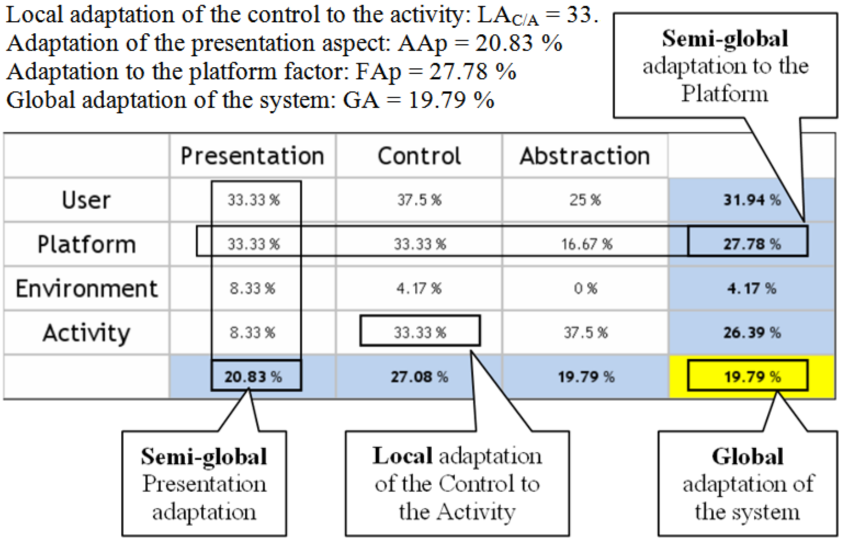
$$AA_B = \frac{\sum LA_{B/j}}{n} \qquad \forall j \subset \{\text{factors}\} \tag{2}$$

$$FA_C = \frac{\sum LA_{i/C}}{n} \qquad \forall i \subset \{\text{aspects}\} \tag{3}$$

Finally, the global adaptation degree, GA, is determined by taking the average of the semi-global adaptation degrees - either of all the aspects or of all the factors - as shown in equation n°4.

$$GA = \frac{\sum AA_i}{n} = \frac{\sum FA_j}{n} \qquad \begin{array}{l} \forall j \subset \{\text{factors}\} \\ \forall i \subset \{\text{aspects}\} \end{array} \tag{4}$$

To enable easy understanding of these adaptation degrees, we then identify the aspects and the factors by using the first letter of their name. Also, the adaptation degree relevant to *adaptivity* (self-adaptive) will be marked with an apostrophe (LA'$_{C/A}$, GA'…). Fig. 3 illustrates an example of these equations:



Local adaptation of the control to the activity: LA$_{C/A}$ = 33.
Adaptation of the presentation aspect: AAp = 20.83 %
Adaptation to the platform factor: FAp = 27.78 %
Global adaptation of the system: GA = 19.79 %

	Presentation	Control	Abstraction	
User	33.33 %	37.5 %	25 %	**31.94 %**
Platform	33.33 %	33.33 %	16.67 %	**27.78 %**
Environment	8.33 %	4.17 %	0 %	**4.17 %**
Activity	8.33 %	33.33 %	37.5 %	**26.39 %**
	20.83 %	**27.08 %**	**19.79 %**	**19.79 %**

Semi-global adaptation to the Platform
Semi-global Presentation adaptation
Local adaptation of the Control to the Activity
Global adaptation of the system

Fig. 3 : Example of local, semi-global and global adaptation degrees.

4. Discussion

The advantages of the AnAmeter tool are:

1) Its simplicity. The tester fills out the grid by answering simple Boolean questions of the following type: does a precise aspect of the system adapt to a precise factor? Clear examples with references are available for each type of adaptation.
2) Its precision. The tool provides precise local evaluations. This is very useful for

systems that specialize on a few adaptations. It is also possible to evaluate the system from two fundamental different points of view: adaptability (user-initiated adaptation) or adaptivity (automatic system-initiated adaptation).

3) Its completeness. We have tried to build a complete evaluation protocol that systematically considers all the possible adaptations so that none are left out. We hope the list of adaptations and the 300 examples of their implementation available on the AnAmeter platform will also inspire system designers.
4) Its ease for comparing adaptations. AnAmeter can be used to compare two systems of the same family but also to measure the evolution of a system by evaluating the effect of adding or withdrawing adaptations.
5) Its extensibility and flexibility. Our idea was to offer a robust basis for the community to build on. The architecture of our tool makes it easy to extend by adding other elements, by dividing the sub-categories or extending the measuring scale.
6) Its accessibility. AnAmeter is freely accessible on the web along with a selection of completely tested systems (http://liesp.insa-lyon.fr/AnAmeter). This makes it possible for the same system to be tested by several evaluators who could then combine their results to obtain a mean value for the adaptation degree.

Although AnAmeter has many advantages, the fact that the approach tries to be as complete as possible extends the time required to evaluate a system to approximately one hour. Indeed, this first version of the grid contains 22 aspect elements and 59 factor elements which add up to more than a 1000 Boolean questions to answer for a highly adaptive or adaptable system. Of course, for most of the systems, entire sections of the grid will be left out or marked as non-applicable, greatly reducing the amount a work. By creating an online tool that enables easy manipulation of the grid and calculates the adaptation degree automatically, we have lightened the task but it is still represents quite an investment of time and effort. We hope it will be possible to improve the grid with the scores and the comments of people who use it.

5. Future work

We believe that building an evaluating tool, widely accepted by the community can only be done in a cooperative way with the help of the members of this community. AnAmeter was created to serve as a basis for building on and this is why we created an open, extensible and flexible online framework.

In the near future, we plan on adding a measure of the adaptations usability to establish a global evaluation mark as seen in the first section (see Fig. 1). In order to do so, indicators such as utility, quality of the implementation or the added efficiency brought by the adaptation will have to be measured. These indicators can be found with different techniques such as interviewing users, analyzing tracking data or setting up a test session with specific tasks to be accomplished. If the adaptations of the system have already been identified and quantified with AnAmeter, the evaluators will have precious information to build an efficient protocol for this third step. Indeed, they will already have a global view of the important adaptations featured by the systems and have an idea of the factors and aspects to modulate so as to observe usability indicators.

The next step is to test the AnAmeter tool for ease of use by asking other people to use it to evaluate systems on their own and send us feedback. Now that AnAmeter is available on the web is should be easy to launch an evaluation campaign.

Finally, we plan to ask people with different user profiles to test the same system in order to see if the results coincide or not to measure AnAmeter's reliability.

6. Conclusion

In this paper, we present AnAmeter, a tool to characterize the multiple facets of adaptability and a quantification technique to measure the adaptability degree of an interactive system. We discuss the multiple facets of adaptation, primarily the aspects and factors of adaptation that serve as parameters. Then, we suggest the use of a scoring matrix to evaluate local, semi-global and global adaptation of an interactive system. We provide a first version of the scoring technique and simple formulas for calculating these adaptation degrees. The AnAmeter tool is presented as a starting point for the community to cooperatively build a widely accepted framework for evaluating any kind of adaptable or adaptive systems.

7. References

1. Brusilovsky, P., "User Modeling and User-Adapted Interaction", Adaptive Hypermedia, Vol. 11, Issue 1-2, 87--110 (2001)
2. Jameson, A., "Adaptive interfaces and agents", J. A. Jacko & A. Sears (Eds.), The Human Computer Interaction Handbook, Mahwah, NJ: Erlbaum, 305--330 (2003)
3. Calvary G., J. Coutaz , D. Thevenin, Q. Limbourg, L. Bouillon, J. Vanderdonckt, "A unifying reference framework for multi-target user interfaces", Interacting With Computers, Vol. 15 No. 3, 289--308 (2003)
4. Vanderdonckt J., D. Grolaux , P. Van Roy, Q. Limbourg, B. Macq and B. Michel, "A Design Space For Context-Sensitive User Interfaces", Proceedings of IASSE (2005)
5. Kobsa, A., J. Koenemann and W. Pohl, "Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships", The Knowledge Engineering Review, 16(2), Cambridge University Press, UK, 111--155 (2001)
6. IEC CDV TR 61997, "Guidelines for the user interfaces in multimedia equipment for general purpose use" (2000)
7. ISO 9241, "Ergonomic requirements for office work with visual display terminals" (1998)
8. De Bra, P., A. Aerts, B. Berden, B. De Lange, B. Rousseau, T. Santic, D. Smits and N. Stash, "AHA! The Adaptive Hypermedia Architecture", Proceeding of the ACM Hypertext Conference, 81--84 (2003)
9. Habieb-Mammar H., F. Tarpin-Bernard "CUMAPH: Cognitive User Modeling for Adaptive Presentation of Hyper-documents: An experimental study", AH 2004, Eindhoven University of Technology, Netherlands, 23--26 (2004)
10. Picard, R.W., "Affective Computing". In The MIT press, United State (2008)
11. Coutaz J., "PAC: an object oriented model for implementing user interfaces", Bull., vol. 19, 37--41 (1987)

An Experiment to Evaluate how to Better Present User Models to the Users

Fabiana Vernero, Alessandra Petromilli, Federica Cena and Cristina Gena

Dipartimento di Informatica, Università di Torino
Corso Svizzera 185, Torino, Italy
{vernero,petromilli,cena,gena}@di.unito.it

Abstract. The externalisation of user models may allow users to understand how a user-adapted Web application makes its adaptation decisions, enabling them to inspect and modify the values stored in their user model. When externalising a user model both the underlying representation of the user model and the visualization used to present it have to be taken into account. In this paper we present a study aimed at evaluating different ways to represent and visualize user models. We use a social recommender system in a cultural events domain (iCITY) as case study. To our purposes, we conducted two experiments: i) a large between-subjects on-line evaluation aimed at confronting different representation and visualization modalities; ii) a within-subjects experiment aimed at confronting the same experimental condition.

1 Introduction and Related Work

Usually, user models tend to be hidden and out of the user access and control [9]. However, many systems have started to involve users in the maintenance of their model, especially in educational context, for example by enabling them to edit it [8], or to negotiate the contents of the learner model with the system [7, 4].

Open user models are models of the users that are available for viewing, and sometimes maintaining them by the users themselves (and sometimes also by other users, such as peers and teachers in educational context) [1]. A further step is the *scrutable user model* [9, 10], an open user model containing not only the user model data but also the evidence about how such data have been derived by the adaptive system. A *transparent system* [6] allows the user to understand the way it works and explains system choices and behaviour. Understanding, accepting and trusting a personalisation system may additionally improve the user-system interaction.

Bull and Kay [1] sustain that a model should be available in a form similar or identical to its underlying representation for greater accuracy. However in case of complex representation the similarity is not mandatory. What is important is that the user might understand the model. Thus, in case of complex underlying representation, a simpler representation, and consequently visualization, could be preferred.

Concerning the representation of the user models, and the visualization used to present them, several solutions have been proposed in the past. The visualization of user models can take a simple textual form such as in ELM-ART [17], Personis [11], SIV - Scrutable Inference Viewer [12]. Other systems visualize the user model content in graphical ways. Some systems use very simple and intuitive visual representations, such as *sliders* (LOZ [13]), *emoticons* (Subtraction master) [3], *stars* (UMPTEEN [2]), *colors* (The Fractionator [2]), *bar charts* (PSAT/NMSQT [19]. In other cases, the information presented can be more complex, such as a graphical externalisation in the form of a Bayesian network [18], a hierarchical tree structure (Viewer[9]), a conceptual graph [7], multiple views (Flexi-OLM [2]). Finally, other systems exploit special metaphors such as *magic wands* (Wandies [2]) or *cups* (INSPIRE [15]).

In this paper we present a set of evaluations aimed at identifying the best modality of representation and visualization of an open user model in an existing social recommender system in a cultural events domain, iCITY[1]. The paper is structured as follows. In Section 2 we introduce our motivations and background. In Section 3 we present our experiments, describing in detail how we conducted the evalations, and which results we obtained. Section 4 presents the conclusions we draw from the experiments.

2 Motivations and Background

As highlighted by Norman [14] it is important to explain *why* and *on what basis* an application shows an adaptive behaviour. Knowledge about the inner working of an application helps users in interpreting the answers it provides, especially when personal data are manipulated. In order to reach these goals, we decide to externalize the user model of iCITY [5], a social recommender system in a cultural events domain which integrates adaptivity principles with Web 2.0 social features. In iCITY users are allowed to publish and share their own events, as well as rating, commenting, bookmarking and tagging other content; moreover, part of the events are provided via RSS-feed by the Turin Municipality. Great emphasis is also put on social networking. As regards adaptation, events are recommended according to their estimated interest for a certain user, balanced with their average rating and also considering the event date and location.

In iCITY, the user model maintains different types of information, such as user level of participation, user skills, and user preferences for the classes of the domain taxonomy[2]. A probability distribution of user interests is associated with each class of the taxonomy. Notice that the values in a probability distribution always sum up to 1. This means that if the value expressing user interest in a class increases, the values representing her interest in the other classes at the same level of the taxonomy proportionally decrease, in such a way that the sum remains equal to 1. For example, considering only two classes, if the user level of interest is 0.2 in “Music” and 0.8 in “Cinema”, and the user changes the first value to 0.3, then her level of interest in “Cinema” should be changed to 0.7.

We have conceptualized this representation of the user model interests overlaying the domain as *“relative representation”*. Given that, our aim is to find the better way to represent this conceptualization to the user, and to allow the user to modify her model maintaining such probability distribution[3]. As far as we know, all the systems in the lit-

[1] http://www.icity.di.unito.it/dsa-en/

[2] Regarding preferences, the iCITY user model overlays the domain model

[3] So far, the section of iCITY user model open to the users regards the user preferences, visualized in a plain textual way. The user is not allowed to modify her preferences (see for details [5])

erature (see Section 1) make use of *"absolute representation"*, wherein the user model values are presented in a scale, and each value is independent from the others. Even if an order can be derived and distances can be measured, such relations are not explicit to the user, who assessed each element separately. None of the reviewed past systems presents values by means of either an *"ordered representation"* (wherein the user model values are ranked in a list) or a relative representation. In an ordered representation, relations such as "superior to" and "inferior to" can be established among the various items. However, it is impossible to measure the "distance" or "difference" between two elements, as well as to assign the same value to different elements, unless co-winners are allowed. Finally, with a relative representation, users have to explicitly assess both order relations and distances among elements. The relative representation provides therefore more information: in fact, not only the relative values express how much a user is interested in a certain topic (or she knows about it) but they also highlight the relation among different elements. For example, the statement: "I assign 20% of my interest to literature and 80% to sport" is more informative than both "I like sport more than literature" (ordered values) and "I rate sport 100 out of 100 and I rate literature 25 out of 100"(absolute values).

However we hypothesized that, even if more informative, the relative representation conveys a more elaborated conceptual model of the inner representation of the system. This could increase the cognitive load of the users. On the contrary the other two representations seem to convey and represent concepts in a way easier to be comprehended by the final users.

Moreover also the visualization modalities chosen to present the user model can have an impact on the comprehensibility of the user model itself. We hypothesized that visualizations which are more often used in social web sites will be the most appreciated since users are already familiar with them. Thus, we decided to perform a set of experiments with the aim of verifying such hypotheses. More in particular, we wanted i) to investigate which graphic visualization, for each of the three representations, is the most understandable and usable. In particular we want to discover which metaphors could be used to better convey the ideas underlying the three representations; ii) to verify whether the "relative representation" is more informative for the user but also more elaborate to manage and to understand in comparison to the absolute and the ordered representation. Furthermore, we wanted to verify if user features (and in particular age, gender, education) influence in some way her preferences in user model externalisation, both regarding the modalities of value representation (ordered, absolute, relative) and the modalities of visualization (e.g., sliders, bar charts, etc). Finally, we wanted to verify whether users really feel useful to inspect and modify their models.

3 The evaluations

With the goal of investigating user preferences both for the user model representation and visualization modalities, as well as user opinions about user model externalisation, we performed in parallel two experiments: i) a large between-subjects on-line evaluation, where we compared different visualizations, given a certain representation modality, and ii) a within-subjects experiment, which allowed an in depth-evaluation and a comparison of the three representation modalities and their corresponding visualizations. In designing the user interfaces used to convey the different visualizations, we took inspiration from the user model of our case-study application. Therefore, we represented user preferences with respect to different categories of events, which correspond to the classes in the taxonomy of iCITY (Appointments, Cinema, Art, Music, Books, Theatre).

The two experiments involved 9 visualizations, implemented as dynamic web pages by means of JavaScript, Ajax, and PHP scripting. Notice that some of them were selected thanks to a previous pilot study based on the paper-prototyping technique. Visualizations were divided into three homogeneous groups, based on the user model representation (see Figure1):

1. ***Ordered representation***
   - *The list*: preferences are represented as an ordered list, sortable at will;
   - *The podium*: each category is represented by a sphere, positioned on a certain step of a podium, according to the level of interest. Preferences can be modified by moving the spheres;
   - *The medals*: preferences are on list where the order is indicated by means of gold, silver and bronze cups and medals; the order can be modified by sorting the names of the categories;
2. ***Absolute representation***
   - *The stars*: each category can be awarded from a minimum of zero to a maximum of 5 stars;
   - *The sliders*: preferences can be adjusted by means of sliders;
   - *The tag cloud*: preferences are represented as tags in a tag cloud: the bigger a tag, the higher the level of interest; preferences can be modified by increasing or reducing the size of the tags;
3. ***Relative representation***
   - *The coins*: each category is associated with a box containing some coins. Preferences are represented by the number of coins; there is a fixed number of coins. Preferences can be modified by moving coins;
   - *The bricks*: user interest in a category is represented by a pile of bricks - the higher the pile, the higher the level of interest. Preferences can be modified by moving the bricks from one pile to another;
   - *The pie chart*: each category is represented by a slice in a pie chart. Preferences are the size of the slice and they can be modified adjusting it.

3.1 On-line Evaluation

The first evaluation was carried out as an on-line test aimed at evaluating the proposed user model visualizations with a large number of users. We wanted to discover i) which visualization is the most appreciated, given a particular user model representation, ii) whether users actually appreciate the possibility to inspect and modify their user models and iii) if significant correlations exist between demographic features and user preferences in visualizations.

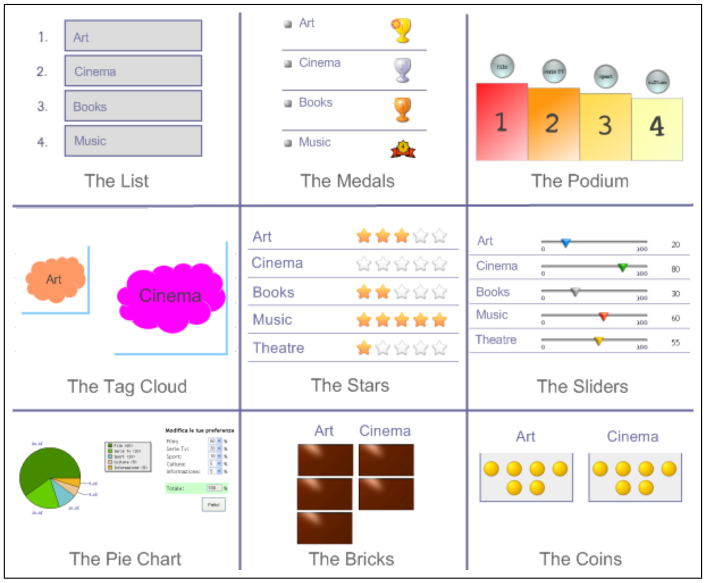


Fig. 1. The figure shows the main features of the visualizations used in the two experiments

Hypothesis. We hypothesized that visualizations which are more often used in social web sites will be the most appreciated since users are already familiar with them. Moreover, we thought that users would prefer prototypical interfaces which give prominence to visual aspects and allow direct manipulation. For such reason, we hypothesized that the preferred visualizations will be the "list" for the ordered representation, and "the stars" for the absolute representation. Since there are no examples of relative representations in existing systems, in this case we hypothesized that only the input modality (textual input vs direct manipulation) will impact on user preferences; consequently, the "coins" could result as the preferred visualization. Moreover, we thought that users appreciate open user models and that demographic variables had some influence in determining preferences.

Experimental design. Multiple factors (user model representations; visualizations) between-subjects design.

Subjects. Subjects were users of Facebook[4] and were therefore familiar with social media, as iCITY. They were recruited among the contacts of the authors and randomly assigned to one of the three groups: in this way, we obtained 100 subjects for the "ordered representation", 96 subjects for the "absolute representation" and 103 subjects for the "relative representation" group (299 subjects in total, 16-65 years old, 133 females and 166 males).

Measures and material. User preferences for the different visualizations were collected through an online questionnaire, personalized according to the group. The 3

[4] Facebook (http://www.facebook.com/) is one of the most popular social networking web sites.

groups of user model representation, and their corresponding visualizations, were made available online.

Experimental tasks. Subjects first accessed a page displaying a short thank-you message and the instructions. They were explained that they would access a series of visualizations of their preferences with respect to different categories of events, as they could have been automatically inferred by the system. They were invited to examine the visualizations and to try to "reply" by modifying/correcting the values. Subjects were also informed that they would be asked to fill in an anonymous questionnaire. After that, users accessed 3 different pages, each one containing an interactive user model visualization. The presentation order of the visualization was randomly changed for each user.

At the end they filled in the questionnaire. The first 4 questions were aimed at collecting basic demographic data (gender, age, education and job); then, users had to indicate the best and the worst visualization, and to give reasons for their choices. In the following 3 questions, subjects had to select one or more adjectives to describe each visualization, choosing from this list: "easy to use", "difficult to use", "pleasant", "unpleasant", "comprehensible", "incomprehensible", "amusing", "boring". The last 3 questions investigated: i) the subjects' opinion about the possibility to correct the values in their user model (answers were collected by means of a 4-point Likert scalewhere the different steps -in ascending order- corresponded to "very negatively", "negatively", "positively", "very positively"; ii) whether subjects would bother to correct their preferences in everyday usage and iii) whether subjects preferred a system where they could modify their preferences or a "traditional" one.

Results. As far as the "*ordered representation*" group is concerned, 63% of the users chose "the list" as their favourite visualization, followed by "the medals" (20%) and "the podium" (17%); this distribution of values is significant ($\chi^2(2) = 28.42$; $p < 0.001$). More than half of the subjects indicated "the podium", as the least favourite one; the distribution of values for this variable is also significant ($\chi^2(2) = 38.36$; $p < 0.001$).

User opinions. The list results "easy to use" (86% of subjects), "comprehensible" (65%) and "pleasant" (25%). Notice that the adjectives "difficult to use" and "incomprehensible" were used only once, while almost the same small number of users (14 and 11, respectively) described "the list" with the two opposite adjectives "boring" and "amusing", suggesting that the corresponding underlying dimension is not relevant. The distribution of values for the description of the list is statistically significant ($\chi^2(7) = 282.31$; $p < 0.001$).

On the contrary, the podium is described with opposite adjectives by almost the same number of subjects: it is "difficult to use" for 23% of subjects, "easy to use" for 28%; "unpleasant" for 18%, "pleasant" for 16%; "boring" for 21%, "amusing" for 15%. The only exception is the "comprehensible-incomprehensible" dimension, where 25% of subjects chose the first adjective and only 5% the second one. The observed values for the description of the podium are significant ($\chi^2(7) = 19.01$; $p < 0.01$).

Finally, "the medals" visualization is strongly and positively characterized on both the "pleasant-unpleasant" and the "amusing-boring" dimensions, with 36% of subjects describing it as “pleasant” and 29% as “amusing”; the distribution of values for the description of this visualization is also significant ($\chi^2(7) = 57.78; p < 0.001$).

As regards the “*absolute representation*” group, there is no clearly defined favourite visualization, since 39% of the subjects expressed their preference for “the sliders”, while 36% chose “the stars” and the remaining 25% the “tag cloud” ($\chi^2(2) = 5.1$; this value is not significant ($p > 0.05$). Notice, however, that users of social media, as our subjects are, are quite accustomed to expressing their preferences by means of stars and sliders, which in fact received most votes. Coherently, the most innovative interface in this group, the tag cloud, was indicated as the worst visualization by almost three quarters of the users (72%) and the distribution of values for the “least favourite” visualization is significant ($\chi^2(3) = 56; p < 0.001$).
User opinions. Both the sliders and the stars, similarly to “the list”, have very high values for the adjectives “easy to use” (chosen by 66.7% of users for the sliders and 77% for the stars) and “comprehensible” (63.5% for the sliders, 57.3% for the stars) and a significant agreement around the adjective “pleasant” - notice that the preferences collected by the sliders and the stars, 36.5% and 49% respectively, are higher than those of the list. Moreover, a few users also described these visualizations as “amusing”. The chi square values for the descriptions of both “the stars” ($\chi^2(7) = 194.54; p < 0.001$) and “the sliders” ($\chi^2(7) = 166.3; p < 0.001$) are significant.
On the contrary, for the “tag cloud”, opposite adjectives obtained almost the same number of preferences (in fact, chi square test is not significant), with the negative adjective prevail for all dimensions. The most unbalanced dimension is “easy to use-difficult to use”, with 38.5% of subjects choosing the negative adjective and only 19.8% the positive one; in contrast, the “amusing-boring” dimension is very balanced, with 28.1% of subjects choosing “boring” and 24% “amusing”, suggesting that this visualization, although considered difficult by most subjects, can prove engaging to some users.
Quite surprisingly, the favourite visualization in the “*relative representation*” group is the “pie chart”, with 47% of the preferences (the value distribution for the “favourite visualization” variable is significant with $\chi^2(2) = 17.67; p < 0.001$), while no clear winner can be identified, as far as the least favourite visualization is concerned: 44% of the users chose “the bricks”, 30% the “pie chart” and 26% the “coins”; chi square value for the “least favourite visualization” variable is equal to 6.07, while the critical value of the chi square distribution is 5,99; therefore, even if significant, it is too close to the critical value to be definitely considered ($\chi^2(2) = 6.07; p < 0.05$).
User opinions. The “pie chart” is described as “easy to use” (44.7% of subjects) rather than “difficult to use” (23.3% of subjects); pleasant (21.4%), rather than “unpleasant” (10.7%) and “comprehensible” (49.5%) rather than “incomprehensible” (3.9%). However, it is considered “boring” (30.1%), suggesting that the input modality, which forces users to correctly define the different percentages so that they sum up to 100, is too demanding. The observed values for the description of the pie chart are significant ($\chi^2(7) = 79.84; p < 0.001$).
The coins are positively assessed on all dimensions, even if they are judged a little less “comprehensible” (41.7% of subjects) in comparison with the “pie chart”; the observed values for the description of the “coins” are significant ($\chi^2(7) = 118.73; p < 0.001$).
“The bricks” are described as “difficult to use”, “unpleasant” and “boring”, the distri-

bution of values for this description is also significant ($\chi^2(7) = 70.8; p < 0.001$). Finally, almost all users declared to prefer a system where they can access their user model compared to a "traditional" one ($\chi^2(1) = 143.36; p < 0.001$) and that they would like to inspect and modify their preferences also in their everyday usage ($\chi^2(1) = 152.87; p < 0.001$).
A correlational analysis was also performed in order to discover correlations between demographic features and user preferences in visualization. However, no significant correlations were found, disconfirming our hypothesis of a relation.

3.2 Empirical evaluation

The second evaluation aimed at gaining a deeper insight about i) user preferences in specific visualizations; and ii) their opinion about the possibility to inspect and modify their models. With respect to the first experiment, we also have the goal to investigate iii) which type of user model representation (ordered, absolute or relative) is the most meaningful and user-friendly.

Hypothesis. In comparing different user model representations, we hypothesized that "relative" representations would be considered more difficult, but also more informative. The easiest-to-use visualizations should be those based on the "absolute representation", which is normally used for the externalization of user models. As far as goals i) and ii) are concerned, we expect to confirm the results of the previous experiment.

Experimental design. Multiple factors (user model representations, visualizations) within-subjects design.

Subjects. We selected a group of 28 subjects, 16-45 years old, 12 females and 16 males, among colleagues and students at the Computer Science Department, University of Turin, according to an availability sampling strategy[5]. All subjects were frequent Internet users, familiar with social media.

Measures and material. We measured user opinions by means of an on-line questionnaire. Oral comments were elicited through *thinking aloud* technique. Both the subjects' comments and their performance were recorded by means of a screen capture software, as a support for thinking aloud. The nine visualizations were made available online and shown to the subjects by means of a laptop computer.

Experimental tasks. The experiment, which took approximately twenty minutes to each subject, was carried out in a laboratory at the University, one subject at a time. After being welcomed, subjects were invited to sit in front of the computer, where they could read a short thank-you message and a text with the same instructions of the first experiment. Specifically, users were invited to read and modify their preferences with the proposed interfaces, "thinking aloud" if they felt comfortable with it. Also in this case, they were informed that they would be asked to fill in an anonymous questionnaire.

After that, subjects could autonomously access all the nine visualizations. These were clearly divided into the three groups (ordered, absolute, relative representation) and each visualization was displayed in a separate page. Notice that the experimenters

[5] Notice that, even though non-random samples are not statistically representative, they are often used in much psychology researches, as well as in usability testing, especially in early evaluation phases [16]

carefully observed the users, while they were interacting, without providing any explanations or suggestions, unless they were explicitly questioned.

Finally, subjects accessed an extended version of the previous questionnaire. In particular, 9 further questions were added, aimed at assessing the *task* of "reading and modifying one's preferences" by means of each visualization: they were based on 4-point Likert scales ("very easy", "easy", "difficult", "very difficult"). Notice that no intermediate, neutral option was provided, in order to force the subjects to express a precise opinion. Users were also asked to choose which type of user model representation (ordered, absolute, relative values) was the most meaningful to them.

Results. As far as the best visualization is concerned, users indicated "the stars" (25%), which were never mentioned as the least favourite one, either. "The list" (18%), "the podium" and "the medals" (both 14%) follow (See Table1). Although these data seem to suggest appreciation for well-known, commonly used visualizations and confirm the evidence collected in the first experiment, the chi square test relative to the distribution of values for the favourite visualization is not significant.
The least appreciated visualizations were "the pie chart" and "the tag cloud", both with 28.6% of votes. However, they were still indicated as the best visualization by 10.7% and 7.1% of users, respectively.

	list	medals	podium	cloud	stars	sliders	pie chart	bricks	coins
Favourite	5	4	4	3	7	2	2	0	1
Least favourite	2	5	0	8	0	2	8	2	1

Table 1. Distribution of values for the favourite and least favourite visualizations

	list	medals	podium	cloud	stars	sliders	pie chart	bricks	coins
Very difficult	1	1	0	0	0	2	1	10	0
Difficult	0	8	7	9	0	14	15	11	4
Easy	11	14	14	10	9	7	10	5	18
Very easy	16	5	7	9	19	5	2	2	6

Table 2. Distribution of values for the task evaluation

An analysis of the comments collected by thinking aloud highlighted that directly manipulating shapes in order to change their size, as in "the tag cloud" visualization, is considered intuitive by some subjects, but not precise enough, according to others (in particular, it seemed difficult to correctly perceive and manage the possible small differences among similar-sized objects). On the other hand, the "pie chart" allowed a very fine-grained control, which was appreciated by some users, but also seemed too cumbersome to others. In addition, remember that "the pie chart" emerged as the favourite visualization in the "relative values" group, when it was evaluated with more users, in the first experiment. Finally, it is interesting to notice that "the bricks" visualizations, which only 7.1% of users indicated as the worst, was never mentioned as the favourite. The observed data about the "least favourite visualization" are statistically significant ($X^2(8) = 25,36; p < 0,01$).

Task evaluation. All users judged the task of reading and modifying their preferences with "the stars" as either "easy" or "very easy" and the observed data for the corresponding variable are significant ($\chi^2(3) = 31.14; p < 0.001$) - see Table 2. On the contrary, both "the bricks" and "the tag cloud" were judged as "difficult" or "very difficult" to use by more than a half of the subjects ($\chi^2(3) = 19.14; p < 0.001$ and $\chi^2(3) = 11.14; p < 0.05$, respectively). These two more"innovative" visualizations, which had received negative feedback also in the first experiment, may actually have appeared as more difficult to use in comparison with a standard, familiar interface such as that provided by "the stars". Moreover, thanks to direct observation and thinking aloud, we can notice that an additional difficulty may have been caused by the some drag-and-drop mechanisms we used to implement these visualizations, which resulted unfamiliar to some subjects (direct manipulation was introduced on the web with AJAX is not yet a standard). Such hypothesis is confirmed by the fact that also the podium, which makes use of the same drag-and-drop mechanism, was considered "difficult" by a quarter of the subjects - the distribution of values for the task evaluation are significant also for this visualization ($\chi^2(3) = 14; p < 0.01$)

User opinions. An analysis of the adjectives used to describe the visualizations confirms our idea that simplicity, ease of use and familiarity are fundamental in determining the subjects' preferences. Both "the stars" and "the list", the two most appreciated , are in fact described as "easy to use" and "comprehensible" by most users, as in the first experiment. It is interesting to notice that they are also considered "pleasant", but by far less subjects (14% of the subjects for the list, 18% for the stars). On the contrary, only 5% of subjects describe "the stars" as amusing, while this adjective is never used for "the list": apparently, this feature is less relevant than ease of use. The value distributions relative to the user opinions, both for "the list" ($\chi^2(7) = 101.45; p < 0.001$) and for "the stars" ($\chi^2(7) = 67.9; p < 0.001$) are significant.

Notice that "the tag cloud", although considered "difficult to use" by half the subjects, "boring" by 11.8% and "comprehensible" only by 7.8%, is also described as "amusing" by 13.7% people (one of the highest scores for this dimension): the corresponding value distribution is quite uniform and the χ^2 is not significant.

The "pie chart" is considered very "difficult to use" (33.3% subjects), "unpleasant"' (14%) and "boring" (21%); however, it scores well (12 preferences) as far as comprehensibility is concerned, suggesting that most problems are related to the cumbersome input modality, as previously hypothesized. In this case, the observed values for the user opinions are significant ($\chi^2(7) = 45.03; p < 0.001$).

Preferred user model representation. Subjects favoured "ordered representation" (46.4%) - the only group which contained no strongly disliked visualizations and a very successful one, i.e. "the list" - followed by "absolute representation" (32.2%)- the group containing "the stars", the most appreciated visualization, but also "the tag cloud", which was much criticized. The "relative representation" only obtained 21.4% of preferences, since this group contained two visualizations which were much criticized, “the bricks” and, in particular, the “pie chart”. Notice that this result partially disconfirms our hypothesis, since we supposed that users would favour the “absolute representation” as for ease of use and the “relative representation” as for its capacity to express rich information. “Ordered values” are the simplest user model representation, so the visuations belonging to this group probably require the least effort and time to users. However, the observed values for the preferred user model representation are not statistically significant ($\chi^2(3) = 2.64$). Finally, we notice that almost all users evaluated the possibility of inspecting and modifying their preferences in a positive way ($\chi^2(3) = 35.14$; $p < 0.001$), also declaring that they would prefer a system which offers such functionalities to a “traditional” one ($\chi^2(1) = 11.57$; $p < 0.001$) and that they would examine and correct their preferences also in their everyday usage ($\chi^2(1) = 17.28$; $p < 0.001$).

4 Conclusion

In this paper we described two experiments, and their results, aimed at verifying i) which visual metaphors used to present user models is more comprehensible for final users, given a specific user model representation, and ii) whether the “relative representation” is more informative than “absolute” and “ordered representation”, even if more cumbersome. Regarding the *first experiment*, coherently with our hypothesis, the preferred visualizations are those which are commonly used in social websites, such as the stars and the sliders, for the absolute representation, and the list for the ordered representation. However, for the relative representation, the favourite visualization is the pie chart, disconfirming our *hypothesis* that users would prefer an easy-to-use, direct manipulation-based visualization, such as the coins. The pie chart is more demanding, but also more complete. Especially it allows more precise comparison between the values. Regarding the *second experiment*, our findings about the preferred user model representation are not significant, thus we will replicate this experiment with a larger sample, and only exploiting the visualizations that have obtained more success in the first experiment. This experiment was performed with a small user sample in order to collect a deeper insight in user opinions, which can be better reached through face-to-face interaction and with methods such as *direct observation* and *thinking aloud*. Comments collected through thinking aloud were particularly useful in order to confirm the idea, emerged in the first experiment, that the absolute representation, to which users are quite accustomed, is easy to understand and to use. However, the ordered representation is considered even easier. On the other hand, some users appreciated the visualizations based on the relative representation (the pie chart in particular), because they were more precise and allowed them to explicitly indicate relations among different categories. Noticed that in this experiment the pie chart was also indicated by a lot of users as least favourite visualization, so probably this visualization cause contradictory opinions. Therefore, our idea that the relative representation is more informative has been partially confirmed. We plan to conduct a further experiment with a larger sample, with the goal of statistically confirming these results. Notice that some of our findings may have been influenced by the specific interaction techniques we proposed. Some visualization allows direct manilation, which has been introduced only recently in the web, and can be therefore unfamiliar to some users. However the list, which in the second experiment obtained a lot of preferences, allow direct manipulation. Probably in this case the kind of interaction proposed, even if it is not a standard interaction, is more intuitive than others. Thus this results suggest to carefully implement drag-and-drop mechanisms on the web.

To conclude, we must remember that what is important is that the user might understand the model. Consequently, if a system makes use of a complex internal representation, such as a relative one and if an effective and easy-to-use visualization cannot be designed based on such representation, the choice of a simple absolute or ordered representation, as far as externalization is concerned, can be the most appropriate.

References

1. S. Bull and Kay J. A framework for designing and analysing open learner modelling. In *Workshop on Learner Modelling for Reflection*, 2005.
2. S. Bull, M. Mangat, A. Mabbott, A.S. Abu Issa, and J. Marsh. Reactions to inspectable learner models: Seven year olds to university students. In *Proceedings of Workshop on Learner Modelling for Reflection, AIED*, pages 1–10, 2005.
3. S. Bull and M. Mckay. An open learner model for children and teachers: Inspecting knowledge. In *Level of Individuals and Peers, Intelligent Tutoring Systems: 7th Int. Conf*, pages 646–655. Springer-Verlag, 2004.
4. S. Bull and H. Pain. Did I say what I think I said, and do you agree with me? Inspecting and questioning the student model. In *Proceedings of AI in Education, AACE,*, pages 501–508, 1995.
5. F. Carmagnola, F. Cena, L. Console, O. Cortassa, C. Gena, A. Goy, I. Torre, A. Toso, and F. Vernero. Tag-based user modeling for social multi-device adaptive guides. *User Model. User-Adapt. Interact.*, 18(5):497–538, 2008.
6. H.S.M. Cramer, V. Evers, S. Ramlal, M. van Someren, L. Rutledge, N. Stash, L. Aroyo, and B.J. Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model. User-Adapt. Interact.*, 18(5):455–496, 2008.
7. V. Dimitrova. Style-olm: Interactive open learner modelling. *International Journal of AI in Education*, 13(1):35–78, 2003.
8. Kay J. Learner know thyself: Student models to give learner control and responsibility. In *AACE, International Conference on Computers in Education*, pages 17–24, 1997.
9. J. Kay. *A scrutable user modelling shell for user-adapted interaction*. PhD thesis, Basser Department of Computer Science, University of Sydney, Sydney, Australia, 1999.
10. J. Kay. Scrutable adaptation: Because we can and must. In *AH*, pages 11–19, 2006.
11. J. Kay, R. J. Kummerfeld, and P. Lauder. Personis: a server for user models. In *AH'2002, Adaptive Hypertext 2002*. AH'2002, Adaptive Hypertext 2002, 2002.
12. L. Li and J. Kay. Assess: promoting learner reflection in student self-assessment. In *Proceedings of Workshop on Learner Modeling for Reflection, to Support Learner Control. Metacognition and Improved Communication*, pages 32–41, 2005.
13. S. Mohanarajah, R. Kemp, and E. Kemp. Opening a fuzzy learner model. In *Proceedings of Workshop on Learner Modelling for Reflection, AIED*, page 6271, 2005.
14. D. Norman. *The Design of Future Things*. New York: Basic Books, 2007.
15. K.A. Papanikolaou, M. Grigoriadou, H. Kornilakis, and G.D. Magoulas. Inspire: an intelligent system for personalized instruction in a remote environment. In *In Proceedings of 3 rd Workshop on Adaptive Hypertext and Hypermedia*, pages 13–24, 2001.

16. Singleton A. Royce and Bruce C. Straits. *Approaches to Social Research (3rd Edition).* New York: Oxford University Press., 1999.
17. G. Weber and P. Brusilovsky. Elm-art: An adaptive versatile system for web-based instruction. *International Journal of AI in Education*, 12(4):351–384, 2001.
18. J.D. Zapata-Rivera and J. Greer. Construction and inspection of learner models. In *Computer Support For Collaborative Learning CSCL 2002*, pages 495–497, 2002.
19. J.D. Zapata-Rivera, J.S. Underwood, and M. Bauer. Advanced reporting systems in assessment environments. In *Proceedings of Workshop on Learner Modelling for Reflection, AIED*, pages 23–32, 2005.

Retrospective Evaluation of Blended User Modeling For Adaptive Educational Systems

Michael Yudelson, Sergey Sosnovsky

University of Pittsburgh, School of Information Sciences, 135 N. Bellefield Ave.,
15260 Pittsburgh, PA, USA
{mvy3, sas15}@pitt.edu

Abstract. In this paper, we are presenting a retrospective approach to evaluating user models by utilizing previously collected learning logs rather than setting up a new experiment. This approach is applied in a novel way to modeling heterogeneous types of user activity – problem solving, and browsing annotated examples. We are blending the two types of activity in the user model in an attempt to increase the accuracy of the composite model. Obtained results suggest that such blending, in fact, does make a difference both for users individually and on a global scale.

Keywords: user modeling, evaluation, model blending, adaptive educational systems.

1 Introduction

The best way to determine the quality of an adaptive system is through a carefully planned empirical evaluation with human subjects. The evaluation design can vary from a short-term controlled experiment to a longitudinal study, but before the system is put into use its value is rather unknown. The system under evaluation is usually considered as a black box, that influences the depended variable as a whole. However, it is not always clear what do we really measure when evaluating the quality of an adaptive system. The effect or the value of adaptation observed in such experiments can be attributed to several things: the accuracy of user modeling, the effectiveness of adaptation strategies, or the quality of the content.

One of the known alternatives to the holistic view on adaptive system studies is layered evaluation [1, 2]. It implies that the user modeling component and the adaptation component of an adaptive system are assessed independently. The evaluation of a user modeling component is based on its accuracy, or predictive validity, which defines how well the model represents the actual state of the user and how reliably it can predict user's next action [3]. In the context of adaptive education, it can be interpreted as the model's ability to predict the result of the student's next attempt to apply a concept or answer a problem.

An interesting opportunity that this approach opens for experimenters is the implementation of several modeling algorithms operating on the stored log of users' activity and comparative evaluation of these algorithms based on their predictive validity. Such retrospective analysis allows the reuse of once collected data for multiple evaluation experiments based on “what-if” scenarios aimed at pre-selection of an optimal user modeling approach [13]. Naturally the optimality of such pre-selection is limited to the user modeling layer. The presence of adaptation that is based on the values supplied by the user model, would add an additional factor. An overall cross-layer empirical evaluation would be necessary to make a final assessment.

In this paper, we apply retrospective evaluation to choose the best value for a singe parameter in the modeling formula. The data set is the log of students’ learning activity with two types of education content. The user modeling algorithm used this log to populate overlay models of students’ knowledge. However, different types of activity were processed independently to compute parallel student models on two different cognitive levels: comprehension level (corresponding to example-browsing activity) and application level (problem-solving activity). Our main goal is to find whether a blending of the user models that correspond to the two cognitive layers can result in a better composite model with higher predictive validity.

The rest of the paper is organized as follows. Section 2 talks about the original approach to building user models from cognitively heterogeneous educational activity. Section 3 discusses user modeling without blending. Section 4 proposes a modification to the modeling approach and introduces blended user modeling. Section 5 outlines the hypotheses and goals of this experiment, which is presented in Section 6. Finally, section 7 concludes the paper with an extended discussion of the obtained results.

2 Modeling From Heterogeneous Student Activity

Many e-learning environments provide students with various types of educational content (learning problems, examples, tutorials, interactive simulations, etc.) that contribute to different levels of material understanding. Several adaptive systems integrate or provide means for integrating such components (e.g. [4, 5]). One of the problems for these systems is to incorporate evidence coming from heterogeneous sources into a student model that would help to deliver viable adaptation. Our previous solution was not to fuse these activities, essentially, maintaining a set of parallel models of student knowledge, each populated by a specific kind of learning activity. The levels of student modeling where taken from the Bloom’s taxonomy of educational objectives [6]. For example, reading a textbook would contribute to the “knowledge” level of the Bloom’s taxonomy; exploring examples – “comprehension” level; answering problems and quizzes – “application” level, etc. However this approach does not take into account the transfer between the categories of the Bloom’s taxonomy: mastering a lower level of activity should also influence the higher level(s).

Over the last several years, we have accumulated a rich collection of user activity logs from student of several undergraduate and graduate level courses using a set of our systems in a number of learning domains. A tangible portion of the logs covers problem solving and browsing annotated examples that correspond to the application and comprehension levels of Bloom’s taxonomy. A question for this study is whether modeling the transfer between different cognitive levels of the user model (in this case, the comprehension and application levels) can be quantitatively detected, i.e. whether this transfer would improve the accuracy of our user models. We try to explore this effect by combining or *blending* different tiers of the user model retrospectively and re-evaluating each blend by computing the prediction validity of the composite user model.

3 User Modeling Without Blending

There is an abundance of approaches to user modeling. A great number of them follow the overlay paradigm, when a user model is calculated with respect to a set of concepts, skills, or preferences. The user modeling component processes evidence of a user's interaction with a content item and updates relevant portions of the overlay vector, spanning the domain. One such approach has been implemented in the user modeling server CUMULATE [7].

CUMULATE builds several types of user models resulting from different types of user activity. The ones that are of interest to our discussion here are: the model of example browsing (the comprehension level of Bloom's taxonomy), and the model of problem solving (the application level of Bloom's taxonomy). For each of the models, CUMULATE uses a different technique to compute knowledge levels. In the case of example browsing, CUMULATE tracks percent of example lines explored. When that percentage reaches 80%, all of the concepts relevant to this example are considered known (on the comprehension level).

Modeling problem solving in CUMULATE is done in a more complicated way. Each of the concepts with which a problem is indexed, has a weight. This weight is produced during indexing and denotes the importance of that concept in mastering the problem. Concept weights are used in distributing the total amount of updates a user model receives. CUMULATE also has a safety mechanism discouraging users from over-practicing one particular exercise. This over-practicing gradually decreases the knowledge updates when users solve one particular problem correctly more that one time. Thus users are motivated to attempt solve a diverse set of problems in order for their user models to grow. Refer to equations (1) and (2) for details.

$$k_{n+1} = k_n + res \cdot (1-k_n)^2 \cdot \begin{cases} k_n \le .5 & w/2 \\ k_n > .5 & w \end{cases} \tag{1}$$

$$w = \sqrt[4]{\frac{w_{c,p}}{\sum_i w_{c_i p} \cdot \left({}_{succ}att_p + 1\right)}} \tag{2}$$

The initial value of a concepts knowledge k_0 is 0. With every correct solution of the problem (where *res*=1 in (1)), all of the related concepts receive an update. This update is directly related to: a) the amount of knowledge this concept can grow by squared ($(1 \ -k_n)^2$ in (1)), and b) to a special weighting factor (2). This weighting factor is composed of weights ratio and over-practicing penalty. Weights ratio is the weight of the currently updated concept in the problem ($w_{c,p}$ in (2)) over the sum of all weights involved in the problem ($\Sigma_\mathrm{i}\ w_{c,p}$). The problem's over-practicing penalty is one over number of successful solutions to this specific problem by a particular user plus one ($_{succ}att_p$ + 1). When the prior knowledge level is below 50% the weighting factor is halved (1). This is done to prevent initial leaps in knowledge level.

4 Blending Problem-Solving And Example Exploration

Over several years we collected user activity and modeling user knowledge in CUMULATE. We noticed that, while practicing problem solving does provide a faster way to acquire knowledge, users do spend significant time reviewing annotated examples. This suggests that examples are in fact an important part of learning and that there may be a better way to incorporate example browsing into computing the user model than the one we have described in the previous section.

Intuitively there should be some form of transfer between comprehension and application tiers of the user model. There might not be direct impact, of course, as problem solving requires deeper understanding of the domain than mere clicking and looking could hope to achieve. However, a limited influence of example browsing is not at all impossible.

We have modified equation (1) to reflect the possible comprehension-to-application level transfer. Refer to equation (3). The only difference is a B weight. This weight is 1 for problem solving, making equation (3) identical to equation (1). In the case of example browsing, B would constitute a blending coefficient: value from 0 to 1. 0 – meaning no blending whatsoever – without considering example browsing, and 1 – meaning example browsing is as important as problem solving. Other than the B weight, the updates to the knowledge level of the concepts are done in the same manner on the unified problem- and example-related user model.

$$k_{n+1} = k_n + res \cdot B \cdot (1 - k_n)^2 \cdot \begin{cases} k_n \le .5 & w/2 \\ k_n > .5 & w \end{cases}, \tag{3}$$

After some experimentation, we found that in addition to blending coefficient we should take into account the amount to which the example was explored. Truly, we cannot equally consider user activity in case the example is fully explored and when only say 1 out of ten lines were reviewed. To take that into account, for examples-related activity modeling we have decided to define B in equation (3) as a product of blending coefficient and percentage of example lines explored.

5 Hypotheses And Goals

Our hypotheses regarding blending comprehension and application layers of user mode are the following.

1. In general, blending example activity (evidence of concepts' comprehension) and problem solving (evidence of concepts' application) increases the accuracy of user modeling.
2. Different users benefit from different blends.

The goals that we are trying to reach in this study are.

1. Find a universally optimal blend of comprehension and application levels in the user model, if such exists.
2. If possible, determine and describe groups of users that can benefit from different blending conditions.

6 Experiment

6.1 Experimental Setup

To evaluate our hypotheses and meet our goals regarding blending layers of the user model belonging to different levels of Bloom's taxonomy, we have set up a computational experiment. We used student activity logs that were collected during Fall 2007 and Spring 2008 semesters from 4 database design courses offered at both the University of Pittsburgh (1 graduate and 2 undergraduate courses), and Dublin City University (1 undergraduate). All 4 courses, although slightly different in structure, were roughly identical with respect to the content. Each course consisted of a set of topics. Every topic had a set of SQL writing problems provided by SQL KnoT system [8] and a set of annotated SQL code examples supplied by the system WebEx [9]. Both SQL KnoT and WebEx were introduced to students roughly in the beginning of each of the semesters. The use of these systems was optional and did not impact the students' grades. Overall, there were 48 problems and 64 examples available to the students.

The number of students, as well as their level of participation, varied across semesters and is summarized in Table 1 along with basic usage statistics.

Table 1. Basic user participation statistics across semesters and courses.

School	Semester	Level	No. of users	Avg. problem attempts	Avg. example views	Avg. distinct problems	Avg. distinct examples
U. of Pitt	Fall 2007	U*	27	156.40	189.00	29.96	32.07
U. of Pitt	Fall 2007	G	20	61.70	104.70	29.95	29.10
U. of Pitt	Spring 2008	U	15	26.94	46.65	16.35	10.29
DCU	Spring 2008	U	52	81.68	257.25	22.82	38.63

* U – undergraduate, G – graduate

All student activity with both problems (SQL KnoT) and examples (WebEx) has been logged by the CUMULATE user modeling server. Each problem and example has been indexed with a set of metadata concepts with the help of a semi-automatic grammar parser. The concepts came from an SQL ontology, developed by domain experts. The indexes were double-checked afterwards.

6.2 Experimental Procedures

For each of the semester logs, we have (re)-computed several blended user models. First of all, a 0-blend was computed; here, no example activity was taken into account – only problem solving activity was modeled. 0.1, 0.2, … 0.9, and 1.0 blends corresponded to user models where updates resulting from example activity were weighted from 0.1 to 1.0 with 0.1 steps. This gave us 4 semesters * 11 blends = 44 clusters of user models or 114 users * 11 blends = 1254 user models. A classical accuracy measure (correct predictions over all predictions) was computed for each user model.

Prior to proceeding with testing of our hypotheses, we filtered user models. The filtering condition was that the user had to attempt to solve at least 33% of the problems (15 out of 48) and view at least 33% of examples (22 out of 64). The reason behind this threshold was that, in order to improve problem solving model by blending it with example browsing model, both have to be well populated. Namely, the user had to work with both examples and problems to a significant extent.

After the filtering, the number of users in each semester/class dropped to the values shown in Table 2. Thus, the initial number of 114 users was reduced to 56 users.

Table 2. Number of qualified users after applying filtering.

School	Semester	Level	No. of users	No. of qualified users
U. of Pitt	Fall 2007	U*	27	14
U. of Pitt	Fall 2007	G	20	10
U. of Pitt	Spring 2008	U	15	3
DCU	Spring 2008	U	52	29

* U – undergraduate, G – graduate

6.3 Results

To get a general idea about the usefulness of blended models for each user, we have selected the best non-0% blend (10% to 100%) and ran a left-tailed paired *t*-test. Individual best blends turned out to be significantly better then 0% blends with $t = -5.38$, *p*-value<.001. The average edge of each student's best blend over 0% blend was .015 or 1.5% in terms of accuracy. Mean standard deviation of blended model accuracies across users was .0113 or 1.13%. The minimum standard deviation was 0% and the maximum was 10%.

To select a universal useful blend we ran 10 left-tailed paired *t*-tests, in each case comparing 0% blend to one of 10 non-0% blends. Here, 40% and 50% blends turned out to be the most potent ones and the only ones with significant edge over 0%-blend (both with t = -2.05 and p-value = .023). The average advantage of 40% and 50% blends over 0% blend dropped to .56%. As we can see, "universal" blends lose to individually tailored blends.

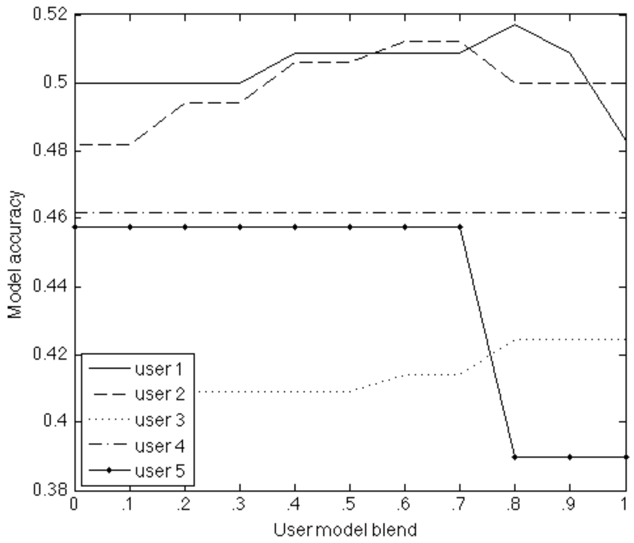


Fig. 1 Examples of blend effect on user model accuracy.

Before further exploring individual user differences with respect to blends, let us refer to Fig. 1, where 5 sample users are represented with a graph of blending percentage vs. accuracy. Here we can see that the model of user 4 is not sensitive to blending whatsoever: the accuracy does not change with respect to blends. In the case of user 5, blending has no effect till 70% blend after which accuracy drops. Blending does help to improve user models for users 1,2, and 3.

One feature of the blended models apparent in Fig. 1 is that different users have different numbers of points of maximum accuracy. Graph of user 4 is flat, giving us 11 points of maximum (or no maximum at all). User 5 has 7 points of maximum, and users 1, 2, and 3 have 1, 2, and 3 points of maximum accuracy respectively. Fig. 2 shows the distribution of the number of maximum accuracy points for blended models of all 56 users.

Instantly, we can notice a group of "no difference" consisting of 15 users for which blending doesn't improve the user model. The rest of the range of the number of maximum blends can be subdivided into the "low" group (1 maximum) of 2 users, the "medium" group (2-4 maximums) of 22 users, and the "high" group (5-9 maximums) of 17 users.

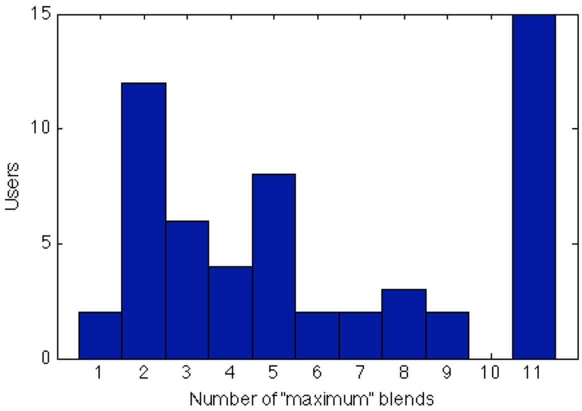


Fig. 2 Distribution of number of users for different numbers of peak (maximum) blends.

The “low” group consists of the two rare cases of a user having just one best blend. Both users prefer high blends of 80% and 100% respectively. Users in the “medium” group have an inclination towards higher blends. Since our data did not meet the requirements of the parametric test (paired *t*-test), we used its non-parametric analog Wilcoxon signed-rank test. Out of 10 tests the most potent belongs to 90% blend with *p*-value = .037.

Users of the “high” group follow the global trend. Out of 10 Wilcoxon signed-rank tests the ones corresponding to 40% and 50% blends turn out to be equally significant. Both with *p*-value = .049.

7 Discussion

We are able to see from the data that blending comprehension and application tiers of user model in fact does make a difference both for users individually and on a global scale. Namely, there is a benefit in (partially) scoring example browsing as if it was problem solving, and there is a transfer effect between cognitive layers of the model. The major downside is that, although statistically significant, the difference is quite small: on the order of few percent.

Nevertheless, there is a clear indication that, with respect to blends, users do differ in what blend works best for the higher accuracy of their model. We also believe that there is a way to pinpoint both individual and global blending effect better.

One of potential ways to improve is to contextualize the model. As described in Section 3, modeling in CUMULATE follows the *one-fits-all* schema. However, as it has been shown in [10] each item of the problem space, as well as each user, possess individual features. With respect to problems, each has its inherent complexity not always captured by the metadata index. Knowledge of concepts does not grow equally fast for all of them and does not always starts from same value (0 in our case).

Making appropriate adjustments in user modeling to accommodate these differences has a chance to improve the modeling itself and help to find an optimal blend of Bloom's user model tiers both on global and individual scale.

Another issue with an exploration of the blending effect is that we had to filter nearly 50% of the users out. Ideally, for the blending to have a tangible effect, both example browsing and problem solving behaviors have to be well established: the user has to work enough with both types of learning resources.

A prospective remedy here could be to shift from number of distinct learning resources covered to the amount of metadata overlap. Instead of counting how many examples were viewed or problems were attempted, it might be more beneficial to trace the overlap of the domain concepts that both examples and problems addressed.

One important thing to mention is that in all of the reported studies some form of adaptive navigation support was available to users and this could potentially have affected our measurements. The navigation support was expressed in the form of a descriptive icon next to the link that opened an example or a problem.

An aspect that still remains unaddressed is the temporal dimension. It might be the case that the optimal blending of the user model layers is not persistent over time. As users progress through the course, the best blend may change for them. It would be challenging to detect these changes, as users would have to stay very active for the whole duration of the course and generate enough log data to analyze. From our own experience, the proportion of such motivated users is very low in every class and often they are outstanding in various regards: both in positive and negative sense.

For our future work, we would like to apply the blending of cognitive layers of the user model in a longitudinal study. This might help us to see a clearer differentiation between blending factors and assist in making cognitive layer blending preferences explainable more transparent.

Also we would like to test our blending approach in different learning domains such as learning C or Java. In addition, we would like to test other approaches to user modeling such as knowledge tracing [11] and/or learning factor analysis [12].

Acknowledgments. We would like to express special thanks and appreciation to Peter Brusilovsky for support, insights, and valuable feedback provided as this work took shape.

References

1. Paramythis, A., & Weibelzahl, S. (2005). A Decomposition Model for the Layered Evaluation of Interactive Adaptive Systems. In Ardissono, L., Brna, P., & Mitrovic, A. (Eds.), Proceedings of the 10th International Conference on User Modeling (UM2005), Edinburgh, Scotland, UK, July 24-29 (pp. 438-442) (Lecture Notes in Computer Science LNAI 3538, Springer Verlag). Berlin: Springer.
2. Brusilovsky, P., Karagiannidis, C., & Sampson, D. (2004). Layered evaluation of adaptive learning systems. International Journal of Continuing Engineering Education and Lifelong Learning, 14(4/5), 402-421.
3. Corbett, A.T., Anderson, J. R. and O'Brien, A. T.: 1993, The predictive validity of student modeling in the ACT programming tutor. In: P. Brna, S. Ohlsson and H. Pain (eds.),

Artificial Intelligence and Education, 1993: The Proceedings of AI-ED 93. Charlottesville, VA: AACE.

4. Denaux, R., Dimitrova, V., and Aroyo, L. 2005. Integrating Open User Modeling and Learning Content Management for the Semantic Web. In L. Ardissono, P. Brna & A. Mitrovic (eds.), Proceedings of 10th International Conference on User Modeling (UM'2005), Edinburgh, Scotland, UK, 23-29 July (pp. 9-18).
5. Trella, M., Carmona, C., and Conejo, R. 2005. MEDEA: an Open Service-Based Learning Platform for Developing Intelligent Educaional Systems for the Web. In Proceedings of Workshop on Adaptive Systems for Web-Based Education: Tools and Reusability at AIED'05, Amsterdam, The Netherlands (pp. 27-34).
6. Bloom, B. S. (1956). Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain. New York: David McKay Co Inc.
7. Brusilovsky, P., Sosnovsky, S. A., & Shcherbinina, O. (2005). User Modeling in a Distributed E-Learning Architecture. In: L. Ardissono, P. Brna, & A. Mitrovic (Eds.), 10th International Conference on User Modeling (UM 2005), Edinburgh, Scotland, UK, 2005 (pp. 387-391). Springer.
8. Brusilovsky, P., Sosnovsky, S. A., Lee, D. H., Yudelson, M., Zadorozhny, V., & Zhou, X. (2008). An open integrated exploratorium for database courses. In: J. Amillo, C. Laxer, E. M. Ruiz, & A. Young (Eds.), 13th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE 2008), New York, NY, USA, 2008 (pp. 22-26). ACM.
9. Brusilovsky, P. (2001). WebEx: Learning from Examples in a Programming Course. In: W. A. Lawrence-Fowler & J. Hasebrook (Eds.), World Conference on the WWW and Internet (WebNet 2001), Orlando, Florida, 2001 (pp. 124-129). AACE.
10. Corbett, A. T. & Anderson, J. R. (1992). Student Modeling and Mastery Learning in a Computer-Based Programming Tutor. In: C. Frasson, G. Gauthier, & G. I. McCalla (Eds.), 2nd International Conference On Intelligent Tutoring Systems (ITS'92), Montréal, Canada, 1992 (pp. 413-420). Springer.
11. Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253-278.
12. Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement. In: M. Ikeda, K. D. Ashley, & T. Chan (Eds.), Intelligent Tutoring Systems, Jhongli, Taiwan, 2006 (pp. 164-175). Springer.
13. Yudelson, M. V., Medvedeva, O., & Crowley, R. S. (2008). A multifactor approach to student model evaluati*on. User Modeling and User-Adapted Interaction, 18(4), 349-382.

Assessing the Effectiveness and Usability of Personalized Internet Search through a Longitudinal Evaluation

Lex van Velsen [1], Florian König [2], Alexandros Paramythis [2]

[1] University of Twente, Dpt. of Technical and Professional Communication
P.O. Box 217, 7500 AE Enschede, the Netherlands
l.s.vanvelsen@utwente.nl

[2] Johannes Kepler University
Institute for Information Processing and Microprocessor Technology (FIM)
Altenbergerstraße 69, A-4040 Linz, Austria
{alpar, koenig}@fim.uni-linz.ac.at

Abstract. This paper discusses a longitudinal user evaluation of Prospector, a personalized Internet meta-search engine capable of personalized re-ranking of search results. Twenty-one participants used Prospector as their primary search engine for 12 days, agreed to have their interaction with the system logged, and completed three questionnaires. The data logs show that the personalization provided by Prospector is successful: participants preferred re-ranked results that appeared higher up. However, the questionnaire results indicated that people would prefer to use Google instead (their search engine of choice). Users would, nevertheless, consider employing a personalized search engine to perform searches with terms that require disambiguation and / or contextualization. We conclude the paper with a discussion on the merit of combining system- and user-centered evaluation for the case of personalized systems.

1 Introduction

Attaining automatically personalized system behavior is, in many cases, a process that can not be considered "complete" at a certain moment in time. On the contrary, personalization often becomes effective only after a certain period of user-system interaction and even after that can be subject to constant changes, as user characteristics and interests may change and expand. In order to explore how these long-term changes affect the (perceived) usefulness of personalized output, longitudinal studies with a partial user focus need to be conducted [1]. However, most evaluations of personalized systems are short term and do not focus on the effects of continued use [2]. Furthermore, most often these evaluations take either a system-centered or a user-centered focus, while a combination of both yields the most valuable evaluation results [3]. System-centered evaluations focus on the quality of system algorithms (to be assessed by means of quality metrics), while user-centered evaluations center on users' subjective experience of their interaction with the system.

This paper discusses a longitudinal evaluation of a personalized search engine, which combined a system- and user-centered approach. The goal of this paper is threefold. First, we want to assess whether personalized Internet meta-search, as provided by Prospector, the system under evaluation, is effective and perceived as useful. The second goal, which is related to the first, is that to determine *why* the system is (regarded as) effective or not, and determine its usability. Third, we want to provide future personalized system evaluators with information that can help them to design their evaluation setup, by reflecting on the experiences we gained in this study.

The rest of this paper is organized as follows. Section 2 describes Prospector, the system which was the subject of this study. Section 3 presents the evaluation setup, followed by the user-centered evaluation results in section 4 and the system-centered ones in section 5. We wrap up this paper with our conclusions in section 6.

2 The Prospector System

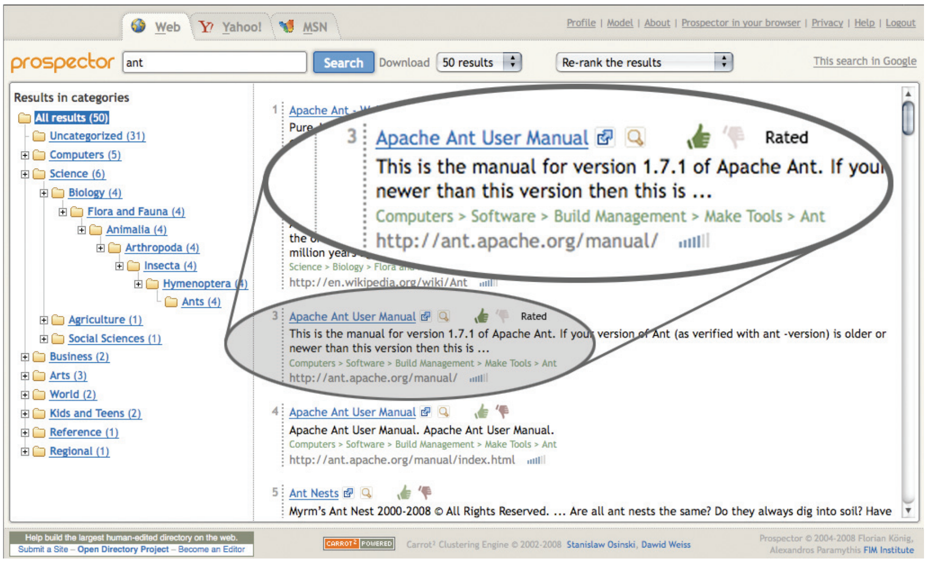
Prospector's personalization algorithm is based on the utilization of the Open Directory Project (ODP)[1] ontology, which provides semantic meta-data for classifying search results. Prospector uses taxonomies as overlays [4] over the ODP ontology for modeling user and group interests and bases the re-ranking of search results on said overlays. The operation of Prospector can be summarized as follows: the underlying search engine retrieves results for a user's query; results are classified into thematic topics using the ODP metadata; user- and group- models maintained by the system are used to determine an appropriate ranking of the results; users are presented with the re-ranked results, which they can rate on a per-result-item basis; the system uses these ratings to update individual and group models. User- and group- models are overlays containing the probability of an ODP topic being of interest to a user or group.

The version of Prospector discussed in this paper has been preceded by two other versions, described in [5, 6]. In this paper, we discuss the third version of the system, largely shaped by the results of an evaluation of the second version, reported in [7]. The most important new features in this version include a more stable ranking algorithm, better use of existing meta-data, and usability enhancements. The rest of this section provides a brief overview of interactive aspects of the system and the result re-ranking algorithm (for additional information please refer to [8]).

In order to get personalized search results users first have to register. At the first login they are asked to specify their interest in the 13 top-level ODP topics. It is explained to the users that this way they will benefit from the ratings of results by users with similar interests. Representative sub-topics are also listed for each topic, to help users form a better mental model of the area a topic covers.

For each search users may choose the underlying search engine to use by selecting the corresponding tab (see Fig. 1): Web (i.e., the www.etools.ch meta-search engine), Yahoo and MSN. Google was not included for technical reasons. When issuing a query this engine is accessed, its results are retrieved and classified (per the ODP ontology). The classification paths are displayed for each result, and the tree control on the left side of the results page lets users filter results by these topical categories.

[1] For information on the Open Directory Project please refer to: http://www.dmoz.org



ant

1 Apache Ant - W
Pure
the
million years
Science > Biology > Flora
http://en.wikipedia.org/wiki/Ant

3 Apache Ant User Manual Rated
This is the manual for version 1.7.1 of Apache Ant. If your version of Ant (as verified with ant -version) is older or newer than this version then this is ...
Computers > Software > Build Management > Make Tools > Ant
http://ant.apache.org/manual/

4 Apache Ant User Manual
Apache Ant User Manual. Apache Ant User Manual.
Computers > Software > Build Management > Make Tools > Ant
http://ant.apache.org/manual/index.html

5 Ant Nests
Myrm's Ant Nest 2000-2008 © All Rights Reserved. ... Are all ant nests the same? Do they always dig into soil? Have

Fig 1. Prospector's main interface.

Re-ranking of results works as follows: The first step is to calculate a relevance probability for each result item, composed from the interest probability of each ODP topic in which the result has been classified. If the user model does not contain an interest probability for a topic, the value is derived from more general topics in the user model, and from the group models. In a second step, the calculated relevance probabilities for each topic are combined into a weighted average. The affinity of the user to the corresponding group is used as the respective weight. In a third step, the relevance probability of each result is combined with its rank as returned by the underlying search engine (both normalized in the value space [0..1]. The normalized rank and score values are then combined by a weighted extension [9] of the *CombSUM* method [10]. Prospector uses this final value for re-ranking the result list accordingly.

By rating individual results positively (thumbs up) or negatively (thumbs down) users implicitly express their preference for certain topics. Quickly evaluating a result is possible by toggling an embedded preview below the result with the magnification glass icon; opening a result in a new window is facilitated by the arrow icon. When previewing or returning from an opened result, the user is notified / reminded of the possibility to rate that result, by pulsating the thumbs a few times.

Each rating modifies the appropriate user- and group- models, thus affecting the calculation of relevance probabilities of classified results in subsequent searches. To give users a way to quickly assess the system-calculated relevance of a result, its probability is visualized next to the URL by means of a "connection quality" indicator, as used in mobile phones. Hovering over the bars with the mouse shows the exact relevance percentage in a tool-tip.

For logged in users the ranking is by default personalized with respect to their user model and the models of associated groups. In addition, users can request that results be re-ranked using a particular group model (e.g., "re-rank for people interested in arts”). This feature is intended to let users focus on the specific high-level topic represented by the group, and is also available for anonymous, not logged in users.

The user models in Prospector are scrutable [11], allowing users to inspect, correct and fine-tune them, while at the same time strengthening their confidence in the system. Affinities to groups, as set when registering, can be changed at any time. The interests inferred from the ratings can be viewed and changed as well. Finally, entire topic “branches” can also be removed from the model, which gives users the means to purge old or invalid interest information.

3 Evaluation Setup

We asked 130 persons whether they were willing to use Prospector as their primary search engine for 12 days, to have their use with the system logged and, finally, to complete three questionnaires. Because the study could be considered privacy infringing, potential participants were informed they would remain anonymous at all times. Twenty-one persons responded positively, including nineteen men and two women, with an average age of 25.8 years (SD = 2.8). Most were students at the Johannes Kepler University in Linz, Austria. They rated their computer skills as high and used the Internet on a daily basis. All participants used Google as their primary search engine, with one third of them performing more than 15 searches a day, one third 2 to 15 searches a day, and the remaining third just a few searches a week.

Besides logging all actions performed with Prospector (as input data for the system-centered evaluation), we distributed a pre-questionnaire, a mid-questionnaire (after five days of use) and a post-questionnaire (after 12 days of use) as input for the user-centered evaluation. For economy of space we will refer to these questionnaires as “preQ”, “midQ” and “postQ” respectively. They addressed the following issues by means of open-ended questions (unless specified otherwise):

1. **Demographics**. In preQ we questioned participants’ demographics, internet use, experience with personalized systems and use of search engines.
2. **Expectations**. The preQ asked for the participants’ expectations of using Prospector and the expected usefulness of a scrutable user model.
3. **Perceived usefulness**. We asked the participants to score their agreement on three statements (7-point Likert scales) on the usefulness of Google (preQ) and on the usefulness of Prospector (midQ and postQ). The statements were based on the perceived usefulness scale of a search engine by Liaw and Huang [12].
4. **Comparisons between Prospector and Google**. In midQ and postQ, we asked the participants to compare the perceived quality of the results they received from Google and Prospector.
5. **Incidents**. We twice (midQ and postQ) asked the participants to describe incidents that made them satisfied or dissatisfied with Prospector.
6. **User modeling**. The transparent user model allowed us to ‘break up’ the evaluation of the personalization done by the system in two parts: user modeling and application of the algorithm. This layered evaluation approach makes it possible to pinpoint the cause of incorrectly personalized output more specifically [13]. Therefore, we asked the participants to inspect their model and questioned its clarity (midQ) and correctness (midQ and postQ).

7. **Usability**. In the postQ, we inquired the participants' experience in relation to usability issues of high relevance for personalized systems, as listed by Jameson [14]. Examples include system predictability, controllability and privacy.

4 User-Centered Evaluation Results

In this section we discuss the user-centered results according to the questionnaire elements, listed in section 3.

Expectations. Most participants expected Prospector to outperform Google, by reducing search times (six participants) and giving better results than Google (six participants). As one participant put it: "Hopefully I will quickly find information that in other search engines is only on the second or third page." Twelve participants were initially positive about the possibility to view and alter their model.

Perceived usefulness. The scale we used to measure perceived usefulness appeared to be very reliable (Cronbach's $\alpha = .95$). On a scale from 1 to 7 (where 7 is very useful), Google scored 6.05 (SD = .83). Prospector scored 3.71 (SD = 1.66) in midQ and 3.70 (SD = 1.55) in postQ. There was no significant difference in Prospector's perceived usefulness between midQ and postQ (t = .12; df = 20; n.s.). To determine whether Google was perceived as more useful after the cold-start problem was overcome, we compared the Google score with the postQ Prospector score. This difference is significant (t = 6.29; df = 20; p<.01): Google was perceived as more useful.

Comparisons between Prospector and Google. Halfway through the study, nine participants preferred Google for searching and one person preferred Prospector. Of particular interest were the answers by six participants that stated their preference depended on the nature of the search task. They liked Google better for searching for simple facts, but thought Prospector had an added value when conducting searches related to their personal interests or study. After 12 days, 19 participants preferred Google. However, several participants apparently did so because Prospector did not offer results in German (the mother tongue of all participants). As one person stated: "I prefer Google, because it provides the possibility to search nationally. With Prospector one doesn't find regional results. The program doesn't like German words."

Incidents. From midQ we could derive two causes that led to dissatisfaction with Prospector: irrelevant search results (mentioned 9 times) and illogical re-ranking of search results (mentioned 6 times). A positive incident that was mentioned more than once regarded Prospector's particular helpfulness when submitting a query containing words with ambiguous meanings. When we asked for these incidents in the postQ the same picture emerged. However, this time more participants mentioned specific searches for which Prospector was useful, like product reviews or scientific articles.

User modeling. When we questioned the participants halfway about the visualization of the user model, 16 participants commented they understood what they were looking at, two 'thought they did', and, finally, three persons stated they did not completely understand what was displayed. Next, we asked whether the user model was a correct representation of their (search) interests. Nine participants stated it was, and six said this was mostly, or for a larger part the case. Three participants answered that they could not judge this as they had not performed enough searches or ratings for a complete user model to be generated. Finally, two participants stated that the user model was not a good reflection of their (search) interests. In the postQ, we asked the participants to judge the correctness of their user model again. Eleven participants said it was correct, three said it mostly was, and one person said it partly was. This time, four persons said they had not provided Prospector with enough feedback to generate a correct user model and two participants considered their user model incorrect. Unfortunately, the data logs cast doubts over the participants' answers. Even though all participants gave their opinion about the user model, the data logs show that only 11 participants inspected their user model before day seven and only 8 participants inspected it between day seven and twelve (with only 3 users making any changes at all). Therefore, the results regarding user modeling remain inconclusive.

Usability. The usability issues predictability, comprehensibility, unobtrusiveness and breadth of experience received mixed results: half of the participants were positive about these issues and half were not. Other issues received more uniform feedback. When asked about controllability, most participants stated they thought they were fully or for a larger part in control over the system. Privacy was not considered a barrier to using Prospector – 16 persons said the search engine does not infringe on their privacy. The last question addressed system competence. A majority believed that Prospector could deliver the results they desired. Interestingly, six participants commented that the system had the potential to deliver relevant search results, but conditionally (offering as an example the inclusion of results in German).

5 System-Centered Evaluation Results

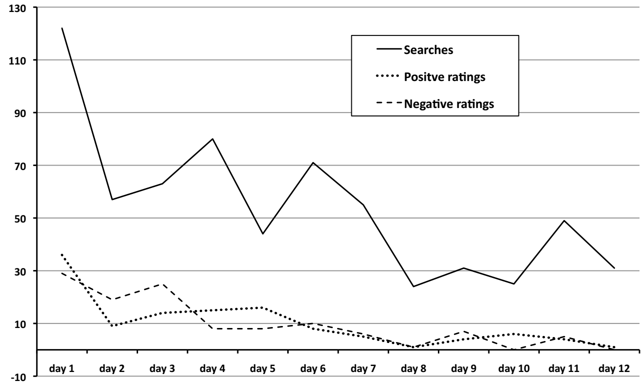


Fig 3. Number of searches, positive and negative ratings

The number of searches, positive and negative ratings over the duration of the evaluation are displayed in Fig. 3. It shows a decrease in all cases. Of note is the fact that there is no significant difference between positive and negative ratings over time.

As a means of determining whether personalization has positively affected the ranking of search results, we examined whether the results participants clicked on were ranked higher than in the original result set. Specifically, for all days, we calculated the distance between the personalized and original ranks of viewed results. This distance was positive if the result had been promoted, negative if it had been demoted, and 0 if the result had retained its rank after re-ranking.

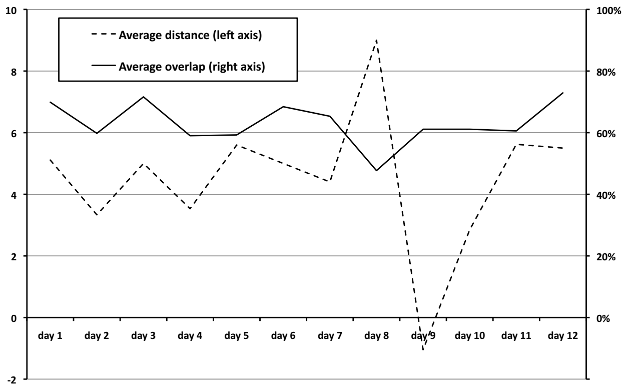


Fig 4. Distances between original and re-ranked results, and percentage of original results still ranked between 1 and 12 after re-ranking ("overlap")

Fig. 4 displays the average distance between these ranks for each day. It shows that for most days, the difference was positive for the personalized rank. The exception is day 9, during which only few searches were performed, while, at the same time, a number of results with a high negative distance were previewed or opened. This combination distorted the overall number for this particular day. For all 12 days, the viewed pages had been promoted by, on average, 4.75 ranks (*SD* = 11.52). To test the re-ranking effect, we compared the average rank distance for each viewed result to 0. This difference is significant (t = 9.14; df = 490; p<.01): Search results that participants viewed were, on average, placed higher up, due to personalization.

Because participants might tend to consult search results ranked highly, regardless of their relevance, we examined whether the first 12 results contained a disproportionately high percentage of items brought there by Prospector. We chose 12, as on a average-sized screen a user would see 6 results in one screen-full and most people do not look beyond the first 10 [15] – we rounded that number up to two full screen. Fig. 4 displays the daily average percentage of results among the first 12 that were originally there. Over 12 days, the mean percentage is 65.10%. This implies that users had a choice between (interspersed) original or re-ranked results, but chose the latter on purpose and not because they were conveniently placed at the top of the list.

In addition to these analyses, the two metrics "Rank scoring" [16] and "Average Rank" [17] were employed. Rank scoring shows how close to the optimal ranking a set of search results is, whereby 'optimal' denotes a situation in which all the consulted results appear at the top of the list. In this metric, the importance of the ranks decreases exponentially (e.g., a correct rank 1 creates a better score than a correct rank 10). We performed a paired samples t-test between the original rank score average (*M* = 5.05, *SD* = .59) and the personalized rank score average (*M* = 6.75, *SD* = 1.19). The averages were calculated from the rank score values of the 12 days. This difference is significant (t = -6.92; df = 11; p<.01): Personalized rank scores were higher than the original ones (see Fig. 5).

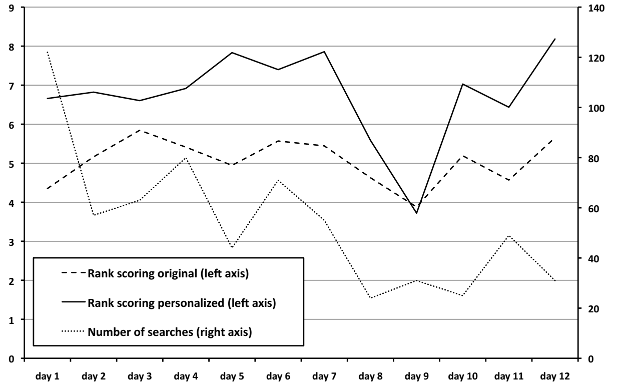


Fig 5. Rank scoring of the personalized and original ranks of viewed results

The average rank measure was calculated for the original and the personalized ranking of consulted results on a per day basis (see Fig. 6). The personalized results had a lower average rank in all cases, except on day 9. A lower average rank means that the consulted results appeared higher up in the result list. The significance of the difference between the average ranks can be derived from the significance of the average distance measure described above.

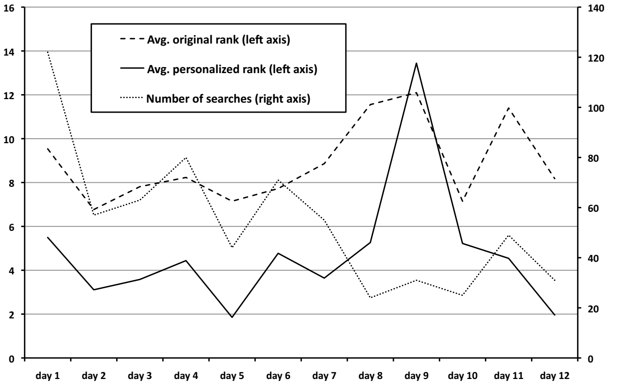


Fig 6. Average rank of the personalized and original ranks of viewed results

6 Conclusions

In this paper we have presented a longitudinal evaluation of the Prospector system, which combined a system- and a user-centered focus. The data that resulted from the system-centered evaluation has shown that Prospector effectively promotes items with relevant informational value in a list of search results. However, user perceptions on system usefulness were not in favor of Prospector: the participants thought their primary search engine (Google) was more useful. Their comments led us to think that this opinion was partly due to missing features popularized by Google (specifically, localized search, spelling suggestion). Prospector offered a different interface with different features which may have biased the participants' perception of its usefulness, regardless of the system's actual value for searching, as people highly value the appearance and features of their primary search engine [15]. A way to design for this implication is to replicate the features and appearance expected of a search engine (e.g., by incorporating the above facilities, and adopting a design similar to Google's).

The participants had very high expectations of Prospector, based on the quality of search results returned by Google, and expected Prospector to outperform Google. Some users explicitly anticipated that the result they needed be listed first or second. These expectations are apparently hard to meet, especially as users will want to see an added value fast and it may take some time for a personalized search engine to deliver top-quality results. When evaluating a personalized system, one has to make sure that participants have a sound mental model of the system so that they can form reasonable expectations. They must understand whether and how much time and effort are required for the system to 'get to know' the user. This issue is not only limited to evaluation of course: for a system to have a fair chance of user acceptance, it must ensure that its users' expectations are realistic from the very beginning.

The evaluation has suggested some circumstances in which personalized search might be more rewarding for users. These are the searches which our participants described as 'personal', or searches without a clear-cut answer. Typical for these searches are, as Marchionini terms it, relatively low answer specificity, high volume and high timeliness [19]: the answer to the search is not easily recognized as being correct (e.g., a suitable hotel in Amsterdam), it has to be found in a large body of information and, finally, the user has to invest some time in finding the right answer. Dou et al. [17] found navigational queries with low click entropy (i.e., most users chose the same result) to be less ambiguous and not suitable for personalization. Teevan et al. [18] proposed methods to detect such queries and predict the usefulness of personalization on the basis of query properties, result quality and search history.

This evaluation reinforces the notion that the application of a dual approach is instrumental in fully understanding a personalized system ([3]): If we had relied on the system-centered approach only, the results would have had a too positive skew, while the results derived from the user-centered approach alone would have put into question the system's effectiveness. In other words, by applying a double focus, one can acquire a more complete view of system usefulness (albeit, potentially with contradicting evidence). Furthermore, in certain cases such an approach makes it possible to cross-validate and ground or disprove findings (e.g., with the user model viewing behavior in this study). Furthermore, the dual focus provides us with the option to not only determine Prospector's effectiveness and chances of acceptance in a real-world setting, it also resulted in redesign input that enables us to further improve the system (e.g., by incorporating spelling suggestions).

In our longitudinal evaluation we have experienced a reduction in user activity as time progressed. This might point to need, for people running similar studies, to actively encourage participants to continue using the system under investigation (e.g., through reminders, or by providing some form of incentive). Last but not least, the application of a longitudinal evaluation setup has yielded insights which, we believe, may not have been attainable otherwise: we have been able to determine whether Prospector "works", but also what users see as the main drawbacks of the system.

Acknowledgements. The work reported in this paper has been supported in part by: (a) the EU-funded "Adaptive Learning Spaces" (ALS) project (229714-CP-1-2006-1-NL-MPP); and (b) the "Adaptive Support for Collaborative E-Learning" (ASCOLLA) project, supported by the Austrian Science Fund (FWF; project number P20260-N15).

References

1. McGrenere, J., Baecker, R.M., Booth, K.S.: A field evaluation of an adaptable two-interface design for feature-rich software. ACM transactions on computer-human interaction 14 (2007) article 3
2. Van Velsen, L., Van der Geest, T., Klaassen, R., Steehouder, M.: User-centered evaluation of adaptive and adaptable systems: a literature review. The knowledge engineering review 23 (2008) 261-281
3. Díaz, A., García, A., Gervás, P.: User-centred versus system-centred evaluation of a personalization system. Information Processing and management 4 (2008) 1293-1307
4. Brusilovsky, P., Millán, E.: User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web. Springer, Heidelberg (2007) 3-53
5. Scholtz, J., Morse, E., Potts Steves, M.: Evaluation metrics and methodologies for user-centered evaluation of intelligent systems. Interacting with computers 18 (2006) 1186-1214
6. Paramythis, A., König, F., Schwendtner, C., Van Velsen, L.: Using thematic ontologies for user- and group-based adaptive personalization in web searching. Adaptive multimedia retrieval, Berlin, Germany (2008)
7. Van Velsen, L., Paramythis, A., Van der Geest, T.: User-centered formative evaluation of a personalized internet meta-search engine. (In review)
8. König, F., Van Velsen, L., Paramythis, A.: Finding my needle in the haystack: effective personalized re-ranking of search results in Prospector. (in review)
9. Renda, M.E., Umberto, S.: Web metasearch: rank vs. score based rank aggregation methods. The ACM symposium on applied computing, Melbourne, Florida (2003)
10. Shaw, J., Fox, E.: Combination of Multiple Searches. Text REtrieval Conference, Gaithersburg, MD (1993)
11. Kay, J.: Stereotypes, student models and scrutability. In: Gauthier, G., Frasson, C., VanLehn, K. (eds.): ITS 2000. Springer-Verlag, Berlin (2000) 19-30
12. Liaw, S.S., Huang, H.M.: An investigation of user attitudes toward search engines as an information retrieval tool. Computers in human behavior 19 (2003) 751-765
13. Paramythis, A., Weibelzahl, S.: A decomposition model for the layered evaluation of interactive adaptive sysems. In: Ardissono, L., Brna, P., Mitrovic, A. (eds.): User Modeling 2005. Springer Verlag, Heidelberg (2005) 438-442
14. Jameson, A.: Adaptive interfaces and agents. In: Jacko, J.A., Sears, A. (eds.): Human-computer interaction handbook (2nd ed.). Erlbaum, Mahwah, NJ (2007) 433-458
15. Keane, M.T., O'Brien, M., Smyth, B.: Are people biased in their use of search engines? Communications of the ACM 51 (2008) 49-52
16. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. Microsoft Research, Redmond, WA (1998)
17. Dou, Z., Song, R., Wen, J.: A large-scale evaluation and analysis of personalized search strategies. WWW, Banff, Canada (2007)
18. Teevan, J., Dumais, S.T., Liebling, D.J.: To personalize or not to personalize: modeling queries with variation in user intent. SIGIR'08, Singapore (2008)
19. Marchionini, G.: Information seeking in electronic environments. Cambridge university press, New York (1995)

Evaluating Recommender Explanations:
Problems Experienced and Lessons Learned for the Evaluation of Adaptive Systems

Nava Tintarev[1], Judith Masthoff[2]

[1] Telefonica Research, nava@tid.es
[2] University of Aberdeen, j.masthoff@abdn.ac.uk

Abstract. We describe the methodological considerations that arose over a series of experiments evaluating the effectiveness of explanations for recommendations. In particular, we look at issues relating to: criteria, metrics, product domain used, choice of materials, possible confounding factors, and approximation of experience versus real experience. We generalize the problems we found and the solutions that we applied to adaptive systems. We illustrate the learned lessons with examples from our previous work on adaptive systems (ranging from adaptive learning to persuasive technologies).

1 Introduction

The evaluation of adaptive systems is not easy, and several researchers have pointed out potential pitfalls when evaluating adaptive systems. Examples of pitfalls mentioned in [1] and [2] include:

- Difficulty in attributing cause: is it the adaptation which is causing the measured effect or something else (such as system usability)?
- Insignificant results due to too much variance between participants. Adaptation is typically used when individual participants differ. However, individual differences are likely to lead to a large variance in results, and this makes it harder to get statistically significant results.
- Difficulty in defining the effectiveness of adaptation. It is sometimes hard to define what constitutes a good adaptation.
- Allocation of insufficient resources. You often need many participants to fully evaluate an adaptive system (in part due to the expected variance between participants mentioned above).
- Too much emphasis on summative rather than formative evaluation. Evaluations often measure only how good or bad a system is rather than providing information on where the problems are and how the system can be improved.

The difficulty in evaluating adaptive systems has led to a series of workshops on this topic such as [3, 4]. This paper contributes to the debate by identifying a number of problems related to and expanding those listed above. This is done in the context of a case study, where we discuss the problems experienced when evaluating explanations of recommended items, and the solutions applied. We also discuss how these problems and solutions are more widely applicable to the evaluation of adaptive systems.

2 Background to Case Study

Recommender systems such as Amazon offer users recommendations, or suggestions of items to try or buy. These recommendations can then be explained to the user, e.g. *"You might (not) like this item because..."*. In experiments, our system generates (given a simple user model) explanations for items using data retrieved from the Amazon website. In a between-subject design, we compared three degrees of personalization in a series of experiments. The example below illustrates the three conditions for *one* experiment in the movie domain:

1. **Baseline:** The explanation is neither personalized, nor describes item features: e.g. *"This movie is one of the top 250 in the Internet Movie Database"*.
2. **Non-personalized, feature based:** e.g. *"This movie belongs to the genre(s): Drama. Kasi Lemmons directed this movie."* The feature 'director' was not particularly important to this participant.
3. **Personalized, feature based:** e.g. *"Unfortunately this movie belongs to at least one genre you do not want to see: Action & Adventure. Also it belongs to the genre(s): Comedy, Crime, Mystery and Thriller. This movie stars Jo Marr, Gary Hershberger and Robert Redford."* For this user, the most important feature is leading actors, and the explanation considers that the user does not like action and adventure movies.

3 Problems, solutions and generalizations

This section discusses some of the problems we encountered in our experiments evaluating explanations, and the solutions we adopted. It also elaborates on what other researchers can learn from our experiences for the evaluation of adaptive systems in general. We illustrate how the lessons can be generalized with example evaluations where the problem appears, and where solutions were applied. Although these problems are common to many evaluations of adaptive systems, it seemed fairer to select examples from our own work, as we have *also* been affected by these problems in our previous research.

3.1 Criteria to use

Problem. The first issue we had to resolve was what we meant by a *good* explanation.

Solution. We surveyed the literature and discovered that explanations can serve multiple aims, such as increasing transparency, trust, satisfaction, efficiency, persuasiveness, and effectiveness [5]. We decided to focus primarily on effectiveness:

how helpful explanations are for users to make good decisions. According to [6], an effective explanation minimizes the difference between the user's rating of an item based on the explanation, and the user's rating of the item after experiencing it. Therefore, in our experiments, participants rated the items based on the explanation, then re-rated the items after having experienced them[3].

There is some evidence to suggest that personalization may increase persuasion, or acceptance of recommendations [7]. In contrast, we wanted to investigate whether personalization would increase *effectiveness*. In addition to measuring effectiveness, we decided to also measure user satisfaction with the explanations.

Generalization of problem. Sometimes when adaptive systems are evaluated, there is a shortage of information about what exactly they are being evaluated on [1]. For example, an adaptive instruction system can be evaluated on how well it keeps the learner motivated, how much it improves the understanding of a weak learner, how much it is appreciated etc. All of these could serve as valid aims for a "good" system, but they are likely to require different types of evaluations. Also, often the discussion on other criteria that may have been relevant is limited, and results are presented as one system outperforming another one without saying on which criterion.

Generalization of solution. In order to achieve a goal of optimization, the optimum or the main evaluation criterion needs to be explicitly formulated prior to evaluation. As optimization in one criterion may damage another, it is also important to understand how the criteria relate to one other. For example, a study of computer generated reports of babies in intensive care found that doctors preferred graphical reports (satisfaction), but made better decisions with textual reports (effectiveness) [8]. In that evaluation, focusing on just one criterion, such as satisfaction, would not have provided a full picture of the situation.

3.2 Avoiding confounding factors

Problem. The measured impact of explanations on effectiveness may be confounded with the impact of the accuracy of the recommender system. If recommended items are meant to be liked by the user (i.e. we do not give predictions for items the user may not like), and the recommender system has poor accuracy, it would be hard to distinguish between the effects of poor explanations and poor accuracy. Most likely, we would not be able to tell which factor was the main contributor to a large change in valuation of an item. This is a problem because in a real recommender system, it would be hard to guarantee comparable recommendation accuracy between participants and between conditions. Also, to obtain reasonable accuracy, participants would need to use the system for a non-trivial amount of time prior to evaluation, in order to provide their

[3] Initially we used approximations of real experience, see Section 3.4 for a justification and further discussion.

preferences.

Solution. We decided not to use a recommender system, and used random item selection instead. This meant that we did not require the experiments to include a training period. Instead, participants' preferences were explicitly requested in order to personalize the explanations. Also, our metric for effectiveness (briefly described in Section 3.1) can be used regardless of whether participants liked or disliked the items; as long as they were able to make an initial assessment based on the explanation, we were able to study their change of opinion after experiencing the item. It was not a problem for us to offer explanations for items participants might not like as participants were told that the explanations were aimed at helping them make decisions (to try or *not* to try) rather than make (only positive) recommendations. This also meant that we did not have to control for recommendation accuracy.

Generalization of problem. As mentioned in the introduction (Section 1), difficulty in attributing cause has been previously discussed as a problem in the evaluation of adaptive systems [1].

Generalization of solution. Layered evaluation has been mentioned as a way to help overcome this problem [9]. In our solution, we effectively used the dicing approach proposed in [1]: we focused on the functionality of interest, and evaluated that functionality in isolation. In a similar way, we have previously evaluated parts of an adaptive learning system [1, 10].

Another solution for an adaptation taking time would be to run an experiment in several installments - first training the system on participants, and then using the adapted system for further evaluations, such as in [11]. If this approach is used, the evaluation design needs to consider that participants may drop out (i.e. consider retention rates), and one may still need to control for confounding factors.

3.3 Domain to use

Problem. The effectiveness of explanations and degree of personalization may well depend on the domain used.

Solution. We surveyed the literature and found that there has been a great deal of debate about classification of products into different categories in economics (see [12] for examples, and [13] for an elaboration on our chosen classifications). For our research, we decided that we should at least distinguish between:

– products which are relatively easy to evaluate objectively and those which commonly require an experiential and subjective judgment
– products which are relatively cheap and those which are more expensive

Ideally, we would have evaluated the explanations in four domains, considering each of the four combinations along these two dimensions. However, this would have been very resource intensive: requiring not only substantially more participants but also detailed investigation into aspects of each domain (such as product features, appropriate material selection, selection of baseline explanations, etc). Lack of resources (e.g. time, sufficient and suitable participants) is often an issue in evaluations [2]. We decided to use a middle ground: instead of fully exploring all domain options, we chose to use two that differed with regard to both of the dimensions mentioned above: movies (cheap and subjective) and cameras (more expensive and objective). We had to perform two user-centered investigations to find appropriate features of movies and cameras (see e.g. [14]). We also had to decide what materials and baseline explanations to use in both domains. However, the additional effort involved in studying two domains was justified: we found the same results for both movie and cameras, providing us with more confidence that our results generalize across domains.

Generalization of problem. Evaluations of adaptive systems tend to focus on evaluating the system in one particular domain, often without mentioning the limitations this puts on the results. For example, when evaluating different algorithms for a group recommender system, we used the domain of video clips [15]. We drew conclusions on what algorithms people preferred (such as avoiding misery for others). However, it is more difficult to say if similar results would have been found if we had used another domain, such as news items, courses of a seven-course meal, etc. The expected duration for items and the expected impact of experiencing an unliked item is likely to differ between domains, and may well affect the final results.

Generalization of solution. The solution of surveying domains, and evaluating in multiple domains is applicable to adaptive systems in general. For example, in an adaptive e-learning system, we evaluated an adaptive item sequencing strategy for two learning tasks: a paired-associate learning task and a concept learning task [10]. This does not cover all possible learning tasks in the learning domain, but does give some insight into how generalizable the results are. So, whenever feasible, adaptive systems should be evaluated in multiple domains. If resources do not permit, at least a discussion of the possible impact of other domains should be included.

3.4 Approximation of experience versus real experience

Problem. It can be very difficult and time consuming for participants to really experience the recommended items. For example, it may take too long (or be too expensive) for participants to read a recommended book or go on a recommended holiday. Participants may also require time to *fully* experience a product (for example, a real experience of a camera may involve using it over a couple of days, so that initial technical difficulties do not overly influence the participants' final evaluation).

Solution. Previous work has approximated experience of the recommended item, for example by letting participants read online reviews [6]. In our initial experiments (for both the movie and camera domains) we used the approximation of reading on-line Amazon reviews (see e.g. [16]). Deciding on an appropriate approximation required careful consideration. For example, for movies we considered using trailers, but decided against this, as these are typically made to persuade people to see the movie rather than help them make informed decisions (which is what we wanted to achieve when we defined our aim as effectiveness). However, online reviews may also be positively biased. Therefore, after several experiments using approximation, we decided to run another experiment where users really experienced the items. We used the movie domain as movies are relatively cheap, and it is easier and faster for participants to judge movies than e.g. cameras. So, our solution has been to approximate for a number of initial studies, in order to adjust and perfect the evaluation, before a costly and time-consuming real experience evaluation.

Generalization of problem. It can be very time consuming for participants to fully experience adaptive systems. For example, a realistic experience of an adaptive learning system involves learners using it over multiple sessions, learning something they would *normally* be learning in another setting. Instead, we often evaluate over one or two sessions, sometimes using a controlled artificial learning domain. For example, when evaluating adaptive navigation in a learning system, we have used the artificial domain of square dancing, with participants learning to operate dancers on the screen using multiple computer-based lessons, but all within a one-hour session [17].

Generalization of solution. Approximation of experience is often a reasonable thing to do in early evaluations, as it requires less time, allows better experimental control, and is sometimes better for ethical reasons. This can then be followed by an evaluation in more realistic settings, often of a longitudinal nature. We have used approximation in many of our studies. For example, we measured whether an emphatic embodied agent influenced participants' mood after inducing a negative mood in an artificial test (rather than say a real course assessment) [18].

3.5 Choice of materials

Problem. For the real experience valuation we needed items that participants had not yet experienced, that had enough interesting features to produce an explanation, would not take too long to evaluate, and that were ethically ok to use.

Solution. To reduce time needed per item, we used short movies instead of full movies. To avoid movies that participants had already seen, an I-have-already-seen-this-movie button that skipped to another movie was added. To avoid exposing participants to sensitive material such as (extreme) violence or sex, only movies suitable for 15 years and over (PG-15) were used. By choosing non-offensive movies, there was a distinct risk that participants' ratings of movies would not be as well spread over possible rating values as they could be. We used a pilot study to confirm that while the distribution may not make full use of the possible values, participants still indicate values that differed sufficiently from the mid-point to warrant an interesting analysis. We also considered the presences of relevant features when selecting the movies. We included movies with actors (e.g. Rowan Atkinson) or directors (e.g. Tim Burton) that were likely to be known. As certification rating (e.g. PG - parental guidance advised) was used as a feature for explanation generation, we also selected movies that had an international certification, which is otherwise often missing for short movies.

We also found that most short movies use less famous actors/directors, and therefore, despite our best efforts, there was a higher frequency of unknown actor and director names than in our previous experiments [16]. Most likely as a consequence of this, we found that participants were less happy with personalized explanations (actors and directors are the most common preferred features aside from genre). Additionally, the ethical considerations had the side effect that the majority of movies watched belonged to the genres comedy, animation and children. So, there is not always a perfect solution, but at least being aware of the potential impact of materials can help explain results and understand the limitations of a study.

Generalization of problem. Fields such as psychology have a common practice of carefully selecting the materials they use in experiments. This is done either because the material may affect the outcome (we also discuss avoiding confounding factors in Section 3.2) or for other reasons such as ethical ones. In the evaluation of adaptive systems, we are often told what materials were used, but not on the basis of which criteria they have been chosen, and whether pilot studies have been done to validate their appropriateness.

Generalization of solution. The criteria on the basis of which materials are chosen need to be clearly defined and stated. In addition, pilot studies need to be performed to test the suitability of materials. For example, when studying the effect of adding a doctor's photo on website credibility, we needed the photo to contain an image of a doctor that would be considered credible in this domain. We found such a photo by running pilot studies in which participants judged the profession of the person depicted, and rated their domain credibility using validated metrics [19].

3.6 Appropriate measurement

Problem. To measure true effectiveness, we needed to distinguish between participants not having formed an opinion of the item, and participants believing the item is kind of average (middle of the scale).

Solution. We decided to add a separate opting-out option as an alternative for rating, for participants who were not able to formed an opinion. In our analyses of effectiveness, we excluded ratings of participants who had opted-out. One thing we noticed in all our experiments was that while baseline explanations did surprisingly well for effectiveness, they also led to a very high opt-out frequency: participants were unable to provide a rating. This means that only considering the change between the before and after rating for those people who opt-in is not a true reflection of effectiveness.

Despite the opportunity to opt-out, we still found that some participants seemed to use the middle of the scale to indicate that they had no real opinion about a movie. This could be seen in the higher frequency of middle of the scale ratings for baseline explanations, or when participants were asked to give a rating based on only a movie title (without the explanation).

Problem. The effectiveness metric from [6] considers the difference between the before and after ratings. However, they do not discuss the effects of over- and underestimation (which could lead respectively to trying an item you may not end up liking and missing an item you may have liked). So, the question arises whether an explanation leading to an overestimation is as bad as one leading to a similarly big underestimation. And how about the position of the gap? Does it matter whether (on a scale from 1 to 5) the pre-rating is 3 and the post-rating 5, compared to a pre-rating of 1 and a post-rating of 3? To complicate matters further, does it all differ per domain type?

Solution. We investigated these questions [13], and found that:

- Overestimation was considered more severely than underestimation
- Overestimation was considered more severely in high investment domains compared to low investment domains (see also Section 3.3 on domain effects).
- Gaps which remained in the negative half of the scale were considered more severely than gaps which crossed over from good to bad (or vice-versa), and gaps which remained in the positive half of the scale.

Generalization of problems. In the evaluation of all systems, including adaptive ones, one has to take care that the metric used is really measuring what you want to find out. For example, when assessing personalisation in interactive TV often the time spent watching a programme is used as an indication of user interest. However, longer viewing times may well have been caused by other factors such as the viewer having a coffee or even being so bored that they fell asleep. Learners spending more time on a lesson may mean that they are more motivated or that they find the lesson harder to understand.

Generalization of solutions. A critical analysis is required of all metrics used, asking whether there are situations when the value given by the metric is not accurate. For a metric to be good, the same value should have the same meaning independent of the circumstances. For example, our experience shows that the effectiveness metric of [6] falls short of this, as e.g. over- and underestimation may lead to the same value while having a different effect on users.

4 Conclusions

This paper illustrates some of the problems we have encountered when investigating the effectiveness of explanations in a recommender system. Our investigations consisted of a series of experiments, and in each experiment we improved our understanding of the evaluation design. As we have discussed, many of these issues are generalizable to other types of adaptive systems. In any evaluation, it is important to:

- Decide which criteria to use
- Avoid confounding factors
- Take into account domain effects
- Build up the experiment gradually, and consider limited resources
- Take into account the effects of the material you select
- Consider if a metric really measures what you want

From this paper it would perhaps be easy to conclude that it is hard (and probably impossible) to design the perfect evaluation. In retrospect, there is always something else that could impact the results. Clearly defining your goals, metrics and refining your design through a sequence of experiments, and gradually investigating different aspects, will however help you avoid the most common pitfalls - may your next evaluation be a successful one!

References

1. Masthoff, J.: The evaluation of adaptive systems. In: Adaptive evolutionary information systems. Idea Group publishing (2002) 329–347
2. Weibelzahl, S.: Problems and pitfalls in evaluating adaptive systems. In: Fourth Workshop on the Evaluation of Adaptive Systems in conjunction with UM'05. (2005) 57–66
3. Weibelzahl, S., A., P., Masthoff, J., eds.: Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, associated with AH'06, Dublin (2006)
4. Weibelzahl, S., Paramythis, A., Masthoff, J., eds.: Fourth Workshop on the Evaluation of Adaptive Systems, associated with UM'05, Edinburgh (2005)
5. Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: WPRSIUI associated with ICDE'07. (2007) 801–810
6. Bilgic, M., Mooney, R.J.: Explaining recommendations: Satisfaction vs. promotion. In: Proceedings of the Wokshop Beyond Personalization, in conjunction with IUI. (2005) 13–18
7. Carenini, G., Moore, D.J.: An empirical study of the influence of user tailoring on evaluative argument effectiveness. In: IJCAI. (2001)
8. Law, A.S., Freer, Y., Hunter, J., Logie, R., McIntosh, N., Quinn, J.: A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. J Clin Monit Comput. **19(3)** (2005) 183–94

9. Paramythis, A., Totter, A., Stephanidis, C.: A modular approach to the evaluation of adaptive user interfaces. In Weibelzahl, S., Chin, D.N., Weber, G., eds.: Evaluation of Adaptive Systems in conjunction with UM'01. (2001) 9–24
10. Masthoff, J.: An Agent-Based Interactive Instruction System. PhD thesis, Eindhoven University of Technology (1997)
11. Amatriain, X., Pujol, J.M., Oliver, N.: I like it... i like it not: Evaluating user ratings noise in recommender systems. In: UMAP. (2009)
12. Cho, Y., Im, I., Hiltz, J.F.S.R.: The impact of product category on customer dissatisfaction in cyberspace. Business Process Managment Journal **9 (5)** (2003) 635–651
13. Tintarev, N., Masthoff, J.: Over- and underestimation in different product domains. In: Workshop on Recommender Systems associated with ECAI. (2008) 14–19
14. Tintarev, N., Masthoff, J.: Effective explanations of recommendations: User-centered design. In: Recommender Systems. (2007) 153–156
15. Masthoff, J.: Group modeling: Selecting a sequence of television items to suit a group of viewers. User Modeling and User Adapted Interaction **14** (2004) 37–85
16. Tintarev, N., Masthoff, J.: Personalizing movie explanations using commercial meta-data. In: Adaptive Hypermedia. (2008) 204–213
17. Masthoff, J.: Design and evaluation of a navigation agent with a mixed locus of control. In Cerri, S., Gouardres, G., Paraguau, F., eds.: Intelligent Tutoring Systems, Berlin, Springer Verlag (2002)
18. Nguyen, H., Masthoff, J.: Designing empathic computer: The effect of multimodal empathic feedback using animated agent. In: Persuasive Technology, Claremont, USA, Springer Verlag (2009)
19. Nguyen, H., Masthoff, J.: Is it me or is it what i say? source image and persuasion. In: Persuasive Conference, Springer Verlag (2007)

User-centered design methods for validating a recommendations model to enrich learning management systems with adaptive navigation support

Olga C. Santos[1], Ludivine Martin[1], Elena del Campo[2], Mar Saneiro[2], Emanuela Mazzone[1], Jesus G. Boticario[1], Helen Petrie[3]

[1]aDeNu Research Group, Artificial Intelligence Dept., Computer Science School, UNED,
C/Juan del Rosal, 16. 28040 Madrid, Spain
{ocsantos, ludivine.martin, emazzone, jgb}@dia.uned.es
http://adenu.ia.uned.es/
[2] Evolution and Education Psychology Department. Faculty of Psychology, UNED,
C/Juan del Rosal, 10. 28040 Madrid, Spain
{mcampo, masterdiscap}@psi.uned.es
[3] Department of Computer Science, University of York, Heslington, York, YO10 5DD, UK
helen.petrie@cs.york.ac.uk

Abstract. Recommendation techniques have shown to be successful in many domains (e.g. movies, books, music, etc.). This success has motivated us to research on how to deploy a recommending system in the eLearning domain to extend the functionality of standard-based learning management systems with adaptive navigation support. An initial model of the recommendation process has been developed from informal discussions with lecturers. This is now being elaborated and validated using a scenario-based user-centered design process. This paper presents the formal methodology to carry out this validation process.

Keywords: User-centred design methods, Elicitation process, Recommender systems, Learning Management Systems, Adaptation.

1 Introduction

Recommender systems (RS) support users in finding their way through the possibilities offered in web-based environments by highlighting information a user might be interested in from the information already available in the system. The first challenge for designing a RS is to define the users and their purposes [1]. RS in education should help and support both learners and teachers [2]. In particular, the RS's goal is to improve learning effectiveness and efficiency, as well as learners' satisfaction, while reducing the teachers' workload related to the follow-up and support of the learners. The approach followed by this research focuses on suggesting learners the most appropriate actions to take by the user in the learning management system (LMS) at each moment (i.e. navigation adaptation), which can vary from reading some specific contents that have been uploaded in the file storage after the course has been packaged to posting a comment in a blog to foster the learner to reflect what has been learnt [3]. The RS takes as input i) the user profile (which can be dynamically built from the users' interactions) and ii) the current context (e.g. course, objective, platform tool, …). With this information, the appropriate actions (e.g. links to objects in the LMS with instructions on what to do and explanations to justify it purpose) are recommended to the current user. In a first phase, the recommendations are obtained from teachers by following the methodology described below. These recommendations serve a double purpose: 1) avoid the cold start problem of RS, and 2) feed machine learning algorithms to tune the recommendations and/or produce new ones from the experience inferred by analysing the interactions of the users in the system. The later is planned for a second phase in the research.

If the RS supplies appropriate recommendations to the learners in the context where they are relevant to help them while interacting with the LMS, the teacher will be relieved from providing this specific type of support and can focus on other educational activities: preparing contents suitable for the learners' needs (e.g. learning styles) or giving more detailed advice to specific situations that are not yet covered by the RS.

In the context of this research, the first efforts undertaken to build a knowledge-based recommender for the eLearning context are described elsewhere [3, 4]. As a result, a recommendations model [3] and a standard-based recommendations service that implements that model [5] were produced. The prototype has been integrated into the dotLRN open source standard-based LMS. A formative evaluation process was carried out, which include some small-scale studies with users that are reported in [4]. Very shortly, users had to interact with a course where they were recommended different actions depending on their learning styles and their situation in the course. Afterwards, they were asked to fill in a questionnaire about their perception of the RS output and their interest for the different types of recommendations.

2 Scenario-based user-centred design approach

When trying to define recommendations using the model, we have found that we lack the content and context to think about meaningful recommendations that address the real needs of learners in eLearning scenarios. Moreover, we realized that although we had tried to involve users in our work, we had not done it properly. At this stage, having already a recommendations model and a RS providing adaptive navigation support in an LMS, we realized that we needed to go back to the users and apply appropriate user-centred design methods to get these meaningful recommendations. The objective was twofold: 1) involve users to validate (and refine if needed) the model previously obtained, and 2) obtain samples of meaningful psycho-educational sound recommendations from current teaching practices. With the collaboration of experts in Human Computer Interaction and Psychology, a formal methodology which applies scenario-based methods, was defined to help us lead this process.

Scenario-based methods [6] are used to elaborate the design. They consist on involving the user in writing stories (i.e. scenarios) about the problems taking place in relevant situations that come to their mind. On top of these scenarios, the design team proposed solutions to these situations.

As commented in the introduction, recommendations address the needs of the learners when interacting with LMS and try to suggest the most appropriate actions to take depending on the current user in the current context. However, unlike RS for the entertainment domain, where the goal is to satisfy the users' preferences, in the educational domain psycho-educational considerations have to be taken into account. What a learner prefers may not be the most adequate for their learning. For this reason, the users involved in our study to elicit these scenarios are teachers and not learners. However, the outcomes of this study (i.e. the recommendations elicited) have to be checked with learners to assure that they are useful to them to reach the desired goal: learning effectiveness and efficiency, as well as learners' satisfaction. The plan established for this study covers the following four stages:

Stage 1: Briefing and initial data gathering on the participants' background

- In an introductory face-to-face session:
  - the aims and objectives of the research are explained to potential participants as well as the nature of the participation expected of them and the benefits for them
  - sample scenarios are presented to the participants, who are asked to think about them in order to obtain other scenarios that have occurred in their work
  - a consent form describing the conditions and requests is provided to the participant, which they can take home to read carefully before signing it
- After the face-to-face session, participants are given time to digest the information.
- If they agree to participate, they have to i) fill in an online questionnaire with demographic information, including information about their teaching experience and ii) sign the consent form and give it to the research team in the next face to face session (next stage).

Stage 2: Eliciting scenarios with the participants

- Individual face-to-face semi-structured interviews are arranged with the research team to build together a couple of scenarios that reflect the teacher's experiences.
- The interview is conducted by a primary researcher who poses the questions and a secondary researcher taking notes. In this way, one of the researcher focuses on following the reasoning of the participant, and the other checks that the relevant information to identify the recommendations is being provided.

Stage 3: Identifying the recommendations in the scenarios by the research team

- From the scenarios built in the interviews, the research team identifies recommendations and attempt to map them onto the recommendations model previously defined. If the information required to describe the recommendation does not map to the model, the model will be revised to include the new information. This process is done by three members of the research team, and then checked for consistency.
- The result of this process is an enriched scenario that includes recommendations that address the problems and situations identified by the teachers in their scenarios.

Stage 4: Review of the scenarios and the recommendations elicited

- In the first step in this stage, participants analyse individually the enriched scenarios proposed by the researcher which include the recommendations identified by the research team. They are asked 1) to state the relevance of each recommendation using a five point Likert scale and 2) to propose new recommendations (or modifications to the existing ones) within the situation described in the scenario.

- After the revision by the participants, the research team aggregates scenarios that share similar situations and present a new (and reduced in number) set of enriched scenarios.
- To validate the results obtained, a focus group is planned. It will include some of the teachers who have built the scenarios, but also we are considering involving other roles, such as experts in the online teaching practices and learners. The goal is to discuss the set of enriched scenarios.

At the end of the four stages of the process, participants are provided with detailed information about the model and how the current prototype is running in dotLRN. The teachers are asked whether they would be interested in continuing the collaboration and applying the recommender system to one of their courses. If they agree, a new face to face session will be arranged to prepare the recommendations for the course. Moreover, if the model has been modified after the study, the previous prototype of the RS has to be modified accordingly.

Although it is still too early to draw conclusions on the application of the methodology, we think that our approach can be informative to other designers, and motivate them to work on a formal methodology that apply user-centered design methods from the very beginning of their research.

Acknowledgments. Authors would like to acknowledge the insightful suggestions to the methodology provided by Chris Power from the University of York. Authors would also like to thank the European Commission and the Spanish Government for funding the research involved in this work, EU4ALL (IST-2006-034778) and A2UN@ (TIN2008-06862-C04-01/TSI).

References

1. McNee, S., Riedl, J., Konstan, J. Making Recommendations Better: An Analytical Model for Human-Recommender Interaction. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2006)
2. Zaine, O. R. Building a Recommender Agent for e-Learning Systems. In Proceedings of International Conference on Computers in Education, Auckland, New Zealand, 3-6 December, pp. 55-59 (2002)
3. Santos, O.C. and Boticario, J.G. Users' experience with a recommender system in an open source standard-based learning management system. In proceedings of the 4th Symposium of the WG HCI&UE of the Austrian Computer Society on Usability & HCI for Education and Work (USAB 2008), p. 185-204 (2008)
4. Santos, O.C., Boticario, J.G. Adaptive accessible design as input for runtime personalization in standard-based eLearning scenarios. In proceedings of the 2nd Conference on Accessible Design in the Digital World 2008 (2008)
5. Santos, O.C., Granado, J., Raffenne, E., Boticario, J.G. Offering Recommendations in OpenACS/dotLRN. Int. Conf. on Community based environments, Valencia (2008)
6. Rosson, M. B. and Carroll, J. M. Usability engineering: scenario-based development of human computer interaction. Morgan Kaufmann (2001)