Utility-Based Evaluation of Adaptive Systems

Eelco Herder

Department of Computer Science, University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
herder@cs.utwente.nl

Abstract. The variety of user-adaptive hypermedia systems available calls for methods of comparison. Layered evaluation techniques appear to be useful for this purpose. In this paper we present a utility-based evaluation approach that is based on these techniques. Issues that arise when putting utility-based evaluation into practice are dealt with. We also explain the need for interpretative user models and common sets of evaluation criteria for different domains.

1 Introduction

As the Internet has become a common source of information and services, the need for web sites to cater a heterogeneous user population has increased dramatically. It has been shown hard to design interfaces that match all user needs in all user contexts, which might be partially due to the lack of well-founded design guidelines [11]. Adaptive hypermedia systems try to bridge the gap between sites and individual users by building models of the goals, preferences and knowledge of each individual user, and use this model throughout the interaction in order to adapt to the user needs [1].

Some decades of research have provided us with a huge collection of different user-adaptive systems. Unfortunately, thus far most adaptive systems are only compared to their non-adaptive counterparts [7]. This makes it hard to compare the results as reported in journals or conference proceedings, as the systems are evaluated against different criteria, by lack of well-defined or common criteria for the success of adaptive hypermedia systems [12].

Recently, the use of frameworks for layered evaluation of adaptive applications and services was advocated by a number of researchers [2][7][12]. Although these frameworks are described at different levels of granularity [12], in essence they separate the process in the evaluation of the *interaction assessment* phase and the evaluation of the *adaptation decision making* phase [7]. The basic intuition behind this approach is that unsuccessful adaptations might be due to incorrect assessment results, or to improper adaptations based on a correct assessment. Layered evaluation of adaptive systems appears to be a promising approach, as shown in a case study described in [2]. In the next paragraph the limitations of user models that are constructed in the interaction phase are described, and how they should be dealt with when evaluating adaptation decisions. In the third paragraph we propose a utility-based approach to layered evaluation. Further, it is argued why common sets of

evaluation criteria are needed in order to objectively compare different adaptive systems. We conclude with a brief summary and some future perspectives.

2 Interaction Assessment and User Models

In this paragraph we describe issues that arise in the *interaction assessment* phase. User models can only contain limited data on the environment. These data are likely to be unreliable, due to limited input data and to imperfect inference mechanisms [12]. Further, the accessibility and interpretability of user models is highly dependent on the representations used [9]. These issues need to be considered when developing an evaluation method for adaptive hypermedia.

2.1 Limited Knowledge and Uncertainty in User Models

As mentioned before, adaptive systems build models of users and their user contexts, which are used for making adaptation decisions. For at least two reasons, these user models only cover data that are expected to be relevant to specific adaptation goals. First, limited computational resources hinder us to cover all factors that influence the interaction between user and system. Second, there exists no integrative model of web navigation. From a hypothetical model of web navigation, as presented in [6], it becomes clear why user models cannot cover all relevant factors. The hypothetical model consists of six categories of predictive factors for user performance and satisfaction:

- cognitive factors (e.g. expertise, working memory, spatial ability)
- affective factors (e.g. mood, trust)
- conative factors (e.g. motivation, interests)
- demographic factors (e.g. age, gender)
- technology factors (e.g. means of interaction, navigation support)
- task/context factors (e.g. time criticality, interruptions)

Given this general framework, one can easily think of zillions of factors that fit into one of these categories and that might have impact on web navigation. Moreover, these factors are expected to be highly correlated: given the fact that a user is a typical male Dutch student, there is a fair chance that he might suffer from a post-weekend hangover on a Monday morning, which does not contribute to his working memory or his motivation. To avoid undesirable biases or omissions, an important aspect of user model evaluation is verifying whether the model reflects the user's actual state [8].

The choice of factors to include in some user model also depends on the *availability* of data. Adaptive systems are limited to using low-level monitoring information as input, which might be unreliable. Further, inference mechanisms - be it hand-crafted knowledge bases or advanced machine-learning techniques - introduce uncertainty to predictions as well [13]. Therefore, evaluation methods should provide measures for the expected accuracy of individual user models [8]. As one cannot infer users' thoughts, feelings or expectations directly from their behavior, empirical evaluation methods that elicit user feedback are needed [8]. Unfortunately, these

methods are not a formal proof of facts [12] and therefore never will be able to completely remove uncertainty in user models. Both levels of uncertainty need to be dealt with in the evaluation process.

2.2 The Influence of the Representation of User Models

User models can be represented in different ways. Traditional user modeling systems make use of handcrafted knowledge representation techniques, with clear semantics that enable interpretation of the user model. However, machine-learning techniques have become very popular in the adaptive hypermedia community [9][13]. The representation of learning results is highly dependent on the technique being used (e.g. decision trees, probability tables). This implicit representation makes it hard to interpret the data inferred [9]. Naturally, one can judge any user model on how well it differentiates between users. More discriminative user models are likely to be better than those that are unable to model the differences between users. But one cannot tell from an implicit representation whether it correctly models a user context.

Explicit representation of user models enables evaluation of the adaptation decision phase, independent of the interaction assessment. As user models that make use of implicit representations cannot be evaluated on correctness – or only partially – layered evaluation becomes cumbersome. This does not mean that implicit user models should be avoided. However, one should be aware of the implications. This issue is dealt with in the next paragraph.

3 A Utility-Based Evaluation Approach

In this paragraph we propose a utility-based approach to layered evaluation of adaptive systems. We explain how this approach benefits from explicit representations of user models. Further, we explain the need for a common set of evaluation criteria in different domains.

3.1 A Utility-Based Approach to Layered Evaluation

The current evaluation practice attempts to evaluate adaptation as a whole [2], with user satisfaction or performance as the overall criterion – based on selected, measurable criteria [7]. In other words, the evaluation can be seen as an utility function U [10] that maps a system, *given some user context*, to a quantitative representation of user satisfaction or performance. If one compares an adaptive system with its non-adaptive counterpart, the value of adaptation is the difference in utility between the two systems.

The main advantage of layered evaluation methods is that it breaks the utility function in several functions. In the introduction the basic intuition behind this approach is explained. This can be observed from the utility function as well. Suppose

there is a utility function U_1 that maps the interaction assessment and the resulting user model to a real number that represents its correctness. Suppose there is also a utility function U_2 that maps a system, *given some user model*, to a real number that represents user satisfaction or performance. We can then express the whole utility function as $U = U_1U_2$. It is clear that the latter utility function better indicates the usability of an adaptive system. As one would expect, an adaptive system that coincidentally makes correct decisions based on wrong assumptions will be rated poorly by this function [10].

3.2 Interpretative User Models and Utility-Based Evaluation

As explained in the previous paragraph, a user model needs to be interpretable – at least to a certain extent – in order to judge its correctness. This evaluation process can be divided into two phases [8].

The first phase involves the modeling process itself. Accuracy measures of interaction monitoring and inference methods are needed for this purpose. The representation format – both its interpretability and the balance between factors that are included – is of influence as well, as both aspects add uncertainty to the evaluation of the actual models. The second phase involves the evaluation of the actual models, given the limitations of the methods used. As mentioned before, empirical evaluation methods – such as controlled experiments – appear to be the most fruitful approach. Combining both phases yields U_1, which actually indicates the joint uncertainties as introduced by the approach chosen, by the level of interpretability of the approach and by the actual interaction assessments using this approach.

Empirical evaluation of the actual user models can be done through user tests in which the assessment conclusions are compared to the opinion of the user or an expert [7]. Evaluation of the modeling process itself has not been addressed thus far – at least to the best of our knowledge – and therefore remains an open issue.

3.3 The Need to Decide On Evaluation Criteria for Adaptation Decisions

In the first section of this paragraph it was discussed how a utility function can be used for the evaluation of adaptive systems. In this section we deal with the issue what such a utility function should look like. Unfortunately, this function is highly dependent on the criteria employed. For different domains, different criteria can be thought of. The benefits of adaptation of educational hypermedia can be expressed by an increase in learning rate or examination results. The benefits of adaptation of e-commerce systems can be expressed by an increase of sales or customers returning to the vendor more often. However, more generic adaptation goals are hard to evaluate.

A common problem in hypermedia is users getting ‘lost in hyperspace’. At some point users may not know where they are, how they got there and where they should go next. As a result, navigation performance and user satisfaction drop dramatically [4]. But how can one observe lostness from user actions? Users who are exploring a document may be rated as disoriented, even though they may be experiencing no

disorientation at all [5]. One of the main benefits of hypermedia is that it facilitates both goal-directed activities and open-ended browsing [11]. A user who is browsing a site can be regarded as a tourist wandering through a city center, looking for unknown places of beauty. This type of navigation implies some sort of *voluntary lostness*. However, gradually a tourist might feel more uncomfortable not knowing where she is. Given these varying navigation strategies and user goals, standard usability measures, such as performance, are not suitable for evaluation purposes.

Recently, a new criterion called behavior complexity has been proposed [12]. User satisfaction is reported to improve as interaction complexity decreases. However, browsing is expected to produce more complex navigation behavior than goal-directed interaction [5]. Moreover, browsing can be encouraged or discouraged by the structure of a site. Based on this observation, we proposed metrics for the evaluation of user navigation that take both site structure and navigation complexity into account [4][5]. In experimental settings, techniques such as observation, questionnaires and thinking-aloud protocols [6] can be employed as well.

From the above it can be concluded that many different evaluation criteria can be thought of. In order to compare systems or approaches, one needs to decide on sets of criteria to be used in several domains. These domains can range from broad (e.g. hypermedia in general) to more narrow (e.g. educational hypermedia). In this process, previous work on general usability matters, as carried out by e.g. the W3C, as well as overviews of the current state of the art in adaptive hypermedia – e.g. [1] – should be taken into account. Once such common criteria are established, researchers will be able to employ them to guide them in their research and to compare results.

4 Summary and Future Perspectives

In this paper a utility-based approach to evaluation of adaptive systems is proposed, mainly inspired by previous work on layered evaluation [7][8][12] and theories on uncertainty and utility from the field of artificial intelligence [10]. We pointed out how interpretable user models will facilitate evaluation; for this reason, when choosing a modeling technique that produces implicit representations, researchers should weigh its advantages against the loss of interpretability. We also indicated why researchers should decide on common sets of evaluation criteria and methods that are used by researchers in some domain (e.g. hypermedia in general, web sites, educational hypermedia, e-commerce systems).

Many open issues have become more apparent from the utility-based perspective on layered evaluation of adaptive systems. The utility functions – although they are not clearly defined at the moment – connect the separate parts of layered evaluation frameworks as described in [7][8].

We expect that the utility-based approach can also be used *within* adaptive systems. *Decision networks* [10] can be constructed to choose the most promising adaptation decision from several alternatives, based on an individual user model and the system's judgement on its correctness. This form of meta-reasoning paves way to more versatile and robust adaptive systems.

5 References

1. Brusilovsky, P.: Adaptive Hypermedia. *User Modeling and User-Adapted Interaction 11* (2001) pp. 87-110
2. Brusilovsky, P., Karagianidis, C. & Sampson, D.: The Benefits of Layered Evaluation of Adaptive Applications and Services. *Empirical Evaluation of Adaptive Systems. Proc. of workshop at the 8th Intl. Conference on User Modeling UM2001*. Freiburg (2001) pp. 1-8
3. Catledge, L.D. & Pitkow, J.E.: Characterizing Browsing Strategies in the World-Wide Web. *Computer Networks and ISDN Systems 27 (6)* (1995) pp. 1065 -1073
4. Herder, E. & Van Dijk, B.: From Browsing Behavior to Usability Matters. *Workshop on Human Information Processing and Web Navigation, HCI 2003*, Crete (to appear)
5. Herder, E.: Modeling User Navigation. *UM2003 User Modeling: Proceedings of the Ninth International Conference* (to appear)
6. Juvina, I. & Van Oostendorp, H. Human Factors in Web-assisted Personal Finance. *Workshop on Human Information Processing and Web Navigation, HCI 2003*, Crete (to appear)
7. Karagiannidis, C. & Sampson, D.: Layered Evaluation of Adaptive Applications and Services. *Adaptive Hypermedia and Adaptive Web-Based Systems* (2000) pp. 343-346
8. Paramythis, A., Totter, A. & Stephanidis, C.: A modular approach to the evaluation of Adaptive User Interfaces. *Empirical Evaluation of Adaptive Systems. Proc. of workshop at the 8th Intl. Conference on User Modeling UM2001*. Freiburg (2001) pp. 9-24
9. Pohl, W. & Nick, A.: Machine Learning and Knowledge Representation in the LaboUr Approach to User Modeling. *Proceedings of the 7th International Conference on User Modeling*. Banff, Canada (1999) pp. 197-188
10. Russell, S. & Norvig, P.: Artificial Intelligence: a Modern Approach. Prentice-Hall, Inc. (1995)
11. Shneiderman, B.: Designing Information-Abundant Websites: Issues and Recommendations. *International Journal of Human-Computer Studies 47 (1)* (1997) pp. 5-29
12. Weibelzahl, S., & Weber, G.: Advantages, Opportunities, and Limits of Empirical Evaluations: Evaluating Adaptive Systems. *Künstliche Intelligenz 3 (02)* pp. 17-20
13. Zukerman, I. & Albrecht, D.W.: Predictive Statistical Models for User Modeling. *User Modeling and User-Adapted Interaction 11* (2001) pp. 5-18