Weibelzahl Stephan, Paramythis Alexandros,
and Masthoff Judith (Eds.)

Fourth Workshop on the Evaluation of Adaptive Systems

Held in conjunction with:

10th International Conference on User Modeling (UM'05)
Edinburgh, UK , July 24th to 30th, 2005

Preface

The Fourth Workshop on the Evaluation of Adaptive Systems (EAS) follows in the tracks of three very successful workshops held in conjunction with UM2001, UM2003 and AH2004. The workshop's guiding perspective is that adequate methods and reliable criteria are prerequisites towards increasing the quantity and quality of evaluation studies on adaptive systems. The workshop aimed to contribute to the exploration and discussion of suitable methods and criteria in various domains with differing user modeling and adaptation techniques. It further aimed to encourage researchers to perform evaluation studies with their own systems.

The workshop, continuing in the steps of its predecessors, encouraged submissions on the following general themes:

- Criteria and methods for evaluating adaptation
  - Which of the existing criteria and methods are appropriate for the evaluation of user models and adaptive systems?
  - What new criteria need to be introduced to specifically cater for the presence of user modeling and adaptation in the evaluated systems?
  - What empirical methods are appropriate (or how do existing methods need to be modified so as to be suitable) for assessment against the new sets of criteria?
- Adaptation metrics
  - What aspects of user modeling / adaptation offer themselves to assessment through qualitative or quantitative measures?
  - Can metrics be developed to facilitate the comparison between different (versions of) adaptive systems, or between adaptive and non-adaptive systems?
  - Can metrics be developed to assess adaptation in “absolute” terms (e.g. on the basis of well-defined scales), thus making it possible to conduct studies that do not involve comparisons between systems?
- Encouraging and supporting evaluation studies
  - How can we foster an increase in the volume and quality of empirical evaluations of adaptive systems?
  - What are the most common pitfalls that can be identified in previous studies?

Expanding upon the scope of its predecessors, the workshop also had an explicit focus on non-empirical approaches to evaluating user models and systems that employ them as a means towards exhibiting intelligent / adaptive behavior. Of particular interest were metrics / approaches that would enable one to judge factors such as correctness, effectiveness, efficiency, etc., during early stages of the development cycle of user modeling systems, where, typically, prototypes are still not functional or mature enough to involve users in the evaluation process. Also of interest were approaches that can be employed to lower the inarguably high costs and complexity involved in performing full-scale evaluations of adaptive / intelligent systems.

The workshop also included a talk delivered by the workshop organizers, on the topic of common pitfalls and problems when using “traditional” approaches to evaluate adaptive systems. This talk was intended as a condensed overview of the results of previous workshops, and as a basis for discussion among participants.

Further to the above, this workshop introduced a new “evaluation challenge” track. The challenge concerns an adaptive system, which recommends sequences of music clips to groups of users. Details on the submission requirements and received entries are included in these proceedings.

Program Committee

We would like to take this opportunity to thank the members of the program committee for the time and effort they have put into reviewing the papers for this workshop. Their support in making this workshop a worthwhile event is greatly appreciated:

Peter Brusilovsky, University of Pittsburgh, USA
David Chin, University of Hawaii at Manoa, USA
Betsy van Dijk, Twente University, The Netherlands
Judith Masthoff, University of Aberdeen, UK
Alexandros Paramythis, Johannes Kepler University Linz, Austria
Gerhard Weber, University of Education Freiburg, Germany
Stephan Weibelzahl, National College of Ireland, Ireland
Frank Wittig, SAP, Germany

Organizers

Dr.Stephan Weibelzahl
National College of Ireland Dublin
Mayor Street IFSC, Dublin 1, Ireland
sweibelzahl@ncirl.ie
http://www.weibelzahl.de

Alexandros Paramythis
Institute for Information Processing and Microprocessor Technology (FIM)
Johannes Kepler University Linz
Altenbergerstr. 69, A-4040 Linz, Austria
alpar@fim.uni-linz.ac.at
http://www.fim.uni-linz.ac.at/staff/paramythis/

Dr.Judith Masthoff
Department of Computing Science
University of Aberdeen
Aberdeen AB24 3UE, Scotland, UK
jmasthoff@csd.abdn.ac.uk
http://www.csd.abdn.ac.uk/~jmasthof/

Table of Contents

Potentials of Eye-Movement Tracking in Adaptive Systems

Roman Bednarik

Department of Computer Science, University of Joensuu,
P.O. Box 111, FI-80101, Joensuu, Finland
bednarik@cs.joensuu.fi

Abstract. Eye-movement tracking proved its potentials in many areas of human-computer interaction. Resting on a hypothesis that eye-direction and mind are linked, some of the HCI researchers have employed eye-movement trackers to investigate the visual attention focus of the participants completing their tasks. Others have used the eye-movement tracking in real-time applications, either as a direct interaction device or as an input to gaze-aware interfaces. Inspired by the previous HCI applications, we propose to utilize eye-movement trackers in adaptive systems research and development in two ways. First, the evaluations of adaptive systems could get an access to the information otherwise unavailable, as for instance to how the visual attention and cognitive processing are influenced by an adaptivity implemented into the evaluated system. Second, we propose to employ the eye-movement tracking technologies for a real-time registration of users' loci of visual attention, therefore increasing the awareness of the adaptive systems about their current users. We discuss possible potentials, difficulties and pitfalls of eye-movement tracking when applied to adaptive systems. We argue that a methodological framework of applying eye-tracking into adaptive systems shall be developed.

1 Introduction

In various areas, eye-movement tracking systems have provided researchers an access to underlying cognitive processing of users completing their tasks. Typical areas where eye-trackers have been successfully employed are studies of reading [13], scene perception, visual search, and eye-based interaction as for instance eye-typing [11]. Others have used eye-tracking in usability evaluations of computer interfaces [5]. For a recent survey of applications of eye-movement tracking, see [4].

Surprisingly little is known about how eye-movement patterns, and therefore also cognitive processes, are influenced when the environments exhibit some kind of adaptive behavior. Although eye-tracking technology has achieved a certain degree of maturity, its applications to the adaptivity research were rare. Evaluations of adaptive systems which would employ eye-movement tracking are indeed hard to find. Moreover, even though the gaze location recorded by eye-trackers has been used in real-time application, the aims were either to save the bandwidth of a channel through

which a graphics is transferred or to enhance the interaction by direct manipulation of cursor through new gaze-modality. In the present paper, we propose to fill the gap of knowledge about the eye-movement patterns in adaptive systems by suggesting two possible directions for future research into 1) evaluation of adaptive systems with a help of eye-movement tracking and 2) using the gaze as a new adaptation source. However, we believe that neither of the proposed directions can be taken immediately, without developing a methodological framework sensitive to the specific area of adaptive systems.

In the rest of this section we briefly introduce eye-movement tracking as a powerful tool for investigating visual attention. In section 2 we propose how eye-tracking could be integrated into evaluation of adaptive systems. The possibilities of using real-time gaze direction as a new adaptation source are outlined in section 3. In section 4 we discuss some of the problems and pitfalls of the proposed approach, and we present our conclusions in section 5.

1.1 Eye-movement tracking

Eye-tracker is a device that registers the movements of eyes via processing of reflections from infrared light shone to eyes. Two types of eye-trackers exist, (1) a remote, table mounted version, making no contact with users, or (2) a head-mounted version with a see-through mirror. In addition to the measurements of the movements of the eyes, most of the current eye-trackers can also provide an estimate of pupil size, users' distance from the eye-camera, and validity codes indicating the presence of the eyes within the field of view. Current eye-trackers are relatively cheap and can deliver the gaze location precisely; the data are usually sampled at rates between 50-250Hz.

Eyes are never perfectly still. In general, two types of eye movements, saccades and fixations, are identified from the protocol recorded by the eye-tracker [14]. *Saccade* is rapid and ballistic eye-movement that serves for repositioning the eyes onto a new location. Human visual system does not extract any information during a saccade; a phenomenon known as saccadic suppression. Information from a stimulus is extracted only during fixation, when the image of the investigated object falls onto the fovea. *Fixation* is a relatively stable position of eye, lasting about 300ms. During fixation the information is extracted from the observed object. Because the retina needs to be continuously refreshed, even during a fixation eyes perform microscopic movements. Other types of eye movements exist, for instance microsaccades or pursuits; for the description of these, see [8].

Another division of eye-movements can be done in terms of how they are initiated and controlled. We recognize either voluntary eye-movements, as for instance when one wants to keep a certain object on the retina, or involuntary, reflexive eye-movements, such as changes in the pupil size or the microscopic movements serving to refresh the image on the retina.

Studies investigating the allocation of visual attention of users completing experimental tasks have confirmed a strong relation between the direction of gaze and focus of visual attention. Particularly, the link between eye fixations and cognitive processes has been investigated [10], [13]. From these and other studies a general assumption has been derived that eye and visual attention are tightly linked. It is

believed that attention precedes the eyes so that after the information is extracted and current feature is processed, the attention is shifted to a new location and a saccade to the location is programmed and executed. Duration of a fixation has been shown to correlate with participants' difficultness to process the fixated object [5], indicating therefore the depth of the processing required to encode the information or the experience level [1]. Number of fixations, on the other hand, has been shown to reflect the importance of the interface object to the participants. Finally, the patterns of eye-movements, in terms of sequence of fixations and saccades to different object of a scene viewed, differ for the same scene when the task given to participants changes [15].

2 Eye-movement tracking and evaluation of adaptive systems

Eye-movement based evaluation of interaction is often conducted either in a retrospective way or it is based on some underlying cognitive model and hypothesis. A typical scenario includes experimental participants conducting their tasks while their eye-movements are measured. Experimenters manipulate with the features of the investigated task and then examine the eye-movements for significant patterns related to the manipulation.

Apart the domain of studies using eye-movement patterns, it is often the case that an evaluation of benefits of new techniques concentrates on measures of performance, completion time, frequency of errors, or preference. However, as the adaptive technologies aim to support users in carrying on their tasks, for instance during learning, we shall pay attention to how the underlying cognitive processing is influenced by the adaptivity.

It has been previously suggested that an evaluation of adaptivity shall be conducted at two distinct phases, recognized as interaction assessment and adaptation decision making [3]. Considering the former, eye-movement tracking itself is the source of rich interaction information and provides data with a high level of detail. In the latter phase, eye-trackers could be used to quantify whether the decision of the adaptive system were visually attended by the users.

In the following, we illustrate how the evaluation of adaptive systems could benefit from employing the eye-movement tracking. Two main approaches to adaptation, namely the adaptive navigation support and the adaptive presentation technologies, can be identified in current adaptive systems [2]. By involving the adaptive navigation support, the directions a user can take during learning are limited and proposed, or guidance is given to better support the learning process. In adaptive presentation of content, the materials shown to the user are modified to better suit the user according to the user model built. In any case, the user model is built and updated, as accurately as possible, so the adaptive engine can act upon it and provide the users the most relevant information to support achieving their goals.

2.1 Evaluating adaptive presentation of content

Some adaptive systems make the decision of what content shall be displayed to users, based on knowledge the user acquired during a previous interaction with the tool. For instance, the adaptation mechanism of a tool aiding the understanding of mathematical expression evaluation decides whether certain parts of an expression are understood well enough, so some other parts can be displayed with a focus. We propose that eye-movement tracking could help to estimate automatically the focus of attention on certain elements of such an expression. That means a study would compare the patterns of eye-movements on the elements or operations that are recognized by the adaptive engine to be already comprehended to those fixations falling on the elements that are thought to be not yet fully understood. The difference in the eye-movement patterns shall then indicate whether the decisions of adaptive algorithm indeed correlate with interests of users, and with problematic and less familiar parts the users were attending.

2.2 Evaluating adaptive navigation support

The decision made by the adaptive navigation support can be investigated using eye-movement tracking. Typically, some links are hidden when the adaptive system comes to a decision that the user is not ready to follow the links. On the other hand, some links are generated and/or annotated dynamically when the adaptive component decides that the user might benefit from the information behind the links. Two systems could be compared (with and without adaptive component), in terms of whether the annotation is attended by the users, or whether the presence of additional links creates a confusion or disturbance to otherwise unaltered cognitive processing. Clearly, eye-movement tracking can be employed in such studies; comparison of two or more adaptive systems or adaptive vs. non-adaptive system comparison with respect to the eye-movement data can be conducted. For example, eye-movement trackers allow for a measurement of cognitive workload, through the dilatations of the pupil [12]. These dilatations happen involuntary, and therefore provide an objective measure of users' cognitive processing and changes in the mental workload during competition of a task [6].

Previous eye-tracking research established and applied numerous eye-movement metrics (for an overview see [4]); however, not all of them may directly apply to adaptive system evaluation. Therefore, studies of interaction with adaptive systems and comparative studies with/without adaptive features have to be conducted in order to establish a body of knowledge about typical eye-movement patterns produced during the interaction and during the adaptation decision making. With conjunction with other data collected during the interaction, eye-movement tracking can then deliver a powerful tool for evaluators of adaptive systems.

3 Real-time gaze registration and adaptive systems

Knowledge level, learning style, preferences, goals, user and usage (interaction) data, are all typical sources of adaptation [2]. We believe, however, that the collection of users' actions cannot be complete without the awareness of what features of the interface were visually attended, what strategies the users exhibited, and what cognitive efforts they had to exert while completing their tasks. Considering the eye-movement tracking as a source of adaptation, the tool provides instant information about the location of user's visual attention. Presumably, the object under one's visual investigation is most of the time also located on the top of his/her cognitive processing stack. Therefore, knowledge about the location of the gaze in time and space helps us in understanding what features of the interface were of interest to the user, in what order, when, and how long the user needed to attend each of the objects.

Considering again the example of an adaptive tool for expression evaluation learning, the user modeling mechanisms of the tool could be enriched by knowing what parts of an actual expression caused users the greatest problems, measured, for instance, as the number of fixations paid to a certain element of the expression. Further, if some component of a complex expression was not attended at all during the learning process and the knowledge level related to the component indicates it shall be still processed before continuing, the tool can immediately act upon this information and ask the user whether he/she wishes to overcome the problem.

We see a great potential of using eye-movement tracking as a real-time adaptation source. However, similarly as in the previous section concerning the evaluation of adaptation systems, we believe that a thorough investigation of what type and patterns of eye-movements could be used and how they can be used has to be carried out first. We suggest that the eye-trackers shall be used first for evaluating the outcomes of adaptivity to create a set of measures appropriate for a specific application domain. Once the metrics are developed, it shall be possible to employ the real-time gaze collection and use the gaze data and inferred cognitive processing as a new source of adaptation.

4 Difficulties and pitfalls of eye-movement tracking

Although eye-movement tracking provides information which is inaccessible via other measurements, its application also involves certain difficulties that have to be taken into consideration prior the technique is applied. Although a number of issues exist, we briefly introduce here the most significant problems that hinder the widespread of eye-trackers.

4.1 Methodological issues

First and foremost, the methodological issues of proper analysis and interpretation of eye-movement data seriously influence the outcome. Given the typical sampling rate

of 50Hz, a 10 minutes recording generates as much as 1.5 Mbytes of eye-movement data to be analyzed. Although the methods for an automatic eye-movement data extraction (in terms of fixations and saccades identification) have been developed [e.g. 14], there is no standard way of interpreting the protocols and establishing their relation to the investigated task.

Eye movements are both voluntary and unconscious, although we usually execute them automatically. When fixations are used as means for selecting/controlling objects in an interface and for information acquisition, another methodological consideration arises, known as Midas touch [7]: the interface cannot certainly determine whether a fixation at an interaction widget is meant to issue a command or to purely extract information. Avoiding the Midas touch remains one of the greatest challenges in the eye tracking research.

4.2 Technological issues

The eye-movement data often contain a great deal of noise. The noise can be attributed to participants moving excessively their heads, wearing glasses, blinking, or to other factors, such as drift and inaccurate calibration. These all cause the eye-trackers to fail to obtain a video-image of the eye(s) and as a result not to report any eye-movement data for some time.

Another problem can be seen in the accuracy of the present eye-tracking systems. Most of the technologies can achieve the accuracy between 1 to 2 degrees of visual angle (the size of thumb-nail at about 90 cm distance), which is not enough considering the resolution of current displays. Therefore, it is hard to investigate how the visual attention is allocated to the small areas like a single line and words in this paragraph. However, both the previous problems are technological issues that can be solved in the next versions of currently available eye-trackers. A limitation in accuracy will, nevertheless, persist, due to the size of human fovea.

5 Conclusion

The aim of this paper was to raise a discussion and interest about an intersection of areas of eye-movement tracking research and adaptive systems evaluation and development. Although the ideas presented in this paper are not necessarily novel, we still believe that their consideration contributes to a more holistic approach to adaptive systems evaluation. We presented two of possible directions for future research.

First, we suggested employing visual attention tracking as one of the methodologies for evaluating adaptive systems. As the eye-tracking can be conducted without any interventions to users and their tasks, it is a powerful tool to investigate the patterns of visual attention and therefore related cognitive processing influenced by adaptive mechanisms.

Second, we proposed to use the gaze direction for building gaze-aware adaptive environments, where the eye-movement patterns are used as a new adaptation source. Adaptive systems could become aware of the intentions and attention of their users to

different parts of the interface. The process of modeling the users could benefit from such information to create more accurate user models.

We call for developing of the methodology enabling adaptive system research to fully utilize the potentials of eye-tracking. By doing so, the pitfalls and problems related to application of eye-movement tracking could be reduced. As we expect the price of the eye-tracking equipment to drop and making thus the technology available to a wider public, the eye-trackers will become a standard and common part of personal computers and other ordinary video-based systems. General-purpose adaptive systems could make a great use of the technology.

Acknowledgement

The author would like to thank Minna Kamppuri and Niko Myller for the kind comments on the earlier version of this paper.

References

1. Bednarik, R., Myller, N., Sutinen, E., Tukiainen, M.: Effects of Experience on Gaze Behaviour during Program Animation. Accepted to the 17th Annual Psychology of Programming Interest Group Workshop (PPIG'05), Brighton, UK, June 28 - July 1, (2005).
2. Brusilovsky, P.: Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction, 6 (2/3), (1996), pp. 87–129.
3. Brusilovsky, P., Karagiannidis, C., and Sampson, D.: Benefits of Layered Evaluation of Adaptive Applications and Services. In S. Weibelzahl, D. Chin and G. Weber (Eds.) Empirical Evaluation of Adaptive Systems. Proceedings of workshop held at the Eighth International Conference on User Modeling in Sonthofen, (2001).
4. Duchowski, A.: A Breadth-First Survey of Eye Tracking Applications. Behavior Research Methods, Instruments, and Computers, (2002).
5. Goldberg, J. H., & Kotval, X. P.: Computer Interface Evaluation Using Eye Movements: Methods and Constructs. International Journal of Industrial Ergonomics, 24, pp. 631-645, (1999).
6. Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., and Brian P. Bailey: Towards an index of opportunity: understanding changes in mental workload during task execution. In proceedings of CHI '05: Proceeding of the SIGCHI conference on Human factors in computing systems, Portland, Oregon, USA, pp. 311-320, (2005).
7. Jacob, R. J. K.: Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces. In Advances in Human-Computer Interaction, Vol. 4, ed. by H.R. Hartson and D. Hix, pp. 151-190, Ablex Publishing Co., Norwood, N.J., (1993).
8. Jacob, R. J. K.: Eye Tracking in Advanced Interface Design. In Virtual Environments and Advanced Interface Design, ed. By W. Barfield and T.A. Furness, Oxford University Press, New York, 258-288, (1995).
9. Jacob, R. J. K., Karn, K. S.: Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises (Section Commentary). In The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research, ed. by Hyöna, J., Radach, R. and Deubel, H., pp. 573-605, (2003).

10. Just, M.A., and Carpenter, P.A.: Eye Fixations and Cognitive Processes, Cognitive Psychology, 8:441-480, (1976).
11. Majaranta, P., and Räihä, K.-J.: Twenty Years of Eye Typing: Systems and Design Issues. In Eye Tracking Research and Applications (ETRA) Symposium, (2002).
12. Marshall, S. S.: The Index of Cognitive Activity: Measuring Cognitive Workload. In Schmorrow, D.: Tomorrow's Human Computer Interaction from Vision to Reality: Building Cognitively Aware Computational Systems. Symposium presented at IEEE 7th Conference on Human Factors and Power Plants, (2002).
13. Rayner, K.: Eye Movements in Reading and Information Processing: 20 years of Research. Psychological Bulletin, 124, 372-422, (1998).
14. Salvucci, D.D. & Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the Eye Tracking Research and Applications (ETRA) Symposium 2000, pp. 71-78, (2000).
15. Yarbus, A.L.: Eye Movements and Vision. NY, Plenum Press, (1967).

The Evaluation of In-Vehicle Adaptive Systems

Talia Lavie, Joachim Meyer, Klaus Bengler, Joseph F. Coughlin

[1]Department of Industrial Engineering and Management
Ben Gurion University of the Negev
Beer Sheva 84105, Israel
{tlavie, Joachim}@bgu.ac.il
[2] BMW Forschung und Technik, GmbH, Munich, Germany
Klaus-Josef.Bengler@bmw.de
[3] AgeLab, Massachusetts Institute of Technol ogy (MIT),
Cambridge MA, USA
(jmeyer, Coughlin)@mit.edu

Abstract. Although research on adaptive systems has begun only recently, studies have shown the benefits of using adaptive systems. However most of those studies have examined systems with and without adaptive qualities, disregarding additional factors that may influence the interaction. This study presents a first step towards a more comprehensive evaluation of adaptive systems. We assert that adaptive systems should be examined with regard to different types of tasks, different situations and using various users to be able to determine the conditions in which adaptivity will be beneficial. A preliminary study evaluated adaptivity when performing routine and infrequent tasks. The study showed that adaptivity is beneficial for routine tasks, and that adaptivity impairs performance of infrequent tasks. The study proposes a method to calculate the point at which adaptivity ceases to be beneficial as a function of the relative frequencies of different tasks and provides a starting point for a more comprehensive understanding of the subject.

1. Introduction

Adaptive user interfaces (AUI) are designed to support users in performing their tasks by adapting to their individual characteristics. AUIs can facilitate user performance, make the interaction more efficient, improve ease of use and assist the user in overcoming information overflow and help them use complex systems [2]. However, adaptation has also some limitations, usually related to usability problems, as the user her/himself is an adaptive "system". Such problems include: lack of control the user may feel regarding the system appearance or functions [6], [7], lack of consistency [8], [9], [11], and lack of transparency and predictability [6], [7]. In addition to these problems, [7] suggested two other problems. First, the adaptive system might place demands on the users' attention, therefore reducing the capability to focus on the system's main task. He referred to this problem as Unobtrusiveness. Secondly, he mentioned that the adaptive system might impair the user's breadth of experience because some types of adaptive systems assist the user by acquiring

information or perform parts of the task instead of the user. Therefore, users may become less knowledgeable in a certain domain (i.e. knowledge degradation). This may also lead to over reliance on the system by the user who believes that the choices made by the system are always relevant and good [5], [7].

Given that adaptive systems have the limitations mentioned, it is valuable to demonstrate that adaptivity indeed improves the interaction with the system, and under what circumstances such an improvement will occur. Therefore, the evaluation of such systems is of great importance and should be as comprehensive as possible.

Our evaluation of the benefits of adaptivity will focus on adaptivity in in-vehicle telematic systems. These systems are now standard equipment in high-end cars. They combine a variety of functions in a single user interface, including access to the navigation system, traffic advise, entertainment (CD, radio, satellite radio, MP3, etc.), climate control, communications (cellular phone, SMS, email access, web access, etc.). The population of drivers is a highly diverse user population in terms of age, cognitive abilities, skills, computer experience, etc., and it therefore is very appealing to adjust systems to the properties and preferences of the individual driver. One way to achieve this goal would be by incorporating adaptive functions in such systems. These functions need to meet two basic requirements: 1. They should facilitate the interaction with the system and improve driver satisfaction with it. 2. They should not increase the distraction (and consequent safety problems) caused by the system, and should ideally even lower distraction. A number of papers have addressed the use of adaptivity in in-vehicle devices (e.g., the "adaptive route advisor" by [10]).

1.1 The Evaluation of Adaptive Systems

The evaluation of AUIs refers mainly to the effectiveness of the systems and whether they meet usability criteria. The effectiveness is usually determined by the quality of the information the system provides, its accuracy, performance time and users' subjective evaluations. This off course depends on the specific characteristics of the adaptive system. The extent to which these systems meet usability criteria is usually evaluated through traditional HCI usability variables such as consistency, transparency, learnability, predictability etc. To date, the evaluation of adaptive systems is still in its infancy and only few studies have evaluated such systems empirically. Additionally, most studies up to now compared an adaptive and a non-adaptive system on a number of variables, examining whether the adaptive system has some advantage over the non-adaptive system (e.g., [3], [4], [6] [12], and [13]).

The evaluation of adaptive systems needs to cope with a number of problems that are particularly crucial in this context. First, the benefits and limitations of using adaptivity are likely to appear only after fairly prolonged use. Short experiments or observations may fail to provide an adequate picture. Second, the dynamic changes in system properties that result from adaptivity may have differential effects in different situations, while performing different tasks and on different users. For instance, adaptivity may be more beneficial for complex tasks. Similarly, while some users may consider adaptivity to have advantages, others may find it confusing and prefer to have it turned off. Therefore, the evaluation might consider properties of the user,

the task and the situation, since it is not enough to establish for a specific system whether the adaptive version is better compared to the same system without adaptivity. Additionally, there are various types of adaptive systems, ranging from systems that support system use, like adaptive menus, to systems that support information acquisition, such as adaptive filtering systems (see [7]). Fig. 1 presents the variables we claim influence the interaction with an adaptive system and therefore should be considered when evaluating a system.

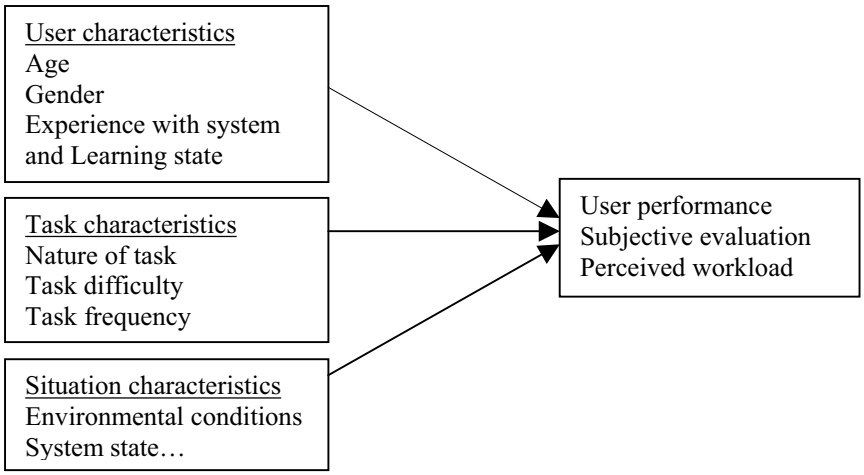


Fig. 1. The variables assumed to influence the interaction with an adaptive system

We assert that a framework describing the conditions in which adaptivity will be most beneficial should be generated and examined. Such a framework requires the evaluation of adaptive systems in a number of steps. The first step should examine adaptivity when performing different types of tasks, such as routine and uncommon tasks, tasks with different levels of difficulty, etc. The second step should examine adaptivity in different situations, such as different environmental conditions. Finally, the third step should evaluate system use by various types of users differing in age, level of expertise with the system, etc. We claim that the frequency at which a task needs to be performed has major influence on the degree in which the system will be beneficial.

1.2 Task Frequency

The aspect of task frequency has been examined before in other domains, such as adaptable systems and mainly in the study of automation, but has not yet been examined with relation to adaptive user interfaces. Previous research has raised the value of task frequency. [1] For example, compared two interfaces. In one interface the user can customize all the items in the menu (the user adds all features to the menu) for both frequent and infrequent tasks he or she will need to perform. In a second interface the user customizes only the features necessary for the more

frequently performed tasks (the user needs to switch to the full interface to perform infrequent tasks). [1] Found that when users perform a task infrequently, adding all items is not always as efficient as adding only those from the frequent task. Adding infrequently used items depends on a number of factors including the number of infrequently used features, where these features will be located in the menus, the ratio at which the infrequent features will be used compared to the frequent features and the user's expertise.

1.3 The Study

This paper will describe a step towards developing a framework for evaluating adaptive systems by examining the first variable we assert influences the interaction. We assert that adaptive systems should be more beneficial to the user when performing routine and frequent tasks. On the other hand, when the user is required to perform an uncommon and infrequent task, the adaptive system will most likely cease to be advantageous and even may become a burden on the user. The purpose of this paper is to discuss the issue of task frequency and more specifically to provide a method to calculate the point from where adaptivity will no longer be beneficial and may impair performance.

2. Method

2.1 Participants

Twenty engineering students at Ben-Gurion University of the Negev, Israel, served as paid participants in this study.

2.2 Apparatus

An experimental system, which consists of two subsystems, was developed: a driving simulator and a telematic system simulating in-vehicle devices. The system was PC based and was developed in Visual Studio.Net 2003. It displayed a road scene on a 21-inch monitor located in the center of the participant's visual field. The simulator showed a two-lane curved road without additional traffic. The car position in the lane was controlled through a steering wheel and it drove at a constant speed of approximately 30 km/h. The in-vehicle telematic system was simulated through a visual display (16 cm wide X 9 cm high) that was displayed on a separate 15-inch screen to the right of the driving simulator screen. The telematic system included three subsystems: a communication system (including SMS, Outlook, News Updates), an entertainment system (including radio and CD) and a navigation system (including traffic updates). Participants controlled the telematic system using buttons located on

the steering wheel (left, right buttons for navigation in the telematic system and a button for selection). The integrated system was connected to an output data file that contained data on driving performance (the driver's steering actions and lateral lane position were recorded every 200 mSec) and on task performance with the telematic system. To assess the drivers' subjective evaluations of the system they were asked to respond to three questionnaires, one at the end of each drive.

2.3 Experimental Design

A 2 X 3 X 5 X 2 between-within experimental design was employed. The between subject variable was the manual condition compared to the adaptive condition. The within-subject variables included: the number of drives (2 routine and 1 uncommon drive), 5 tasks (traffic updates check, SMS reading, news updates reading, e-mail checking and CD change) and 2 occurrences of all tasks (first, second). The dependent variable was the performance with the telematic system. Performance time was measured in milliseconds for all participant actions with the telematic system.

2.4 Procedure

Participants were requested to perform tasks with the telematic system while driving the car. The experiment began with an introduction drive in which the user became acquainted with the system and the tasks in the manual mode. After completing the introduction drive, participants drove 2 routine drives and 1 uncommon drive. Each routine drive included 12 tasks the user was asked to perform (5 tasks that occurred twice and additional 2 tasks in which lane shifts were required). The uncommon drive included 3 uncommon tasks in addition to the routine tasks. Ten participants performed the tasks in the manual mode and ten in the adaptive mode. During the drive the participants received a text message in the top section of the telematic system that specified the required task. For example, the system notified the participant that she received an SMS message and she was requested to reply with a message "I'm driving". In the manual condition, the participants were requested to reply manually by typing their reply on a virtual keyboard. In the adaptive condition, the system automatically sent the participants' usual response. The text messages were always accompanied by an auditory message. The appearance of the next task was conditioned on the successful completion of the previous task. All 4 drives (introduction, 2 routine and 1 uncommon) took place in one experimental session that lasted about 90 minutes, with 5-minute breaks between the drives and time for filling out the questionnaires. The participants performed the following tasks with the telematic system: Receiving an SMS message and sending a reply, reading e-mail from the inbox, receiving news updates, receiving traffic updates, and changing from radio to CD. The participants received some instructions prior to their drive, informing them about their regular use of the telematic system.

3. Results

Performance time was measured in milliseconds as the time from the moment the message appeared on the screen until the participant completed the task. Performance time was analyzed for two types of tasks:

1. Constant tasks: tasks that did not include uncommon actions and therefore did not change along the three drives. These tasks included checking traffic updates, checking news updates and one instance of reading an email message.
2. Changing tasks: tasks that included uncommon actions during the third drive. These tasks included the 2 SMS messages received and one instance of checking email in which the participants were required to change the user in the inbox.

Analyses on both tasks used a 2-way ANOVA with repeated measures on the number of drive variable (3 drives). The between factor was adaptivity (Manual and Adaptive).

The results of the ANOVA performed on the constant task showed an interaction Adaptivity X Drive ($F(2, 28)= 5.99$, $p<0.0001$). Fig. 2 presents the results. The results show that in all drives performance times were better in the adaptive condition, although they significantly improved in the third drive in the manual condition.

The results of the ANOVA performed on the changing tasks showed an interaction Adaptativity X Drive ($F(2, 28)= 74.45$, $p<0.0001$). Fig. 3 presents the results. The results show that in the first two drives, which include the routine tasks, performance times in the adaptive condition were much faster, while in the third drive, where the tasks were infrequent, performance times in the adaptive condition were significantly longer, compared to the manual condition.

We propose that an additional factor related to the frequency of the task relates to the ratio at which routine and infrequent tasks occur. The greater the ratio of the routine tasks to the infrequent tasks, the more adaptivity should be beneficial. The calculation of the costs and benefits of adding adaptivity to a system can be demonstrated on the results of our experiment.

As can be seen in Fig. 3, the mean time required to perform the tasks in drives 1 and 2 (in which only routine tasks needed to be performed) was 43.25 seconds for the manual condition. The use of adaptivity improved the time to 22.63 seconds, so that we can state that introducing adaptivity shortened the performance times to approximately half their value without adaptivity. In drive 3, when non-routine tasks needed to be performed, times remained approximately the same for the manual condition (49.31 seconds) whereas performance time increased to 72.69 seconds for the adaptive condition. Thus adaptivity increased performance times in non-routine tasks by approximately 50%. We do not state that the symmetry in the effects (where adaptivity shortens times for routine tasks by half and raises times for non-routine tasks by half, as well) will be found whenever adaptive systems are evaluated, but it can serve as a convenient first approximation of the effects of adaptivity.

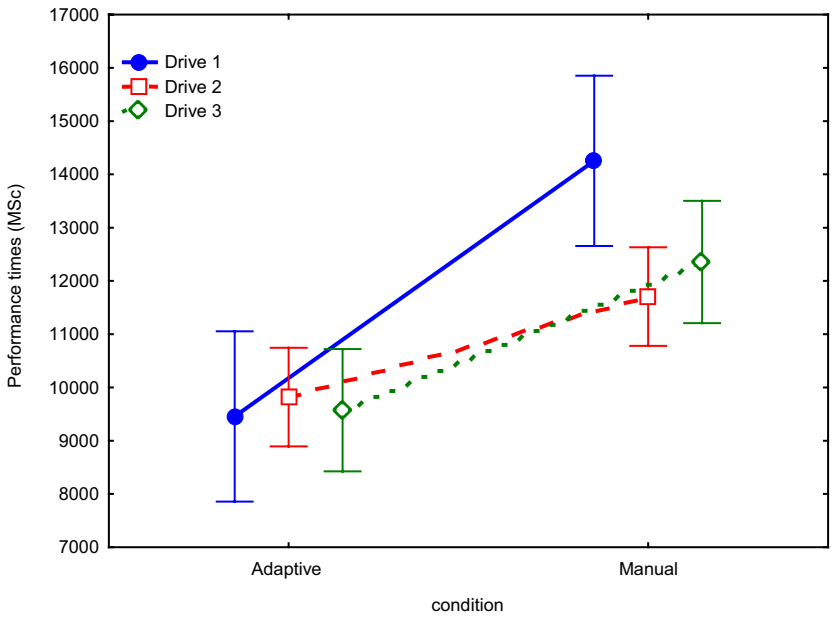


Fig. 2. Mean time to perform constant tasks in the manual and adaptive conditions

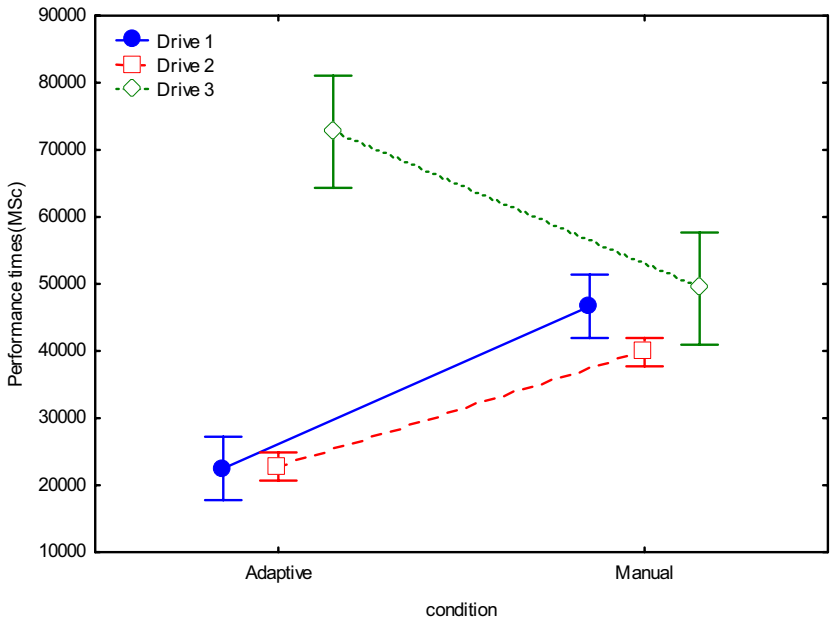


Fig. 3 Mean time to perform changing tasks in the manual and adaptive conditions

It is now possible to compute some estimate for the effects of adaptivity as a function of the proportion of frequent tasks out of all tasks that need to be performed. A measure of the total performance time T_{Total} can be computed from the expression

$$T_{Total} = p_{Freq}(1 - B) + (1 - p_{Freq})(1 + C) .$$

where p_{Freq} is the proportion of frequent tasks, B is the benefit from adaptivity for frequent tasks, and C is the cost of adaptivity for non-frequent tasks. Costs and benefits in our case are the degree of change in performance time after introducing adaptivity for frequent and infrequent actions. T_{total} in this case is the ratio between the performance time with adaptivity and without it, so that $T_{total}=1$ when adaptivity has no effect on performance time, $T_{total}<1$ when adaptivity shortens performance times, and $T_{total}>1$ when adaptivity lengthens performance times. In our case we can set B=C=.5. The resulting computation is shown in Fig. 4. Clearly, in this very simple case, system performance will benefit from installing adaptivity if the proportion of frequent tasks out of all tasks that need to be performed with the system exceeds 50%.

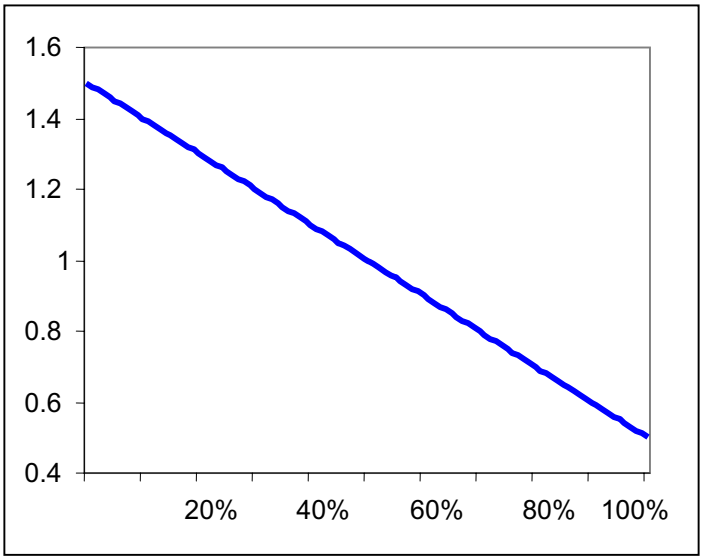


Fig.4. Performance time ratio as a function of the ratios of routine and non-routine tasks. The value 1 represents the performance level without adaptivity. Shorter times indicate faster and therefore better performance.

4. Discussion and Summary

Adaptive user interfaces were shown to be beneficial in empirical studies that compared an interaction concept in an adaptive versus non adaptive version. However, a number of additional factors influence the interaction with adaptive systems and are likely to affect the value of adaptivity. We call for a more comprehensive examination of adaptive systems that should lead to the development of guidelines that specify the conditions in with adaptivity will be beneficial. These conditions should be based on the analysis of the set of tasks that need to be performed with the system, the various usage situations, and the characteristics of the individual users.

Our study is a first step towards achieving this goal. It evaluates adaptivity as a function of the ratio of routine and infrequent tasks. We suggested that adaptivity will be beneficial when routine tasks are to be performed and will impair performance when infrequent tasks arise. The results of our study support our assumptions and showed that indeed adaptivity improves performance of routine tasks and impairs performance of infrequent tasks. We demonstrate that for a given adaptivity algorithm the relative value of adaptivity can be assessed, given the benefits of adaptivity for the performance of frequent tasks, the costs due to adaptivity for infrequent tasks, and the relative frequency of frequent tasks.

This study presents only a preliminary evaluation of the subject. Clearly more experiments are needed to replicate and expand the results. For instance, we assume here that the costs and benefits are independent of the relative frequency of the frequent task. This assumption may hold after prolonged experience in using a system, but may actually require closer scrutiny in the early stages of learning system usage. Future research should examine adaptivity using different frequencies of routine and infrequent tasks. Also, the effects of the different categories of variables that we identified as affecting the performance with adaptive systems should be examined empirically. We hope that by gradually accumulating a set of empirical results on the performance with adaptive systems for different tasks in different usage situations and by different users, we will be able to develop a comprehensive model from which the outcome of installing adaptive functions in a system can be predicted. Such models should have great value for system designers in general. They may have particularly great appeal for the interaction design of e.g. in-vehicle systems, where issues of adaptivity and user performance have major impact both on the appeal of the system to the driver population and the safety of the use of the system.

5. References

1. Bunt, A., Conati, C. and McGrenere, J.: What Role Can Adaptive Support Play in an Adaptable System? IUI 04, January (2004) 13-16
2. Engström, J., Arfwidsson, J., Amditis, A., Andreone, L., Bengler, K., Cacciabue, P.C., Eschler, J., Nathan, F., Janssen, W.: Meeting the Challenges of Future Automotive HMI

Design: An Overview of the AIDE Integrated Project. In Proceedings of ITS congress Budapest (2004)
3. Gong, Q., Salvendy, G.: An Approach to the Design of a Skill Adaptive Interface. International Journal of Human-Computer Interaction, Vol. 7(4) (1995) 365-383
4. Greenburg, S., Witten, I.H.: Adaptive Personalized Interfaces – A Question of Viability. Behavior and Information Technology, Vol. 4 (1985) 31-45
5. Hook, K.: Evaluating the Utility and Usability of an Adaptive Hypermedia System. Knowledge Based Systems, Vol. 10 (1998) 311-319
6. Hook, K.: Designing and Evaluating Intelligent User Interfaces. In Proceedings of the IUI 99 ACM Press (1999)
7. Jameson, A.: Adaptive Interfaces and Agents. In Jacko, J.A. Sears, A. (eds.): Human-Computer Interface Handbook. Mahwah, NJ, Erlbaum (2003) 305-330
8. Keeble, R.J., Macredie, R.D.: Assistant Agents for the World Wide Web Intelligent Interface Design Challenges. Interacting with Computers, Vol. 12 (2000) 357-381
9. Kuhme, T.: A User-Centered Approach to Adaptive Interfaces. Proceedings of ACM IUI '93 (1993) 243 - 245
10. Rogers, S., Fiechter, C., Langley, P.: An adaptive Interactive Agent for Route Advice. Proceedings of the Third International Conference on Autonomous Agents, Seattle: ACM Press (1999) 198-205
11. Shneiderman, B.: Direct Manipulation for Comprehensible, Predictable and Controllable User Interface. Proceedings of the 1997 International Conference on Intelligent User Interfaces. ACM Press (1997)
12. Te'eni, D., Feldman, R.: Performance and Satisfaction in Adaptive Websites: An Experiment on Searches within a Task-Adapted Website. Journal of the Association for Information Systems, Vol. 2/(3) (2001)
13. Trumbly, J. E., Arnett, K. P., Johnson, P.C.: Productivity gains via an adaptive user interface: an empirical analysis. Human-Computer Studies, Vol, 40, (1994) 63-81

Is the ACT-value a valid estimate for knowledge? An empirical evaluation of the inference mechanism of an adaptive help system

Dorothea Iglezakis

Cath. University Eichstaett-Ingolstadt,
Department for Applied Computer Science,
Ostenstr. 14, 85072 Eichstaett, Germany,
dorothea.iglezakis@ku-eichstaett.de

Abstract. This paper reports the results of an empirical study that evaluates the inference mechanism of an adaptive help system for web-based applications. The help system adapts to a measure for the procedural knowledge that is computed from activity logfiles according to the ACT-theory of Anderson and Lebière [1]. The results of the study show that the ACT-value of procedural knowledge correlates with subjective and objective measures of performance and proves itself as a better estimate of the procedural knowledge than general computer knowledge, a measure often used by other adaptive help systems.

1 Introduction

Help systems are not always as helpful as they should be. In a survey, users of a newly introduced software rated human helpers much higher than the online help, but asked for a better help function [7]. The help of SPSS even worsens the performance of the users [6]. On the other hand, a good online help is an important source of contentedness of customers. Adaptive help systems are one of many approaches to enhance the quality and the helpfulness of help systems. Existing systems either try to adapt the contents of the help by adapting to the knowledge of the user or try to reduce the information by adapting to the goals and plans of the user [5]. The used models of users knowledge are either global stereotypes or overlay models. Global stereotypes are only capable of domains, where the concepts are clearly ordered by difficulty. Overlay models assign a value or a probability to each concept, how well or how probable a user knows this concept. The overlay models entail much more information than the global stereotype models, but also use static knowledge. If a user proves to know a concept, the model assumes this knowledge until the user proves the opposite. None of the adaptive help systems use forgetting in their user models.

Cognitive psychology offers a lot of research about knowledge and forgetting. The adaptive help system evaluated in this paper tries to use the results of this research to offer an alternative way to model the knowledge of the user.

2 CHEetah—An Adaptive Help System for Web-Based Applications

CHEetah is an adaptive help system for web-based applications that adapts to the procedural knowledge of the user. The user model is a long term model that contains a measure of the procedural knowledge of the user for every functionality in the target application. The adaptation happens through showing and hiding different parts of the help items.

2.1 Adaptation target

CHEetah adapts to the procedural knowledge of the user. To model the knowledge of the user as accurately as possible, results from cognitive psychology are used. The ACT-theory from Anderson and Lebière [1] delivers an empirically founded theory[1] about the learning and the forgetting of memory contents. The activation of a memory item corresponds to the accessibility of a concept and is influenced by the frequency of access. Each access of a memory item increases the activation of the corresponding memory trace. Without access, the activation fades over time.

Knowledge items differ in the type of knowledge. Procedural knowledge covers knowledge about actions and activities, whereas declarative knowledge covers knowledge about facts and world knowledge. Especially for help systems that have to explain and to support the execution of tasks, procedural knowledge is more central than declarative knowledge.

A knowledge item in the context of CHEetah is the procedural knowledge about one functionality of the target application. For example, if the target application is the configuration menu of an Internet provider, one knowledge item would be the knowledge about adding a new e-mail address. Declarative knowledge items can function as prerequisites or additional knowledge to the procedural knowledge items, but the procedural knowledge items are the focus of adaptation.

2.2 User Model

The ACT-theory incorporates the strength accumulation equation that computes the activity of a memory item in dependence on the time stamps of the last contacts with this memory item. A memory item in the context of CHEetah is the procedural knowledge about one functionality of the target application. The user model entails a ACT-measure of every functionality of the system for every user. If a user has never used a functionality, the corresponding ACT-value is 0. Otherwise, the ACT-value is computed according to the following equations.

$$a_i^z(T) = \sum_{j=1}^{n} t_j^{-d} \tag{1}$$

[1] See http://act-r.psy.cmu.edu/publications/ for a list of publications about the ACT-theory and evaluations with the ACT-theory

The strength accumulation equation (1) computes the activation a_i of a knowledge trace at the time stamp z in the following way. T is a set of time stamps that contains all contacts with a functionality i. t_j is the time in minutes that passed since the j^{th} contact with the functionality i. The parameter d can be chosen from the interval $]0;1[$ and is dependent on the application. From literature, a $d = 0.5$ is a good estimate for many applications.

If the user model already specifies an ACT-value from a computation on a time stamp z_0 in the past, the value is updated at the time stamp z according to the following equation 2:

$$a_i^z(a_i^{z_0}, T) = a_i^{z_0} * (z - z_0)^{-d} + \sum_{j=1}^{n} t_j^{-d} \tag{2}$$

$a_i^{z_o}$ is the old value, T again the set of contacts with the functionality i between the old time stamp z_0 and the actual time stamp z.

The strength accumulation equation builds on the general activity function of the ACT-theory that we used in [5], but omits the logarithm. Therefore, the strength accumulation equation avoids the problem of negative ACT-values for very old experiences.

2.3 Input Data

The ACT-theory needs the time stamps of every contact with a memory item—in our context a functionality of the target application—to compute the ACT value. One contact with a functionality is a complete execution of this functionality. As CHEetah concentrates on web-based applications, the input data of CHEetah are logfiles of the web server on which the target application runs. We use an extended logfile format in XML that saves for every visited page the user-ID, the page name and the used parameters.

From these data, CHEetah recognizes the execution of functionalities through predefined processes. Each process specifies one possible execution of a functionality in form of a regular expression. The output of the process recognition is a process ID, a start time-stamp and an end time-stamp of every successful execution of a functionality of the target application. A contact with the functionality in the sense of the ACT-theory is therefore the end time-stamp of a successful execution of a corresponding process.

2.4 Adaptation

The adaptation of CHEetah happens through online assembling of different components of help items. This form of adaptation has proven successful in other adaptive help systems such as PUSH [4] and EPIAIM [2].

The main advantage lies in the simplification of writing and maintaining the help items. As Höök [4] states, it is very difficult to write and to maintain different versions of the same help items. The online assembling of the help items from standardized components like EPIAIM or PUSH is therefore more promising.

The help items of CHEetah are written in an XML format that bases on the Docbook [8] format. Each help item corresponds to one functionality of the target application. The adaptation happens through online assembling of the different components to one

help item, dependent of the actual activation of the target process and its subordinate concepts and processes.

2.5 Summary

CHEetah has a concept of knowledge that is oriented on the results of cognitive psychology. The most important difference to other adaptive help systems is the fact, that CHEetah incorporates forgetting. If a user does not use a concept for a longer time, the knowledge about this concept is no longer available but nevertheless not vanished. The key concept of CHEetah is to show the user only these parts of the help information that are most valuable for him at this specific moment. This knowledge concept does not only allow to help novices, but also differentiates between different grades of expertise.A user that has forgotten the use of a functionality needs a different kind of help than a user that has never used this function before.

3 Empirical study

Weibelzahl [9] developed a framework for evaluations of adaptive systems that recommends evaluations on every step of the system:

- Evaluation of input data
- Evaluation of the inference mechanism
- Evaluation of adaptation decisions
- Evaluation of total interaction

In my opinion, an evaluation of the input data mechanism is not necessary, because there is no inference or uncertainty in this step. CHEetah uses the logfiles of the webserver that protocols the page visits of every user.

In contrast, the computation of the ACT-values as a measure of procedural knowledge needs evaluation. Therefore, we conducted an empirical study to answer the following questions:

- Is the ACT-value a reliable and valid measure of procedural knowledge?
- Does the ACT-value correspond with subjective and objective measures of the knowledge of the user?
- Is the ACT-value a better estimate of procedural knowledge than general computer experience?

To answer these questions, a web-based questionnaire was conducted. Participants of the study were 16 employees of the hmd software-AG, 6 male and 10 female. The target application for these evaluation was WebTime, a web-based calendar and task organization tool. The usage of WebTime in the hmd software-AG was logged over the period of 11 months. From these logfiles, the ACT-value of two different functionalities from the target application was computed. The functionality NewAppointment is a very basic task used very often by all participants. The functionality NewTodo was used more rarely and by a lower number of participants.

3.1 Design

The study wants to show if the ACT-value is a valid measure of the procedural knowledge of the user. The procedural knowledge according to one functionality is operationalized through subjective and objective measures of performance of this functionality.

Variables Therefore, the dependent variables were the subjective and the objective measure of the performance of the participants in the two different functionalities of the target application. Independent variables were the ACT-values of the participants of the two functionalities and—as a control variable— the general computer expertise of the participants.

Operationalization For the subjective performance measures the participants rated their expertise of the two functionalities on a six-step rating scale.

For the objective measurements, the participants were asked to perform the two functionalities on two examples. The performance of the users was logged and afterward rated on correctness and time. The correctness was rated on the following four-step scale:

- The task was successfully executed
- The task was executed with errors
- The task was started, but not finished
- The task was not started

Time was measured through the difference between the first step of the functionality and the final step of the functionality.

Procedural knowledge was represented by the ACT-values, computed from the log-files of the last eleven months according to equation 1.

Measure	*M*	*SD*	Min	Max
subj. performance NA	3.23	1.54	0	5
obj. performance NA	3.00	0.61	1	4
ACT value task NA	0.89	1.19	0.02	6.12
subj. performance NT	2.20	1.62	0	4
obj. performance NT	2.89	0.80	1	4
ACT value NT	0.02	0.04	0	0.17
$comp_h$	0.51	1.02	−1.25	3
$comp_t$	4.26	1.91	1	6

Table 1. Mean values and standard deviation for performance measures and procedural knowledge for the tasks NewAppointment (NA) and NewTodo (NT) and computer experience

General computer experience was rated through two different measures, a scale for helplessness with the computer ($comp_h$) and a guttmann scale [3] for activities ($comp_t$)[2].

Table 1 shows the mean values and standard deviation for the different variables.

3.2 Results

To get the relationship between the dependent and independent variables, correlations between subjective and objective measures of performance and the ACT-values and the correlations between the performance measures and general computer expertise were computed. Tables 2 and 3 show the results.

Correlation between	**r**	**Significance**
ACT — subjective expertise	0.19	n.s.
ACT — objective performance	0.09	n.s.
ACT — time	−0.06	n.s.
$Comp_h$ — subjective expertise	−0.32	n.s.
$Comp_h$ — objective performance	−0.59	$p < 0.01$
$Comp_h$ — time	−0.27	n.s.
$Comp_t$ — subjective expertise	−0.27	n.s.
$Comp_t$ — objective performance	−0.48	$p < 0.05$
$Comp_t$ — time	−0.27	n.s.

Table 2. Results of functionality NewAppointment

For the functionality NewAppointment, none of the correlations between ACT-value and performance measures is significant. The correlation between objective performance and ACT-value is near zero, but the correlation between subjective expertise and ACT-value shows an expected trend, though not very high. The correlations between general computer expertise and performance measures are higher, but in an unexpected direction. The correlation tendentially shows that the higher the general computer expertise, the lower is the subjective and objective performance.

For the functionality NewTodo, the results are more as expected. The correlation between subjective performance and ACT-value is significant ($p < 0.05$) and with a value of $r = .66$ relatively high. So the higher the ACT-value, the better the participant rates his expertise in the corresponding task. The correlation between objective expertise and ACT-value is not significant, but shows the right direction and has a reasonable value of $r = .30$. The correlation between ACT-value and time shows similar results. The higher the ACT-value, the faster the participants executed the task. The corresponding correlation of $r = -.23$ was also not significant.

[2] The questionnaire with the scales can be found at http://o14-0er-inf1.ku-eichstaett.de/hmd/befragung/index.php. Log in with the user ”testuser” and the password ”testpass”

Correlation between	r	significance
ACT — subjective expertise	**0.66**	$p < 0.05$
ACT — objective expertise	0.30	n.s.
ACT — time	−0.23	n.s.
$Comp_h$ — subjective expertise	−0.25	n.s.
$Comp_h$ — objective expertise	−0.15	n.s.
$Comp_h$ — time	0.34	n.s.
$Comp_t$ — subjective expertise	−0.36	n.s.
$Comp_t$ — objective expertise	−0.16	n.s.
$Comp_t$ — time	0.30	n.s.

Table 3. Results of functionality NewTodo

The correlations between general computer expertise and performance showed again unexpected results. All correlations showed an unexpected direction. The higher the general computer expertise, the lower the subjective and objective performance of the users.

3.3 Discussion

The results were not as clear as desired, but nevertheless show interesting effects. The ACT-value seems to predict better the subjective expertise of a user than the objective performance. But all correlations were in the expected directions. The low results for the task NewAppointment can be explained by a ceiling effect. As this task is often used by all users, every participant mastered this task without problems. In contrast, the results for the task NewTodo show, that the ACT-value is capable to predict the subjective expertise quite well, independent of the general computer expertise. The unexpected results can be explained by the unusual composition of participants. The participants with high computer expertise used the target application less than the participant with low expertise. Therefore, it is not so unexpected that the participants with high computer expertise performed worse than the participants with low computer expertise. Unfortunately, this fact complicates the interpretation of the positive correlations between ACT-value and performance measures. As the values of these correlations stay nearly the same for partial correlations, where the influence of general computer expertise is deducted, the general statement lasts.

All in all, the results are promising. The procedural knowledge represented by the ACT-value seems to be a better performance estimate than general computer expertise, especially for subjective performance. The results are better for tasks that are used infrequently than for tasks that are used very often. But infrequently used tasks are much more the focus of a help system, than task that are used very often. The low number of participants can explain the lack of significance of most of the correlations, but also limits the power of the results. The modeling of the user's knowledge through ACT-values seems to go in the right direction. Further evaluations will show, if help systems that adapt to ACT-values successfully support the user in a helpful and non-frustrating way.

3.4 Further Work

According to the framework described in section 3, the next steps are the evaluation of the adaptation decisions of the system and the evaluation of the total interaction between user and system. The evaluation of the adaptation decisions is accomplished already. The first results show that people with different ACT-values prefer different components of the help items. The evaluation of the total interaction between user and system is in the planning phase.

4 Acknowledgments

I am grateful to the management and especially the employees of the hmd software-AG, who participated in this study and allowed me to protocol their actions in WebTime. It is not easy to find long-term participants, so I am very thankful for their cooperation. I am also thankful to Jörg Desel, Ioannis Iglezakis, and Richard Mote for patient support and fruitful discussions.

References

1. John R. Anderson and Christian Lebiere. *The atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ, 1998.
2. Fiorella de Rosis, Bernardina de Carolis, and Sebastiano Pizzutilo. User-tailored hypermedia explanations. In *Adaptive hypertext and hypermedia - Workshop held in conjunction with UM'94*, 1994.
3. L. Guttmann. A basis for scaling qualitative data. *American Sociological Review*, 9:139–150, 1944.
4. Kristina Höök. *A Glass Box Approach to Adaptive Hypermedia*. PhD thesis, Stockholm University, Swedish Institute of Computer Science, 1996.
5. Dorothea Iglezakis. Adaptive help for webbased applications. In *Adaptve Hypermedia and Adaptive Web-based systems*, pages 304–307, 2004.
6. Mark Neerincx and Paul de Greef. How to aid non-experts. In *Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems*, pages 165–171, New York,NY,USA, 1993. ACM, ACM Press. SPSS help even worsens performance of the user.
7. Yvonne Waern, Nils Malmsten, Lars Oestreicher, Ann Hjalmarsson, and Anita Gidlöf-Gunnarsson. Office automation and users' need for support. *Behaviour & Information Technology*, 10(6):501–514, 1991.
8. Norman Walsh and Leonard Muellner. *DocBook: The Definitive Guide*. O'Reilly, 1999.
9. Stephan Weibelzahl. *Evaluation of Adaptive Systems*. PhD thesis, University of Freiburg, Germany, 2003.

Impacts of User Modeling on Personalization of Information Retrieval: An Evaluation with Human Intelligence Analysts

Eugene Santos Jr. Qunhua Zhao, Hien Nguyen, and Hua Wang

Computer Science and Engineering Department
University of Connecticut
191 Auditorium Road, U-155, Storrs, CT 06269-2155
{eugene,qzhao,hien,wanghua}@engr.uconn.edu

Abstract. User modeling is the key element in assisting intelligence analysts to meet the challenge of gathering relevant information from the massive amounts of available data. We have developed a dynamic user model to predict the analyst's intent and help the information retrieval application better serve the analyst's information needs. In order to justify the effectiveness of our user modeling approach, we have conducted a user evaluation study with actual end user, three working intelligence analysts, and compared our user model enhanced information retrieval system with a commercial off-the-shelf system, the Verity Query Language. We describe our experimental setup and the specific metrics essential to evaluate user modeling for information retrieval. The results show that our user modeling approach tracked individual's interests, adapted to their individual searching strategies, and helped retrieve more relevant documents than the Verity Query Language system.

1 Introduction

It is both critical and challenging for analysts to retrieve the right information quickly from the massive amounts of data. The task of designing a successful information retrieval (IR) system for intelligence analysts is especially difficult, considering that even when given the same search task, each analyst has different interests and almost always demonstrates a cognitive searching style that is different from others analysts. Clearly, a user model of a intelligence analyst is essential to assisting the analyst in his/her IR task. Since the early 80s, user modeling has been employed to help improve users' IR performance [3]. In our recent efforts, we developed a dynamic user model that captures an analyst's intent in order to better serve his/her information needs [14, 15] in an IR application.

In order to properly assess the effectiveness of a user model, we need to measure how an analyst's performance and experience with an IR system are affected. Intelligence analysts are personnel for collecting and compiling information for government, law enforcement and defense, etc. They are trained be self-conscious of their reasoning process [12], which includes the IR process. One major barrier for

evaluating a system designed for analysts is the limited accessibility to working intelligence analysts, and the nature of the information used in the evaluation.

To assess the effectiveness of our user modeling approach, we have conducted an evaluation with three working intelligence analysts. The objectives of this evaluation are: 1) to evaluate how our user model enhanced IR system performs when compared against a traditional IR system implemented with a keyword based query language, the Verity Query Language (VQL) [16]; 2) to study the impacts of our user model on augmenting personalization in IR; and, 3) to get feedback from the evaluators on user performance. The results show that our user modeling approach tracked the analyst's intents and adapted to the individual analyst's searching styles which helped them retrieve more relevant documents, especially those relevant to each analyst than the system implemented with VQL.

This paper is organized as follows: We first briefly present related work on user modeling in IR and its evaluation. Next, our user modeling approach is described, followed by our prior work on IR evaluation. Our evaluation methodology is then presented and our results are reported. Finally, we present our conclusions and future work.

2 Background and Related Work

The main objective of IR is the retrieval of relevant information for users. It is not an easy task, not only because of the explosive amounts of available information (especially unstructured information), but also the difficulty in judging the relevance, which can be objective or subjective in nature (reviewed by Borlund [2]). User modeling techniques attracted much attention in efforts at building a system for personalizing IR [3, 14]. However, proper evaluation of the user model for IR remains a challenge [4, 10, 17].

In the IR community, various methodologies, procedures, and data collections for evaluation of IR performance have been developed. In a typical experiment with data collections like Cranfield [5], a set of relevant documents is picked up by human assessors for a certain query (topic). Their judgments are considered to be objective [2]. The sets of relevant documents are then used for calculation of precision and recall. The criticism is that these experiments ignore many situational and mental variables that affect the judgment on relevance [8].

Besides applying metrics developed in the IR community, such as precision and recall, for measuring the effectiveness of the user model for IR [4, 7], efforts have also been made to study the impacts of the different systems on user behaviors. The emphasis was on the interaction between the user and the system. In a study done by Koenemann et al [7], the influence of four interfaces, which offered different levels of interaction in relevance feedback supported query formulation, to the user searching behaviors are studied. They found that different interfaces shaped how the users constructed their final queries over the course of the interaction. When the users could view suggestions and had control on the final actions, they needed less iterations to form good queries.

Recently, researchers in the user modeling community have focused on the development of general frameworks to conduct usability tests, which involves various forms of aptitude tests, cognitive tests and personality tests through surveys and questionnaires [4]. In IR domain, these results should be carefully considered, since previous research showed that user preference is not correlated with human performance [17]. Therefore, reliable conclusions could not be obtained solely based on either performance or user satisfaction. As Chin [4] pointed out, the difficulty lies with the evaluation approaches and study with real users to justify the overall effectiveness of a user model.

We attempted to evaluate our user modeling approach for IR, which is described in the next section, by measuring improvement in system performance, system adaptation to the user, and the user's experience with the system.

3 IPC User Model

Our user modeling module consists of three components: Interests, Preferences and Context, which is referred as the IPC model [14, 15]. Interests capture the focus and direction of the individual's attention; Preferences capture how the queries are modified and if the user is satisfied with the results; and Context provides insight into the user's knowledge. We capture user Interests, Preferences and Context in an Interest set, a Preference network and a Context network, accordingly. Interest set is a list of concepts, each of them associated with an interest level. Initially determined based on the current query, it is then updated based on the intersections of the retrieved relevant documents. The Preference network is captured in a Bayesian network [11], which consists of three kinds of nodes: pre-condition nodes, goal nodes and action nodes. Pre-condition nodes represent the environment in which the user is pursuing the goal. Goal nodes represent the tools that are used to modify a user's query; and action nodes represent how the user query should be modified. The Context network is a directed acylic graph that contains concept nodes and relation nodes. It is created dynamically by finding the intersections of retrieved relevant documents. The user model captures the analyst's intent and uses this information to modify analyst's query proactively for the IR application, please see [14, 15] for details.

The user model module has been integrated into an IR system. In the IR system, a graph representation for each document (called a document graph) is generated automatically in an offline process. The document graph is a directed acyclic graph consisting of concepts and the relations between concepts [14]. The query is also transformed into a query graph, which is then matched against each document graph in the collection. To speed up the matching process, only 500 documents that contain at least one term that exists in the query will move into the graph matching process. If there are more than 500 such documents, then the documents containing less terms from the query will be removed. The similarity measure between document graph and query graph is modified from Montes-y-Gòmez et al [9], also see [14].

4 Evaluation Methodology

Previously, we evaluated our user modeling approach by using evaluation measures, procedures and data collections that have been established in the IR community [10]. These experiments demonstrated that our user modeling approach did help improve the retrieval performance. It offers competitive performance compared against the best traditional IR approach, Ide dec hi [13], and offers the advantage of retrieving more quality documents quickly and earlier [10].

As such, we would like to compare our user model enhanced IR system to a more traditional system implemented with a keyword based query language. Furthermore, we would like to have an opportunity to study the impacts of our user modeling approach on augmenting personalization within the IR process, and get feedback from real intelligence analysts about their personal experience. A data collection from the Center for Nonproliferation Studies (CNS, Sept. 2003 distribution. http://cns.miis.edu/) has been chosen as the testbed for this evaluation. It contains 3,520 documents on topics of country profiles concerning weapon of mass destruction (WMD), arms control, and WMD terrorism. It was chosen because its content and its built-in commercial query system from Verity, Inc. [16] that can be used as a baseline system for comparison. In the following text, we will refer to our user model enhanced IR system as the UM system, and CNS with VQL as the VQL System.

The evaluation took place at a laboratory of the National Institute of Standards and Technology (NIST) in May, 2004. The UM system package, which includes the pre-processed CNS database, was delivered to and installed at the NIST laboratory. Three evaluators, who are naval reservists currently assigned to NIST with intelligence analysis background participated in the experiments. Since only three analysts were available, to obtain some fair comparison data, we have to run the UM system and the VQL system side by side during the evaluation. The same queries were input into both systems and the retrieved documents compared. For the VQL system, analysts needed to note on paper which documents were relevant to their interests for each query; for the UM system, in addition to recording the relevancy, they were asked to mark checkboxes beside the documents if they were relevant ones. There was a short tutorial session to show the analysts how to work with the UM system, such as indicating the relevancy. For the VQL system that has a graphic user interface (GUI) similar to Google, it is straightforward to use.

The experimental session lasted about 4 hours for each analyst due to analyst availability and laboratory scheduling. Participants were asked to carry out a searching task on “research and development in Qumar that supports biological warfare” (Note that some of the location names have been replaced). Because of this timing constraint, the participants were asked to check the first 10 returned documents for relevancy only, and the task was limited with just 10 fixed queries (Table 1). For any empirical study, one challenge lies in the large numbers of variables to control (including the human factors). By scripting the queries, we can avoid introducing more variables into our experiments, such as different queries, different number of query inputs, and error in natural language processing. It allowed us to have a better control on the experiment in such a short session, and focus on the main objectives of the evaluation, which is to study the impacts of user model on the IR, as described in the introduction. The queries were extracted and modified from a database that

collected other intelligence analysts' IR activities at the NIST laboratory, which allows us to construct a realistic evaluation session. The UM system started with an empty user model, which means that the user model initially knew nothing about the analyst, and had to start learning about the user from the very beginning.

Table 1. The 10 queries used in the evluation experiments

1	Qumar research biological warfare
2	Qumar research institute, university biological warfare
3	Qumar biological research and biological warfare
4	Biological research facilities in Qumar
5	Intelligence assessment on Qumar biological research
6	Qumar foreign connections in biological weapons program
7	Bacu, Qumar and Russia connections to WMD
8	Qumar's biologists visits Bacu
9	Russian biotechnology, missiles, aid to Qumar
10	China supply and Qumar biological weapons program

Besides the IR task, analysts were asked to fill out an entry questionnaire about their background and experience with searching programs; and, respond to an exit questionnaire about their experience on working with the UM system.

5 Results

The experience in intelligence analysis for the three participants ranged from 5 months to 7 years. Two of them use computers as a tool in their analysis job, while one does not (Table 2). They all felt comfortable with using search tools like Google, and considered themselves well-informed on the topics of WMD and terrorism. Analyst 3 stated that he has never used a system that requires feedback for annotating relevancy (Table 3). The most interesting observation is that the three analysts tend to take different approaches in IR. Analyst 2 looks at the big picture first; while analyst 3 likes to start with the details. Analyst 1 practices a mixed approach that depends on his knowledge of the topic. If much was already known, then he would try to create an outline of the useful information; otherwise, he would look for some details first (Table 3).

After 4 hours, two analysts finished 10 queries that we provided, and Analyst 3 finished 9 queries (Table 4). All of them managed to identify more relevant documents when working with the UM system than they did with the VQL system (Table 4). The precision were 0.257 and 0.312 for the VQL system and the UM system respectively. Since a document could be returned and identified multiple times as relevant for different queries, we also counted the numbers of unique (or distinct) documents that have been returned by the system and found as relevant by each participant. The data showed that when they were using the UM system, each of them was presented with more unique documents, and selected more unique documents as relevant (Table 4). The total number of unique relevant documents for all 10 queries

returned by the UM system is 39, while the number is 27 by the VQL system, a 44% increase (Table 5).

The number of documents selected as relevant by more than 2 analysts are 15 in the UM system and 19 in the VQL system, respectively. Notice that the number of documents marked as relevant by just one analyst is 24 when using the UM system, while this number is only 12 for the VQL system (Table 5). This suggests that more information that is specifically relevant to each analyst's individual interests had been retrieved by the UM system. By using the UM system, the analysts displayed their differences in identifying the documents that were relevant to their individualize interests and searching style.

Table 2. Demographic data

	1	2	3
Highest degree	JD	MS	BA
Length of time doing analysis	7 years	5 years	5 months
Computer expertise	novice	medium	medium
Use computer to do analysis	not at present	yes	yes
Experience doing queries	yes	yes	yes
Query expertise	novice	medium	medium

Table 3. Questions on information seeking behaviors of three participants.

	1	2	3
What is your overall experience with systems using ranked outputs and full-text databases, such as Google? (1-7) 1 is very experienced, 7 is no experience	3	1	1
Have you ever used a system that asked you to indicate whether a document or other system response was relevant? Yes, No	Y	Y	N
When faced with a search problem do you tend to: (a) Look at big picture first, (b) Look for details first, (c) Both	c	a	b
What is your knowledge of Terrorism (1-7) 1 very experienced, 7 no experience	2	3	2
What is your knowledge of WMD? (1-7) 1 very experienced, 7 no experience	3	2	2

By the end of the experiment, the analysts were asked to fill out the exit questionnaire. Generally, they agreed that the scenario used in the evaluation experiment was very realistic, and gave an above average score for feeling comfortable at preparing a report on their task after querying for information. When asked about the system performance and their satisfaction, they scored the UM system as above medium (3.7/5.0) (Table 6 and 7). Notice that they felt the UM system was somewhat demanding, especially in mental effort and the temporal effort. Since relevancy assessment is a mentally demanding process by itself, and the analysts were required to finish the experiment in about 4 hours, which included 10 queries (i.e., more than 100 documents to review, of which some of them may be quite long), and

working with 2 different systems at the same time, we think this is a result of the workload the analysts had in the experiments. As the data shows, the UM system presented more unique documents to the analysts, and helped analysts retrieve more relevant documents. In particular, it helped them retrieve more information that is relevant to their individual interests, which suggests that the user model was tracking the user's personalized interests.

Table 4. Number of documents presented to the analysts, and number of documents marked as relevant with each of the systems.

	VQL system			UM system		
Analyst	1	2	3	1	2	3
Documents presented	100	100	90	100	100	90
Relevant documents	11	31	33	16	41	36
Unique document presented	49	49	45	67	72	54
Unique relevant documents	9	19	21	10	29	23

Table 5. Unique relevant document retrieved by two systems.

	UM System	Verity System
Total unique relevant documents	39	27
Documents marked as relevant by all 3 analysts	8	3
Documents marked as relevant by more than 2 analysts	15	19
Documents marked as relevant by only 1 analyst	24	12

6 Discussion and Future Work

In this paper, we present our evaluation methodology and the results for our user modeling approach. Since the ultimate goal of IR is to meet the user's information needs, testing by actual end users (the analysts in this case) is an evaluation that can not be replaced by other methods. The involvement of end users can help us avoid problems with traditional IR evaluation metrics which excludes the user's individual information needs. Our evaluation answered the question on impacts of user modeling on the retrieval performance of an IR system by measuring the number of unique documents presented to the analysts and relevant ones have been identified; and studied the impacts of user modeling on the personalization of IR by tracking the difference between the documents retrieved by different analysts. By combining these results, we can judge if the user modeling is actually follows the user's individual interests, and improve the IR performance.

Intelligence analysts are trained experts specialized in IR and information analysis in certain areas. It is very hard to get time from real analysts to test a system in a experimental setting. We are very glad that we have had the chance to perform such an evaluation. Since the experimental time is limited (4 hours), we used a short scripted query sequence to reduce the number of variables in the experiment, which allows us to focus on our main objectives.

Table 6. Average score for performance of the UM system (1)

Question	Score
How realistic was the scenario? 1-5, 1 is not realistic, 5 is realistic	4.7
Did it resemble tasks you could imagine performing at work? 1-5, 1 not realistic, 5 realistic	3.7
How did the scenario compare in difficulty to tasks that you normally perform at work? 1-5, less difficult, 5 more difficult	2.7
How confident were you of your ability to use the system to accomplish the assigned task? 1-5, 1 less confident, 5 more confident	3.0
Given that you were performing this task outside of your standard work environment, without many of your standard resources, were you comfortable with the process of preparing your report? 1-5, 1 less comfortable, 5 more comfortable	3.7
Given that you were performing this task outside of your standard work environment, with access to a restricted set of documents, were you satisfied with the quality of the report/answers that you were able to find for this scenario? 1-5, 1 not satisfied, 5 satisfied	2.7

Table 7. Average scores for performance of the UM system (2)

Question	Score
How satisfied are you with the overall results for this task using OmniSeer? 1-7, 1 most satisfied, 7 least satisfied	4.3
How confident are you with the results that they cover all possible aspects of the task? 1-7, 1 most confident, 7 least confidence	4.7
Regarding this task, do you think the OmniSeer approach helped you to retrieve critical documents earlier in the process than the Verity system? 1-7, 1 strongly agree, 7 strongly disagree	3.7
Please rank the following factor: mental demand 1-7, 1 little 7 high	5.3
Please rank the following factor: physical demand 1-7, 1 little 7 high	2.0
Please rank the following factor: temporal demand 1-7, 1 little 7 high	5.0
Please rank the following factor: performance demand 1-7, 1 little 7 high	4.7
Please rank the following factor: frustration 1-7, 1 little 7 high	5.3
Please rank the following factor: effort 1-7, 1 little 7 high	6.0

Although there were only 3 analysts tested on the system within a limited period of time, the results are encouraging. First, the UM system provided more information to the analysts (returned more unique documents, which is usually can only achieved by asking more queries), and helped them to identify more relevant information. Second, even more importantly, experimental results suggest that the UM system tracked the individual interests of the different analysts, and returned different sets of documents to them individually. We know that the 3 analysts employ different seeking approaches (look for general information first, or look for details first, or use a mixed approach). With the UM system, the query was modified based on the user's

feedback. During the IR process, different analysts considered different documents as relevant based on their own knowledge, experience and goals, which led to the difference in the modification of the queries by the user modeling module. As a result, they were presented with different documents. This demonstrates the impacts of our user model on augmenting personalization in the IR process. With the VQL system, there is no effort to meet the individualized information needs. It is always the same set of documents returned for the same query. Because of the timing constraints, the evaluation only involved one task consisting of 10 queries. Also, the query sequence was fixed. If there were more queries asked freely by the participants, and with a larger database, the UM system would have been able to indicate even more significant differences among the analysts.

VQL is a very successful commercial query language. It has been developed and enhanced over more than a decade. Many advanced functions have been included in VQL, such like proximity, density, frequency, field, concept, word stemming, and word location [16]. It is obvious that our UM system, as a prototype, lacks many advanced features offered by the VQL system. For example, VQL's word location function helps the user find the keywords in the query (or words closely related) by highlighting them in the documents; the UM system does not provide the same convenience although we tried to implement a similar GUI with the intention of minimizing the interface differences. VQL uses keyword or concept indexing to accelerate searching process, which also has a big advantage over the current version of our UM system. These features could affect the evaluation outcomes, and might make the participants feel that the VQL system takes less effort.

In a study by Alpert et al [1], it has been pointed out that users want to feel that they are in control. In our case, the analysts were given a short training session and brief introduction on how to use the UM system before the experiment, and were informed that their feedback will be used by the system to try and improve performance. Unfortunately, it is far less than what is necessary. More work is needed in the future to help users understand how and why the system evolves and behaves, which will grant them more of a sense of being in command, and help users overcome suspicious attitudes, such as a system's ability to do it well enough to be useful.

Currently, in the UM system, relevance is explicitly selected by the analysts at the whole document level. When selected, the whole document is placed into the relevant set. This may introduce noise into the user model, since it is possible that only part of the document is considered relevant by the user. In the future, a system may be implemented with both explicit and implicit feedback mechanisms. Implicit feedback, like Hijikata's work [6], can both lessen the burden of marking the relevancy by the user and also identify the specific part that is of interest within the presented text. Explicit feedback can let the user be in control and indicate to the system what the most important relevant information is. We hope, with more concise feedback, our user model can better infer the user's intent and then assist their information needs.

Acknowledgments : This work was supported in part by the Advanced Research and Development Activity (ARDA) U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government. Also, we express our special thanks to Dr. Jean Scholtz and Emile Morse, who helped organize the

evaluation, collected data, and provide the preliminary data analyses. Without their help, this evaluation experiment would not have been successful. This work is a part of the Omni-Seer project, which involves Global InfoTek, Inc., University of Connecticut, University of South Carolina, and KRM, Inc. [15].

References

1. Alpert, S.R., Karat, J., Karat, C-M., Brodie, C., Vergo, J.G. : User Attitudes Regarding a User-Adaptive eCommerce Web Site. User Modeling and User-Adapted Interaction 13 (2003) 373-396
2. Borlund, P. : The Concept of Relevance in IR. Journal of the American Society for Information Science and Technology 54(10), (2003) 913-925
3. Brajnik, G., Guida, G., Tasso, C. : User Modeling in Intelligent Information Retrieval. Information Processing and Management 23(4) (1987) 305-320
4. Chin, D. : Empirical Evaluation of User Models and User-Adapted Systems. User Modeling User-Adapted Systems 11(1-2), (2001) 181-194.
5. Cleverdon C. : The Cranfield test of index language devices. (1967) Reprinted in Reading in Information Retrieval Eds. 1998. Pages 47-59.
6. Hijikata Y. : Implicit User Profiling for On Demand Relevance Feedback. 2004 International Conference on Intelligent User Interfaces (IUI04). ACM presses, Funchal, Madeira, Portugal. (2004) 198-205
7. Koenemann, J., Belkin, N.: A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In Proceedings of CHI 96 (1996) 206-212
8. Large, A., Tedd, L.A. Hartley, R.J. : Information Seeking in the Online Age: Principles and Practice. (1999) London: Bowker Saur
9. Montes-y-Gòmez, M., Gelbukh, A., Lpez-Lpez, A. : Comparison of Conceptual Graphs. In Proceedings of MICAI-2000, 1st Mexican International Conference on Artificial Intelligence. (2000) Acapulco, Mexico.
10. Nguyen, H., Santos, E. Jr., Zhao, Q., Lee, C. : Evaluation of Effects on Retrieval Performance for an Adaptive User Model. AH2004: Workshop Proceedings - Part I. Third Workshop on Empirical Evaluation of Adaptive Systems. (2004) 193-202
11. Pearl, J. : Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. (1988) Morgan Kaufmann, San Mateo, CA
12. Heuer, R. : Psychology of Intelligence Analysis. (1999) Government Printing Office.
13. Salton, G., Buckley, C. : Improving Retrieval Performance by Relevance Feedback. Journal of the American Society for Information Science. 41(4), (1990) 288-297
14. Santos, E., Jr., Nguyen, H., Brown, S.M. : Kavanah: An Active User Interface information Retrieval Agent Technology. In Proceeding of the 2nd Asia-Pacific Conference on Intelligent Agent Technology. (2001) 412-423.
15. Santos, E., Jr., Nguyen, H., Zhao, Q., Wang, H. : User Modelling for Intent Prediction in Information Analysis. In Proceedings of the 47th Annual Meeting for the Human Factors and Ergonomics Society, Denver, Colorado, (2003) 1034-1038
16. Verity White Paper: The Verity K2 Discovery Tier, The Importance of Advanced, Effective Search Tools. (2004)
http://www.verity.com/pdf/white_ papers/MK0348c_WP_Discovery.pdf
17. Wilkinson, R., Wu, M. : Evaluation Experiments and Experience from Perspective of Interactive Information Retrieval. In Working notes of Empirical Evaluation of Adaptive Systems workshop at Adaptive Hypermedia Conference. (2004) 221-230.

Evaluating Scrutable Adaptive Hypertext

Marek Czarkowski

Department of Computer Science,
University of Sydney, Australia
marek@cs.usyd.edu.au

Abstract. Adaptive hypertext systems personalise documents to meet the individual's particular preferences, knowledge and goals. There is a debate over how much control should be given to the user as well as how much transparency there should be to the inner workings of the system. Some adaptive systems make the user model available to the user. We propose transparency and control should extend beyond this by involving the user in the personalisation process and granting them power to change it. Our previous evaluations of scrutable systems have revealed users have difficulty understanding and controlling personalisation. We have developed SASY with a focus on improving scrutability support tools. This paper describes our design for the evaluation of SASY.

1 Introduction

Adaptive hypertext systems personalise documents to meet the individual's particular preferences, knowledge and goals. This offers benefits including an improved user experience, efficiency in information retrieval and navigation support [3].

There is a growing debate within the field of adaptive systems as to how much control should be given to the user and how much transparency there should be to the inner workings of the system [7, 10]. There is an argument that users must understand, to some degree, the workings of an adaptive system in order to trust it to perform tasks on their behalf [10]. In addition, users must feel as though they have ultimate control over the system if and when they choose to exercise it. There is also literature that shows that systems that expose the user model to the student promote learner reflection and enhance learning [2, 11, 12]. In contrast, an evaluation of a tool to recommend relevant conference events [8] found users do not always exercise control.

We propose providing transparency and control should extend beyond making the user model available to the user by involving the user in the personalisation process and granting them power to change it. Our motivation for increasing transparency and control over personalisation is based on several key drivers:

- Increasing legal requirements for access to personal data [9].
- In non-critical applications, such as movie recommendation systems, people trust machines to personalise information. The user is not too concerned if the personal-

isation is faulty, for example, recommends movies the user has no interest in. However, for more critical tasks the user would want more transparency and traceability, for example, a system that invests a user's money,

- Empowering users to correct the misconceptions a system holds about them which impacts personalisation.
- Support a user's sense of curiosity and exploration of the personalisation.
- Allow users to develop comfort in understanding how the system works [10].

We have developed SASY, a personalised system that allows users to scrutinise, or inspect, the personalisation to understand why it occurred and how to control it. The design was informed by the development and evaluation outcomes of the Tutor series of systems which achieved some success in supporting scrutability [5,6]. From the evaluation of Tutor3, we concluded that although users seem comfortable with the notion of personalisation driven by a simple user model, being able to control the personalisation is foreign. Furthermore, despite the compelling reasons for scrutinising personalisation, we have found users are not typically willing to do so. A key concern was that some users had difficulty finding the scrutability tools when needed.

We had two main goals in mind when developing SASY. Firstly, we wanted to test scrutability support in a more genuinely adaptive environment. In Tutor3 the user model was entirely driven by answers to profile questions and was otherwise static. SASY is adaptive rather than adaptable as the user model is updated behind the scenes based on observations about the user. The user model contains attributes that are inferred by the system and these can also be changed by the user. Secondly, the interface design of SASY addressed the issue that some users of Tutor3 failed to find and activate the scrutability support tools.

This paper introduces SASY and describes our evaluation that is in progress.

2 Overview of SASY

SASY provides a generic framework for the delivery of personalised content over the World Wide Web. SASY is a web application that presents personalised content to end-users through a standard web browser. Authors create content by publishing a set of XML documents that conform to a schema called Adaptive Tutorial Mark-up Language (ATML). ATML (and hence SASY) is domain independent, but has additional features which simplify the creation teaching material. An ATML document is an HTML 4.0 document with additional tags that allow the author to define personalisation rules for adaptive content.

SASY builds a profile of each user's interests, background, goals, etc. from answers the user provides to a brief questionnaire. The profile is also populated with beliefs that are inferred about the user as they use the system. For example, accessing a certain page may cause SASY to update the user profile to capture interest in the subject matter. When a user requests to view a page, SASY evaluates personalisation rules embedded in the ATML page against their user profile. Thus, SASY determines

what adaptive content should be included in the user's view of the page and what should be omitted. This is a standard Adaptive Hypertext technique [3], similar to the approach employed by AHA! [4].

What makes SASY unique is the built-in tools for scrutinisation that allow users to see how content is personalised and change the personalisation should they choose to do so. Each page includes a summary of the personalisation that was performed. The user can also access a detailed explanation of the personalisation. In this mode, SASY highlights content that was included or removed from the user's personalised view of a page. It also indicates the user profile attributes and their values that caused content to be included or removed. The user can not only view and update their user profile, but also view an explanation of how each profile attribute was set. This is particularly useful for user profile attributes that have been inferred by SASY through observations about the user.

As part of our evaluation of SASY, we have developed scrutable, personalised content for a number of different domains: UNIX Security course, Holiday Planner, Television Guide and a Museum Guide. We describe the first three in the evaluation design.

3 The User's view of SASY

To use SASY the user logs in and selects from a list of available topics. Each topic is effectively a separate application, with its own set of content. On selecting a topic, SASY shows the profile page. The first time it is accessed it displays a questionnaire that is used to build a profile of the user's interests, background, goals, etc. For example, the Holiday Planner profile asks potential vacationers whether they seek Adventure or Relaxation. Otherwise, it shows the questionnaire and any other profile attributes that SASY has inferred about the user. The user can click a hyperlink to pop-up an explanation of how and why each profile attribute was set. For inferred attributes, this tells the user what event caused SASY to create this belief about them.

Following the profile page, the user is taken to the Home page. Figure 1 shows the Home page from the personalised Holiday Planner. This is a typical personalised page where adaptive content has been included and/or removed, depending on whether the condition to show the adaptive content matched the user's profile.

Every page contains a panel on the right hand side titled "Personalisation", from which the user invokes tools to scrutinise the personalisation. It includes labels to indicate the number of adaptive content items have been removed and included to create the user's personalised view of the page. For example, the page in figure 1 has had five items removed and one item included. Clicking the hyperlink "Click to highlight removed/included items on this page" reloads the same page but this time shows the raw view of the page without any personalisation applied (Figure 2). Content that was removed through personalisation is highlighted in yellow, included content is highlighted in grey. For example, in Figure 2, the content starting with the title "Wine Country Touring" is highlighted in yellow because it was removed. The page instructs the user to hold their mouse over a highlighted item to see an explanation of why the item was included or removed. In Figure 2, the user had moved their mouse

over the “Wine Country Touring” section to pop-up the text “Not shown to you because your profile says: You’re seeking adventure”. That is, the content was removed by the personalisation because the user’s profile states they are seeking adventure.

Below the “Click to highlight removed/included items on this page” link is a list of user profile attributes that affected personalisation on the page. In Figure 1, the personalisation considered the user is single, seeking adventure and have a low budget. Clicking on a user profile attribute, e.g. the hyperlink “You are single”, pops up a window that explains why SASY holds this belief about the user.

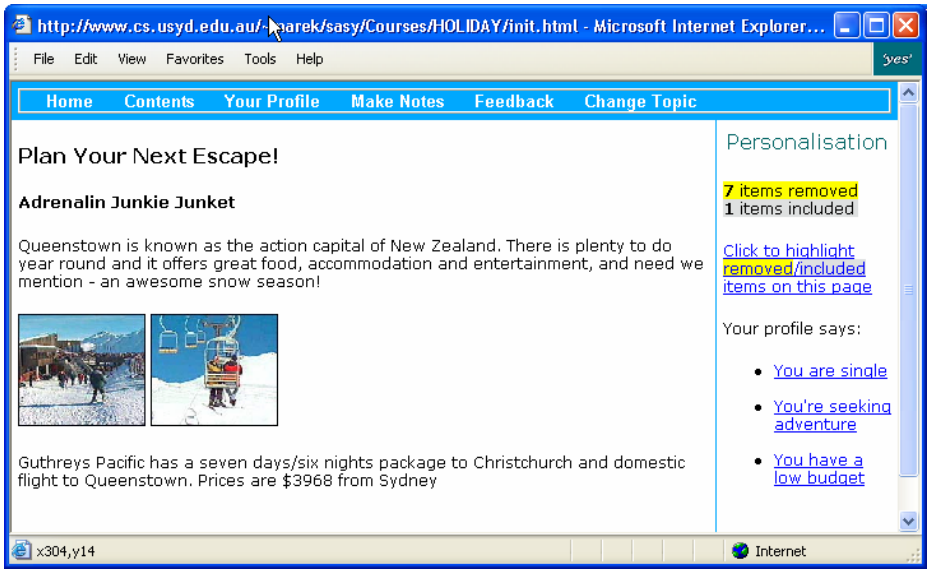


Fig. 1. A typical personalised page in SASY. SASY removed and included adaptive content items to create view of the page.

4. Evaluation Design

We wish to evaluate SASY to determine whether users are able to:

- Understand personalisation is driven by their user profile that may be updated by SASY through their interaction with the system.
- Scrutinise the personalisation on a page and understand why content was included or removed to create their personalised view.
- Demonstrate control over the personalisation by changing values in their user profile to change the personalisation.

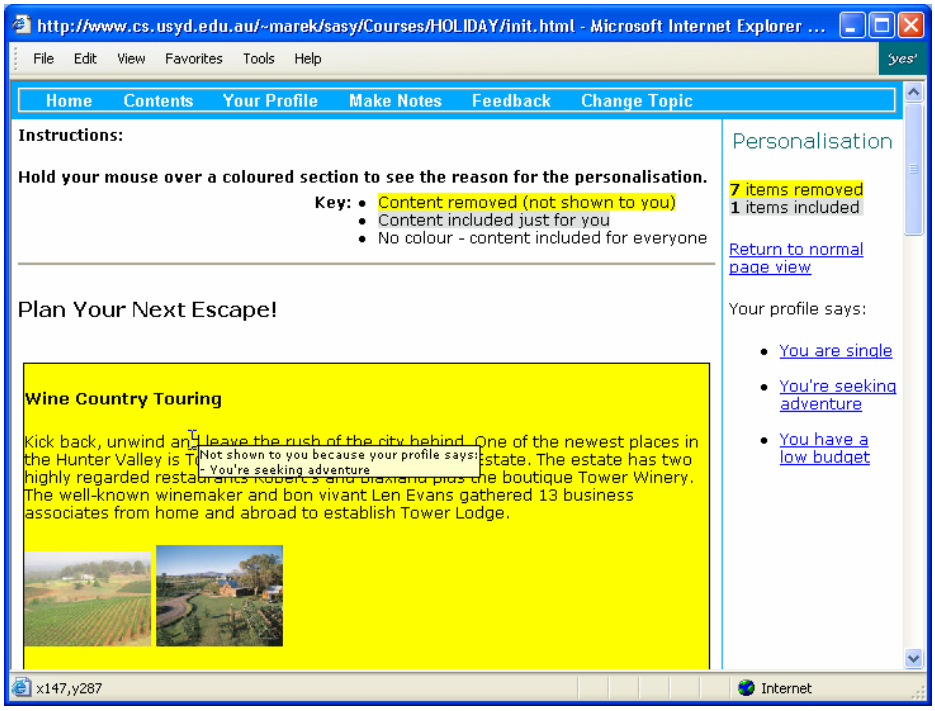


Fig. 2. Typical personalised page in SASY in highlight mode, showing all the content available on the page and explaining why content was included or removed to create the personalised view (Figure 1). Content that was removed is highlighted in yellow, included content is highlighted in grey.

The difficultly in measuring how effectively users scrutinise and control personalisation is that we know from evaluations of the Tutor systems [5, 6] users will not scrutinise often. This is understandable since scrutinisation is not the user's main purpose for using the system. In Tutor evaluations, participants noted the default personalisation seemed appropriate hence they were not motivated to scrutinise.

To counter this difficultly, we base our evaluation of SASY around the most common scenarios where the user might be motivated to scrutinise:

– A user believes the personalisation to be faulty because it produces unexpected results.

– A content author wishes to debug the adaptive content they have created.

– A user is curious as to what the system believes about them or how a page was personalised and wants to explore alternatives.

The evaluation comprises of three separate experiments across different subject matter domains to reduce the effect of a particular domain on the results. Since the personalisation engine is the same in each experiment, results can be directly compared. Each evaluation captures both quantitative results (time to complete task, task

correctness) and qualitative feedback to gauge user satisfaction and capture any concerns raised by users.

5. Evaluation 1 – Personalised TV Guide

This evaluation is inspired by the article "My Tivo thinks I'm Gay" [1]. It describes a situation where Tivo, an adaptive movie recommendation system, consistently recommends movies with a homosexual theme, much to the dismay of its owner who is not interested in this genre. The owner has no way to directly correct the beliefs Tivo holds about them.

This laboratory based experiment models this scenario with a Personalised TV Guide system, where the user has a need to understand the personalisation process and be able to exercise control over it. Unlike Tivo, SASY supports the user in doing this.

5.2 Aim

To measure how effectively SASY supports the user to:

- Scrutinise a page to determine why adaptive content is included/removed in relation to their user profile.
- Explain how/why a belief held by the system was instantiated. In this case the belief is inferred by the system through the user's interaction with the system.
- Demonstrate control over the personalisation by altering their profile to change how content is included and removed.

Additionally we wish to evaluate:

- The effect of the displaying user profile attributes that affected personalisation on each page. We suspect it discourages the use of the content highlighting feature as reading the profile attributes is a quick form of scrutinisation.
- The effect of user training in the form of reading through an online overview.

5.3 Participants

Fifteen randomly selected participants, a small number based on Nielson [13]. It is assumed that all participants have basic familiarity with web browser based applications and interfaces. Users are informed their identity is anonymous throughout the evaluation.

5.4 Method

Participants randomly divided into groups which have different system variants:

- Group 1 – by default the system will display user profile attributes that affected personalisation in the right hand column of the page.
- Group 2 – by default the system will not display user profile attributes.
- Group 3 – same as Group 1 but are given a system introduction to read as training.

All groups have access to online help documentation. During the evaluation user actions are logged to allow us to measure task performance and accuracy. Participants perform tasks described in a worksheet that is presented in an online application separate to SASY. The online worksheet presents one task at a time and logs the time each task is started and completed. The worksheet asks participants to:

- Access the TV Guide and complete a questionnaire to determine their interest in television program genres: Sports, Business & Finance and Current Affairs. SASY builds a user profile and displays a recommended TV viewing schedule.
- Inspect and change the personalisation such that their viewing schedule only includes current affairs programs.
- Use the system to read about any programs they are interested in. The system is crafted so that regardless of the user's selection, SASY will set an attribute in their profile stating they are a member of "Special Interest Group 1", which will cause religious programs to be included in the viewing schedule. This is intentionally obscure so that profile attributes displayed on the page will not directly indicate an interest in religious programs.
- Review the viewing schedule to explain why it now includes religious programs and change the personalisation to return the schedule to its previous state. Participants will need to use the system feature that highlights personalised content and shows the user profile attributes that caused content to be included or removed.
- Provide qualitative feedback to evaluate user satisfaction and usability. Participants will select answers based on a Likert scale.

5.5 Expected Results

We expect participants will be able to scrutinise the TV viewing schedule and change their user profile to control how content is included/removed by the personalisation.

We expect users will rely heavily on the user profile attributes shown in the personalisation panel but will be also able to use the highlight feature to complete tasks.

6. Evaluation 2 – Personalised Holiday Planner

In real world applications of adaptive systems, the content author is required to ascertain whether the content is correct and ready for publication. This evaluation asks the participant to assume the role of an editor, and validate that the personalisation in a Holiday Planning system is correct. This task is conceptually different from those in Evaluation 1 in that the user must be able to answer 'how was this page personalised to me' as well as 'how would this page be personalised to someone else'.

6.2 Aim

Determine whether participants can understand complex personalisation. In Evaluation 1, content is either removed or included based on a single user profile attribute. Here we use complex personalisation rules involving two or more attributes.

6.3 Participants

Ten randomly selected participants. It is assumed that all participants have basic familiarity with web browser based applications and interfaces.

6.4 Method

Participants randomly divided into two groups:

- Group 1 – not provided with an introductory system guide.
- Group 2 – given a system introduction to read as training.

The Holiday Planner presents a personalised holiday recommendation based on the user's vacation style preference (adventure, relaxation, family oriented), budget (low, high) and status (single, couple, family). Participants are asked to validate the personalisation is correct in the Holiday Planner. Unknown to participants, the system has the following errors:

- Simple error: a user with a status of 'single' is shown children's holidays.
- Complex error: a user with a status of 'couple' and a 'low' budget is shown expensive holidays.

All participants have access to online help and user actions are logged to determine task completion time and correctness.

6.5 Expected Results

We expect that users will quickly find the simple errors but may struggle with the complex ones.

7. Evaluation 3 - Personalised UNIX Security Course

Adaptive systems are commonly used as a means of providing personalised instruction. In this class of systems, the user is focused on learning and may not be interested in how the personalisation works unless they perceive value from scrutinising. For example, to see what content they have missed out on or to correct a misconception the system holds about them. In this evaluation, we allow the users to freely use the system to learn UNIX security concepts. However, we have planted adaptive content in the teaching material that would annoy most learners and we wish to evaluate whether we can provoke users to scrutinise content and change default personalisation to remove the annoying content.

7.2 Aim

- Determine whether learners can be provoked to scrutinise content to remove annoying material that is distracting to their learning focus.
- Is there a relationship between scrutinisation and the students learning outcomes?

7.3 Participants

Computer Science Students from the University of Sydney, Australia. Part of the syllabus of several subjects requires students learn about UNIX Security. Over one hundred students are expected to use the UNIX Security course.

7.4 Method

The students are all asked to read a system introductory guide that explains the scrutinisation feature of the system.

Students all sit a pre-test to access their pre-knowledge of UNIX security concepts. Students are then allowed free use of the system and will sit a post test at the end of their session to capture knowledge gains and provide qualitative feedback regarding their use of the scrutinisation tools. The course content includes annoying content in the form of jokes about the UNIX operating system.

7.5 Expected Results

We expect that users will scrutinise the system to remove the annoying content. This will test the notion that users always trust the default personalisation.

Users may also be curious about other aspects of the personalisation. We expect some users will further scrutinise the system out of curiosity or to reflect on their learning.

8. Conclusion

We have found that providing an effective user interface to support the process of scrutinising personalisation is very challenging. Although users seem comfortable with the notion of personalisation, being able to control the personalisation is new and not expected. However, based on user feedback from previous evaluations, users would like to be able to scrutinise an adaptive system. Building effective tools for this is challenging as we must support casual users who will not scrutinise often.

We have described the design for our evaluation of SASY that is currently in progress.

References

1. “Oh no! My TiVo thinks I'm gay”, Jeffrey Zaslow. The Wall Street Journal, Dec. 04, 2002.
2. Bull, S ., Brna, P. and Dimitrova, V. (Eds.) (2003). Proceedings of Learner Modelling for Reflection Workshop, Volume V of the AIED2003 Supplementary Proceedings, 193-298.
3. Brusilovsky, P. (2001) Adaptive hypermedia. User Modeling and User Adapted Interaction, Ten Year Anniversary Issue (Alfred Kobsa, ed.) 11 (1/2), 87-110
4. Calvi L., Cristea, A. (2002). Towards Generic Adaptive Systems: Analysis of a Case Study. De Bra, P., Brusilovsky, P., Conejo R. (Eds.), Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Malaga, Spain, May 2002, Springer:LNCS 2347, 79-89.
5. Czarkowski, M., Kay, J. (2000). Bringing scrutability to adaptive hypertext teaching. ITS'2000, Intelligent Tutoring Systems, Gauthier, G., Frasson, C., VanLehn, K. (Eds.), Intelligent Tutoring Systems, Springer, 423-432.
6. Czarkowski, M., Kay, J. (2003). How to give the user a sense of control over the personalization of adaptive hypertext? Workshop on Adaptive Hypermedia and Adaptive Web-Based Systems, User Modeling 2003 Session, 121-132.
7. Höök, K., Karlgren, J., Waern, A., Dahlbäck, N., Jansson, C.J., Karlgren, K., and Lemaire, B. (1996). A Glass Box Approach to Adaptive Hypermedia. Journal of User Modeling and User Adapted Interaction, special issue on Adaptive Hypermedia, 6(2-3), 157-184.
8. Jameson, A., Schwarzkopf, E. (2003). Pros and Cons of Controllability: An Empirical Study, De Bra, P, P Brusilovsky, R Conejo (Eds.), Proceedings of AH'2002, Adaptive Hypermedia and Adaptive Web-Based Systems, Springer, 193-202.
9. Kobsa, A. (2002). Personalized hypermedia and international privacy. Communications of the ACM archive, V 45 , Issue 5 (May 2002), Special Issue: The adaptive web, 64–67.
10. Maes, P., Schneiderman, B. (1997). Direct Manipulation vs. Interface Agents: A Debate. Interactions, Volume IV Number 6, November 1997, ACM Press, 42-61.
11. Mitrovic, A. & Martin, B. (2002). Evaluating the Effects of Open Student Models on Learning, in P. DeBra, P. Brusilovsky & R. Conejo (eds), Proceedings of Adaptive Hypermedia and Adaptive Web-Based Systems, Springer Verlag, Berlin Heidelberg, 296-305
12. Kay, J, (2001). Learner control. User Modeling and User-Adapted Interaction, Tenth Anniversary Special Issue, 11(1-2), Kluwer, 111-127.
13. Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. International Journal of Human-Computer Studies, 41(1-6), 385-397.

Layered Evaluation of Topic-Based Adaptation to Student Knowledge

Sergey Sosnovsky, Peter Brusilovsky

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

Abstract. A user modeling server is an important part of modern distributed E-Learning architectures. The user modeling server CUMULATE has two main levels: the event storage and multiple inference agents. To evaluate adaptive systems functioning as components of the common distributed architecture and using CUMULATE as the central user modeling server we need to evaluate the adaptation provided by those agents. Unfortunately, there are no commonly accepted approaches to the evaluation of the universal user modeling server. This paper describes the results of layered evaluation of our recent topic-based adaptation engine based on the activity students performed using the system QuizGuide. User modeling and adaptation processes are evaluated separately. While previous evaluation experiments of QuizGuide based on the traditional "with-and-without" approach showed that students like the system and benefit from it, this paper provides evidence of unfitness of large topics as knowledge assessment units used for adaptation, which challenges the reasonableness of the entire adaptation performed by the system.

1 Introduction

A number of researchers in the field of adaptive E-Learning are currently working on distributed component-based architectures for adaptive E-Learning [1], [2], [3], [4]. Such a distributed architecture includes user-adaptive components that could work in parallel with the same user while exchanging collected information about the user for better adaptation. One approach to handling the user modeling needs in a distributed architecture is a centralize user modeling server. Due to diverse needs of various components of a distributed architecture, a user modeling server should be relatively universal and flexible. To explore the problem associated with user modeling servers, we developed student modeling server CUMULATE (Centralized User Modeling for User and Learner-AdapTive Environments). CUMULATE is a user modeling component of the KnowledgeTree [1], a distributed architecture for adaptive E-Learning based on reusable intelligent learning activities. CUMULATE represents information about a student on two levels: the event storage and the user model distilled from event storage by multiple *inference agents*. An ability to define different inference agents is an important flexibility feature of CUMULATE that allows it to accommodate different user modeling needs. This paper focuses on evaluation of user

modeling servers with multiple inference agents. In the two following sections we introduce the specific inferenced agent that performs topic-based modeling of student knowledge and QuizGuide service that uses topic-based modeling. The remaining part of the paper discusses the issue of evaluation of CUMULATE-like servers and presents our attempt to evaluate the performance of CUMULATE in the context of QuizGuide service.

2 Topic-Based Knowledge Modeling

Topic-based knowledge modeling is our most recent attempt to develop an adaptation approach that could be understood, authored, and used by practical teachers. It is a further simplification of *concept-based knowledge modeling* that we explored in the past in InterBook [5] and that we found too complicated for an average teacher. Similarly to the concept-based approach, the student knowledge is represented as a weighted overlay over a set of knowledge elements. However, in topic-based modeling these are coarse-grain elements called topics. We assume that a typical course-level domain model includes just several dozens of topics (in contrast to several hundred concepts). The most important difference is that in the topic-based approach each educational activity contributes to only one topic, while in the concept-based approach it can contribute to multiple concepts (known as outcomes). Our implementation of topic-based modeling follows a transparent approach that was advocated by some instructional designers [6]: for each topic a course author or a teacher identifies several educational activities. Student progress with these activities defines the user understanding of a topic.

3 QuizGuide: An Adaptive Topic-Based Hypermedia Service

We have explored topic-based knowledge modeling in QuizGuide – a value-added service that provides personalized access to self-assessment quizzes for the C programming language. The student interface of QuizGuide consists of two main parts: the quiz navigation area and the quiz presentation area. The quiz navigation area (left on Fig. 1) provides hyperlinks to 44 quizzes organized in 22 topics. Each topic link is adaptively annotated with a target-arrow icon that expresses both the relevance of the topic to the current educational goal and the student's current knowledge. The goal relevance of a topic is indicated by the color of the target or crossed target if the student is not ready for the topic. The number of arrows in the target indicates the topic knowledge state from little knowledge (no arrows) to very good knowledge (three arrows). Goal adaptation is supported by a simple time-based mechanism that switches the relevance of the topics according to the course lecture sequence. Knowledge adaptation is supported by topic-based knowledge modeling. Together, these mechanisms help the student choose the topic to work on by indicating which topics are most important and which need additional work. A click on the selected topic opens links to quizzes for this topic. A click on a quiz link loads the first

question of this quiz in the quiz presentation area (right on Fig. 1). The quizzes in the presentation area are generated and evaluated by QuizPACK system [7].

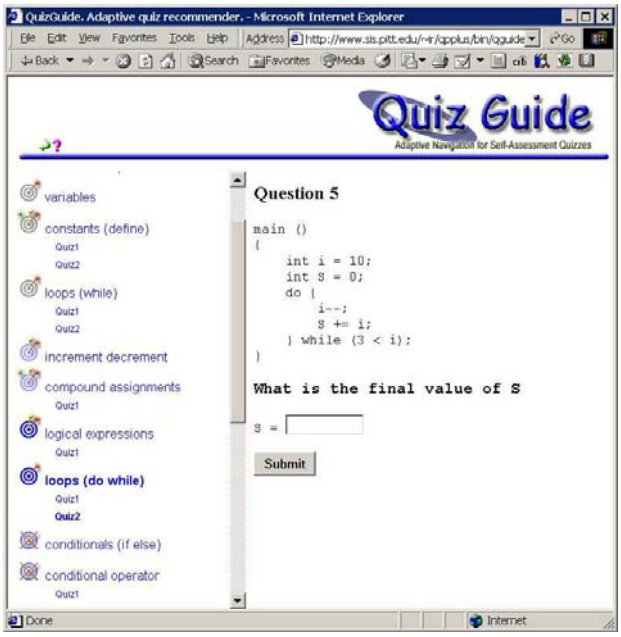


Fig. 1. Student interface of QuizGuide

As a KnowledgeTree service QuizGuide does not change QuizPACK. It stays between the user and the QuizPACK *activity server* providing value-added service – adaptive annotations. To generate the adaptive icons, QuizGuide requests the current student knowledge level of all topics (inferred by the topic-based agent) through CUMULATE query interface. Comparing the knowledge of each topic with three pre-determined thresholds, the system selects an icon with the proper number of arrows (zero to three). Current thresholds are *0.1*, *0.3*, and *0.5*.

4 Layered Evaluation of Topic-Based Adaptation

The user modeling literature provides no guidance how to evaluate a universal user modeling server. What kind of evidence we can provide in favor of CUMULATE and topic-based modeling approach? The evaluation of specific models and adaptive systems has been widely discussed [8] and traditional "with or without" approach is considered as a golden standard. However, what is really evaluated in a "with or without" study of QuizGuide driven by a tunable user modeling server [9]? Are we evaluating the server itself, the topic-based student modeling approach implemented

by one of the inference agents, or just the quality of the job done by the author in defining topics and connecting them with activities? Could we consider as a "proof" the very ability to implement a new student modeling approach and to author the student modeling part of a new application? While appreciating this problem, our paper provides no answer to it so far. Instead, we report our attempts to evaluate the results of our work using layered approach [10], which advocates the need to evaluate separately the user modeling and adaptation parts of an adaptive system.

As the source data for this study we have used two semesters (Spring and Fall of 2004) of student activity with QuizPACK/QuizGuide performed in the context of undergraduate course *Introduction to Programming*. Generally, both systems QuizPACK and QuizGuide were available to students at the same time. Both QuizGuide and QuizPACK transactions (question-answering attempts) were collected by CUMULATE and used as the source for QuizGuide adaptation. To reduce the number of potentially noisy transactions we have filtered out those students who have not performed the minimum required amount of activity with the system (30 questions). Table 1 summarizes the basic statistics of the source data.

Table 1. Quantitative description of the source data

Users	Topics	Quizzes	Questions	QuizPACK Transactions	QuizGuide Transactions
34	22	44	171	4960	5217

4.1 Evaluation of Knowledge Modeling

Evaluation of knowledge modeling process can be further decomposed into two phases: evaluation of knowledge elements describing the domain and evaluation of algorithms/heuristics used for the inference of values characterizing student knowledge for specific knowledge elements.

Topic as an Assessment Unit

To evaluate how well CUMULATE assesses student knowledge we first need to make sure that the units of knowledge measurement are suitable. The following modeling approach as well as the implemented adaptation strategy could be reasonable, however the resulting adaptive behavior of the system might be not adequate to the student's actions and expectations, if the assessment units are wrong.

For evaluation of large topics as assessment units we have applied learning curve analysis [11]. Multiple experiments provide strong evidence that the learning process follows the power law (see for example [12], [13]). In other words, the error rate of some leaning skill decreases as the power function of the number of learning steps involving this skill. Figure 2a demonstrates the dependency between percentage of incorrectly answered questions (served by either QuizGuide or QuizPACK) averaged by topics and students and the number of questions attempted by students on this topic before. Though downward trend witnesses some learning effect, the curve is not smooth and R^2 statistics tells that only about 28% of the variability in the error rate

could be explained by power dependence on the number of steps. At the same time learning curves in the figures 2b and especially 2c, which visualize the same data correspondingly for quizzes and questions instead of topics, have much better fit to the power law. Figure 2b demonstrates the increase of student knowledge on the average QuizGuide quiz. The learning curve in the figure 2c corresponds to separate questions. Hence, when students practice with a specific question the character of learning process is very close to the power law (almost three fourth of the variability in the question performance is explained by the power dependence on the number of steps). However, the topic (combining on average about 8 questions) does not seem to be the best unit on which we can base the assessment and consequently modeling of student knowledge. The main drawback of a topic is the large amount of covered knowledge. While appreciated by a teacher (for whom the authoring time of topic-based intelligent content reduces considerably), this feature results in two challenges that the system developer faces on both stages: modeling and adaptation.

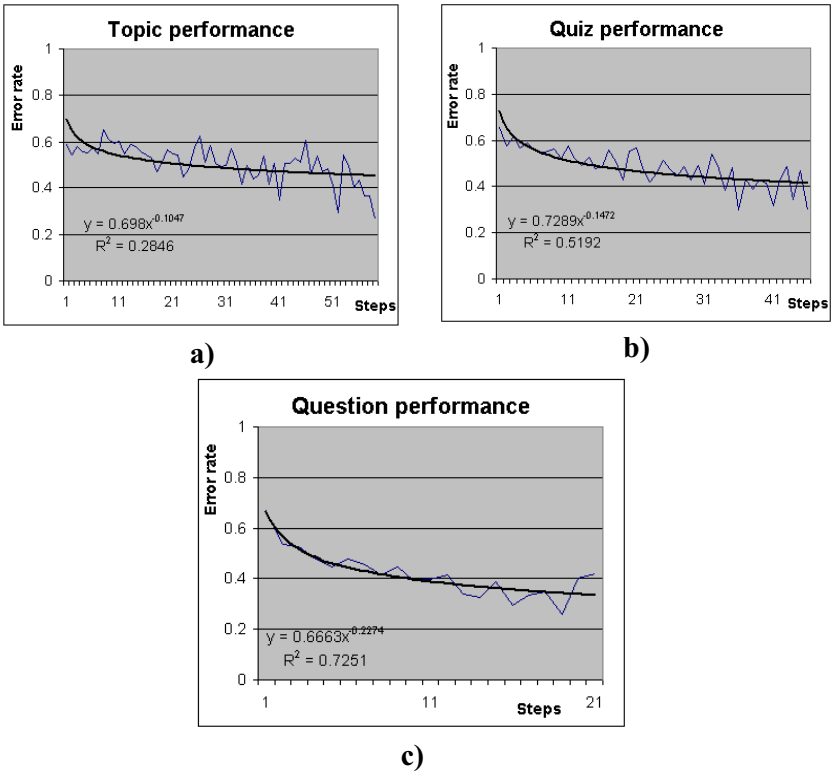


Fig. 2. Learning curves for the average topic (a), quiz (b) and question (c)

First, since topics involve too much knowledge, the precision of the assessment performed by the system is reduced. When taking a quiz on some topic a student can

make a mistake in a number of interrelated concepts used by the questions of the quiz. The smaller the scope of assessment is the closer is the model, built by a system, to the real state of student knowledge and the closer is the learning curve to the power law (figures 2a, 2b, 2c).

On the stage of adaptation large topics reduce the potential accuracy of adaptive interventions, which is inversely proportional to the size of the knowledge element. When adapting to knowledge a system traditionally needs to estimate the best learning activity according to the current levels of student knowledge and calculated measures of difficulty for different activities. In our case QuizGuide informs students about their levels of knowledge for topics; no difference is made for specific quizzes and questions. At the same time, naturally, questions vary in structural complexity and difficulty estimates for different students and for different periods of time. The lack of ability to provide precise information about the difficulty of any specific question results in the situation, when some questions taken by the students are either too easy or too hard.

To investigate this problem we estimated average question difficulty using the traditional measure – the mean value of error rate. The histogram and the boxplot in the figure 3 demonstrate the distribution of questions according to their difficulty. There are no visual difficulty outliers; however, the analysis of influence of the hardest and the easiest questions on the learning curve demonstrates that by filtering out such questions we can make the learning curve considerably closer to the power law.

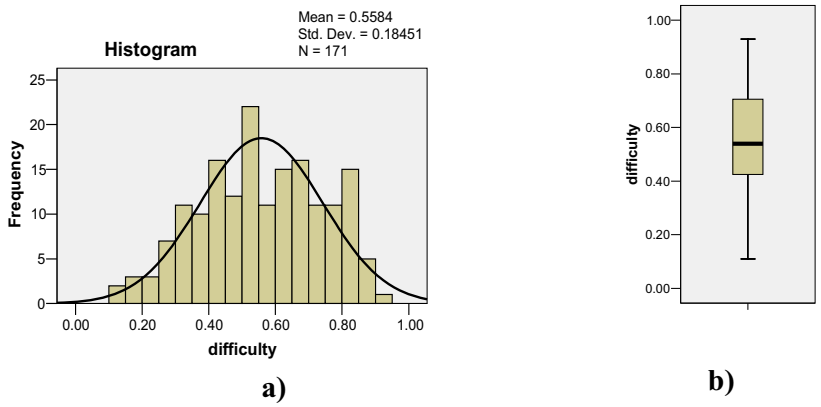
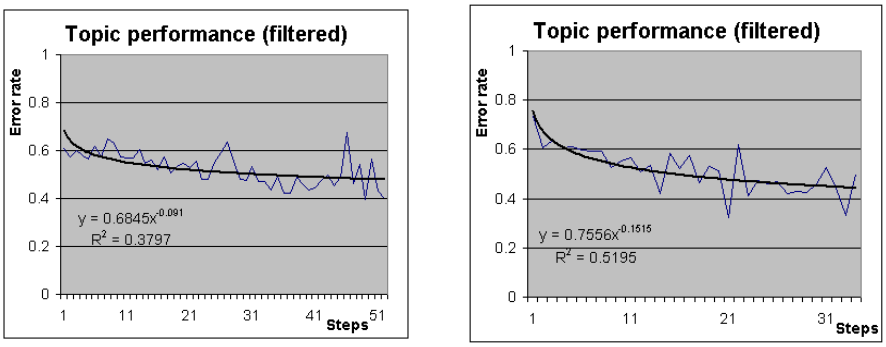


Fig. 3. The histogram (a) and the boxplot (b) of QuizGuide questions distributed according to their average difficulty.

We explored our data on two intervals traditionally used for estimating main trends that could be explained by the central part of the distribution: 90% and 50%. Plot 4a shows the very same learning curve as in the figure 2a, where 5% of most difficult and 5% of least difficult questions are removed from the plot. The hypothesis was that when the question was too hard or too simple the effectiveness of learning is reduced and such questions could be disregarded. As we see the fit is much better then in the

figure 2a. If we filter out all questions below 25th and above 75th percentile according to the error rate distribution, the results are even better (see fig. 4b). More then half of error rate variability is explained by the power dependence on the number of attempts. Hence, the possible outcome of this analysis is: if we adequately manipulated the difficulty of the presented question or at least provided students with reliable information on the question knowledge level to decrease the number of too hard or too simple questions taken, we could improve the learning process.



a) **b)**

Fig. 4. Learning curves for the average topic, where 10% (a) and 50% (b) of "bad" questions are filtered out

AdHoc Knowledge Level Calculation

Next step is the evaluation of the method used for calculation of knowledge levels for specific topics. Current adaptation agent uses fairly simple heuristics – *Average of sums of averages*:

$$K_i = \frac{\sum_{j=1}^{N_i} w_{ij} \frac{\sum_{k=1}^{M_j} x_{jk} / z_{jk}}{M_j}}{\sum_{j=1}^{N_i} w_{ij}}, \text{ where}$$

K_i – current level of knowledge for i^{th} topic,

N_i – number of activities(quizzes) participating in i^{th} topic,

M_j – number of sub-activities(questions) in j^{th} activity,

w_{ij} - weigh of influence of j^{th} activity on ith topic

x_{jk} – number of correct attempts the student has for the k^{th} sub-activity of j^{th} activity,

z_{jk} – total number of attempts the student has for the k^{th} sub-activity of j^{th} activity.

The adequacy of this formula could be estimated by assessing the correspondence between the system's predictions of knowledge levels and the actual results students get. In other words, the correlation coefficient between values computed by the system

to characterize student's knowledge of the specific topic at different times and the results of the attempts students perform on the next steps is expected to be a good measure of this formula's ability to predict student knowledge. However, since the topics are shown to be unsuitable assessment units, we could hardly expect that any topic-based computation might provide a reliable model of student knowledge. The average correlation coefficient (*cor = -0.18*) does not allow us to prove the robustness of used modeling heuristics.

4.2 Evaluation of the Value of Adaptation

While the current settings offered no meaningful way to evaluate the quality of knowledge modeling (the quality of modeling is best evaluated by test questions, but this data is already used by QuizGuide for modeling), we have attempted to measure the value of the adaptation part by checking how different kinds of adaptive icons influenced student behavior. Figure 5 averages demonstrates average number of QuizGuide attempts made by 11 students of the Fall 2004 course to quizzes marked with different number of arrows. QuizPACK transactions here are disregarded when attempts are counted, however they still participate in the calculation of adaptive annotation (number of arrows). We exclude Spring semester students because QuizGuide was introduced only in the middle of the semester and the pattern of system usage was different from "pure" settings of the Fall, when both systems were equally available from the beginning.

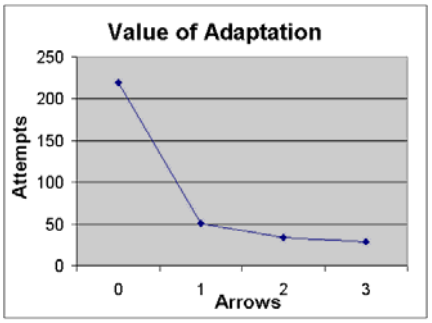


Fig. 5. Value of adaptation provided by QuizGuide

As we see the vast majority of visits (220) were made to topics with little demonstrated knowledge annotated by a target with no arrows. As long as the demonstrated level of knowledge increases, the number of visits decreases to 51 (one arrow), 34 (two arrows), and 29 (three arrows). It seems that the students' motivation to work decreases when they believe that some reasonable level of knowledge is achieved. The appearance of at least one arrow is the most important threshold for them. Difference between visiting quizzes with one or more arrows is very small in

comparison to the difference between quizzes with no arrows and quizzes with one arrow".

5 "With or without" Evaluation of Topic-Based Adaptation

The size and the focus of the paper does not allow us to include detailed traditional "with and without" evaluation. These details can were presented elsewhere [9]. Despite relatively simple user modeling and adaptation techniques used in QuizGuide, the system has achieved a remarkable impact on student learning and performance. Guided by adaptive annotations, students explored 50% more questions, worked with questions more persistently (24 vs. 14 question attempts per session), and accessed a larger variety of questions. There is also an evidence that QuizGuide succeeded in helping the students to select questions of proper difficulty: the percentage of correctly answered questions in QuizGuide sessions is also higher: 44.3% versus 35.6% in QuizPACK sessions. The increase of their work with the system resulted in the larger increase of their knowledge at the end of the course: the average knowledge gain in rose from 5.1 in QuizPACK-only class to 6.5 in a class with access to QuizGuide.

6 Summary

We have presented the student modeling architecture standing behind QuizGuide system that helps students in selecting most relevant self-assessment. QuizGuide uses adaptive annotation technology to show the students their current knowledge level for each course topic and current relevance of these topics. The paper demonstrates our student modeling framework that allows interested authors to quickly implement efficient adaptive systems.

At the same time the reported results of layered evaluation show that using large topics as knowledge assessment units imposes serious problems on both stages: modeling and adaptation. Though the results of evaluation of modeling layer do not allow us to prove the robustness of used heuristics, students seem to benefit from the adaptation provided by the system. The average pattern of usage shows that they follow the navigational guide provided by the system.

References

1. Brusilovsky P (2004) KnowledgeTree: A distributed architecture for adaptive e-learning. In The Thirteenth International World Wide Web Conference, WWW 2004 (Alternate track papers and posters), New York, NY, 17-22 May, 2004, pp. 104-113
2. Carmona C and Conejo R (2004) A learner model in a distributed environment. In De Bra P and Nejdl W (eds) Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2004), Eindhoven, the Netherlands, August 23-26, 2004. Lecture Notes in Computer Science 3137, Springer-Verlag, Berlin, pp. 353-359

3. Conlan O, Wade V, Gargan M, Hockemeyer C, and Albert D (2002) An architecture for integrating adaptive hypermedia services with open learning environments. In Barker P and Rebelsky S (eds) ED-MEDIA'2002 - World Conference on Educational Multimedia, Hypermedia and Telecommunications, Denver, CO, June 24-29, 2002, pp. 344-350
4. Mödritscher F, García Barrios VM, and Gütl C (2004) Enhancement of SCORM to support adaptive E-Learning within the Scope of the Research Project AdeLE. In Nall J and Robson R (eds) World Conference on E-Learning, E-Learn 2004, Washington, DC, USA, November 1-5, 2004, pp. 2499-2505
5. Brusilovsky P, Eklund J, and Schwarz E (1998) Web-based education for all: A tool for developing adaptive courseware Computer Networks and ISDN Systems 30 1-7, 291-300
6. Lundgren-Cayrol K, Paquette G, Miara A, Bergeron F, Rivard J, and Rosca I (2001) Explor@ Advisory Agent: Tracing the Student's Trail. In Fowler W and Hasebrook J (eds) WebNet'2001, World Conference of the WWW and Internet, Orlando, FL, October 23-27, 2001, pp. 802-808
7. Pathak S and Brusilovsky P (2002) Assessing Student Programming Knowledge with Web-based Dynamic Parameterized Quizzes. In Barker P and Rebelsky S (eds) ED-MEDIA'2002 - World Conference on Educational Multimedia, Hypermedia and Telecommunications, Denver, CO, June 24-29, 2002, pp. 1548-1553
8. Chin D (2001) Empirical Evaluations of User Models and User-Adapted Systems. User Modeling and User-Adapted Interaction 11: 181-194
9. Brusilovsky P, Sosnovsky S, and Shcherbinina O (2004) QuizGuide: Increasing the Educational Value of Individualized Self-Assessment Quizzes with Adaptive Navigation Support. In Nall J and Robson R (eds) World Conference on E-Learning, E-Learn 2004, Washington, DC, USA, November 1-5, 2004, pp. 1806-1813
10. Brusilovsky P, Karagiannidis C, and Sampson D (2004) Layered evaluation of adaptive learning systems. IJCEELL 15
11. Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), Cognitive skills and their acquisition. 1-51. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
12. Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R.. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences,* 4 (2) 1995, 167-207.
13. Mitrovic, A., Mayo, M., Suraweera, P and Martin, B. In: L. Monostori, J. Vancza and M. Ali (eds), Proc. 14th Int. Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems IEA/AIE-2001, Budapest, June 2001, Springer-Verlag Berlin Heidelberg LNAI 2070, pp. 931-940.

Acknowledgements

The work reported in this paper is supported by NSF grant # 0310576 *Individualized Exercises for Assessment and Self-Assessment of Programming Knowledge.*

Problems and Pitfalls in Evaluating Adaptive Systems[1]

Stephan Weibelzahl

National College of Ireland, Mayor Street, Dublin 1, Ireland
sweibelzahl@ncirl.ie

Abstract. Empirical studies with adaptive systems offer many advantages and opportunities. Nevertheless, there is still a lack of evaluation studies. This paper lists several problems and pitfalls that arise when evaluating an adaptive system and provides guidelines and recommendations for workarounds or even avoidance of these problems. Among other things the following issues are covered: relating evaluation studies to the development cycle; saving resources; specifying control conditions, sample and criteria; asking users for adaptivity effects; reporting results. An overview of existing evaluation frameworks shows which of these problems have been addressed in which way.

1. Introduction

Empirical evaluation of adaptive learning systems is a very important task, as the lack of strong theories, models and laws requires that we do evaluative experiments that check our intuition and imagination. Researchers from various fields have made experiments and published a considerable amount of experimental data. Many of these data sets can be valuable form adaptive learning systems. Still, most of the results are given in a textual form, while structure of these results is not standardized. This limits the practical value of the results. Therefore, if we want to improve the usefulness of the experimental results, it is important to make more formal descriptions of them. The first step toward this goal is creation of the metamodel of empirical evaluation that should identify concepts such as evaluation style, methods and evaluation approaches. This metamodel serves as a conceptual basis form various applications, such as metadescription of experimental data, and creation of experimental data warehouses. Based on this metamodel various tools can work together on creation and processing and comparative analysis of these experimental data.

Given the observation above, it seems obvious that empirical research is of high importance for the field both from a scientific as well as from a practical point of view because it opens up various advantages and opportunities (Weibelzahl, Lippitsch, &Weber, 2002). For example, empirical evaluations help to estimate the effectiveness, the efficiency, and the usability of a system.

---

[1] This paper is a summary of 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

Adaptive systems adapt their behavior to the user and/or the user's context. The construction of a user model usually requires claiming many assumptions about users' skills, knowledge, needs or preferences, as well as about their behavior and interaction with the system. Empirical evaluation offers an unique way of testing these assumptions in the real world or under more controlled conditions. Moreover, empirical evaluations may uncover certain types of errors in the system that would remain otherwise undiscovered. For instance, a system might adapt perfectly to a certain combination of user characteristics, but is nevertheless useless if this specific combination simply does not occur in the target user group. Thus, empirical tests and evaluations have the ability to improve the software development process as well as the final system considerably. However, they should be seen as complement rather than a substitute to existing software engineering methods such as verification, validation, formal correctness, testing, and inspection.

2. Problems and Pitfalls

In spite of these reasons in favor of an empirical approach, publications on user modeling systems and adaptive hypermedia rarely contain empirical studies: Only about one quarter of the articles published in User Modeling and User Adapted Interaction (UMUAI) report significant evaluations (Chin, 2001). Researchers have been lamenting on this lack frequently (Eklund & Brusilovsky, 1998; Masthoff, 2002), and similar situations have been identified in other scientific areas, too, for instance in software engineering (Kitchenham et al., 2002) or medicine (Yancey, 1996). One important reason for the lack of empirical studies might be the fact that empirical methods are not part of most computer science curricula, and thus, many researchers have no experience with the typical procedures and methods that are required to conduct an experimental study. Moreover, the evaluation of adaptive systems includes some inherent problems and pitfalls that can easily corrupt the quality of the results and make further conclusions impossible. Other problems arise from the nature of empirical work in general. These problems include (Weibelzahl, 2004):

- Formative vs. Summative Evaluation: Often evaluation is seen as the final mandatory stage of a project. While the focus of many project proposals is on new theoretical considerations or some innovative features of an adaptive system, a summative evaluation study is often planned in the end as empirical validation of the results. However, when constructing a new adaptive system, the whole development cycle should be covered by various evaluation studies.
- Allocation of sufficient resources: The fact that evaluations are usually scheduled for the end of a project often results in a radical constriction or even total cancellation of the evaluation phase, because the required resources have been underestimated or are depleted. Empirical work, in particular the data assessment and analysis, require a high amount of personnel, organizational and sometimes even financial resources (Masthoff, 2002). Experiments and real world studies require a considerable amount of time for planning, finding participants, performing the actual data assessment, coding the raw data and statistical analysis.

- Specification of adequate control conditions: Another problem, that is inherent to the evaluation of adaptive systems, occurs when the control conditions of experimental settings are defined. In many studies the adaptive system is compared to a non-adaptive version of the system with the adaptation mechanism switched off (Brusilovsky & Eklund, 1998). However, adaptation is often an essential feature of these systems and switching the adaptivity off might result in an absurd or useless system (Höök, 2000). In some systems, in particular if they are based on machine learning algorithms (e.g., Krogsæter, Oppermann, & Thomas, 1994), it might even be impossible to switch off the adaptivity.
- Sampling strategy: A proper experimental design requires not only to specify control conditions but of course also to select adequate samples. On the one hand the sample should be very heterogeneous in order to maximize the effects of the system's adaptivity: the more the differences between users the higher the chances that the system is able to detect these differences and react accordingly. On the other hand, from a statistical point of view, the sample should be very homogeneous in order to minimize the secondary variance and to emphasize the variance of the treatment. It has been reported frequently that too high variance is a cause of the lack of significance in evaluation studies (Brusilovsky & Pesin, 1998; Masthoff, 2002; Mitrovic & Martin, 2002). For instance, learners in online courses usually differ widely in reading times which might corrupt further comparisons in terms of time savings due to adaptive features.
- Definition of criteria: Evaluating the adaptivity of a system is sometimes seen as a usability-testing problem (Strachan, Anderson, Sneesby, & Evans, 1997). Obviously, usability is an important issue and most adaptive features actually aim at improving the usability. However, there are several aspects of adaptivity that are not covered by usability. For instance, adaptive learning systems usually aim at improving the learning gain in the first place, rather than the usability. The effectiveness and efficiency of other systems are measured in very different ways, as the adaptivity in these systems aims at optimizing other aspects, i.e., the criteria are determined by the system goal and its domain. More details on appropriate evaluation criteria are given below.
- Asking for Adaptivity Effects: In many studies the users estimate the effect of adaptivity (e.g., Beck, Stern, & Woolf, 1997) or rate their satisfaction with the system (e.g., Bares & Lester, 1997; Encarnação & Stoev, 1999; Fischer & Ye, 2001) after a certain amount of interaction. However, from a psychological point of view these assessment methods might be inadequate in some situations. Users might have no anchor of what good or bad interaction means for the given task if they do not have any experience with the 'usual' non-adaptive way. Moreover, they might not even have noticed the adaptivity at all, because adaptive action often flows (or should flow) in the subjective expected way rather than in the static predefined way (i.e., rather than prescribing a certain order of tasks or steps, an adaptive system should do what the user wants to do). Thus, the users might notice and hence be able to report only those events when the system failed to meet their expectations.
- Reporting the Results: Even a perfect experimental design will be worthless if the results are not reported in a proper way. In particular statistical data require special care, as the finding might be not interpretable for other researchers if relevant

information is skipped. This problem obviously occurs in other disciplines and research areas that deal with empirical findings, too. Thus, there are many guidelines and standard procedures for reporting empirical data as suggested or even required by some journals (e.g., Altman, Gore, Gardner, & Pocock, 1983[2]; Lang & Secic, 1997; Begg et al., 1996; Wilkinson & Task Force on Statistical Inference, 1999[3]).

3. Evaluation Approaches

To address at least some of the problems mentioned above, several evaluation frameworks were introduced. These frameworks build upon the idea that the evaluation of adaptive systems should not treat adaptation as a singular, opaque process; rather, adaptation should be "broken down" into its constituents, and each of these constituents should be evaluated separately where necessary and feasible. The seeds of this idea can be traced back to Totterdell and Boyle (1990), who propose that a number of adaptation metrics be related to different components of a logical model of adaptive user interfaces, to provide what amounts to adaptation-oriented design feedback.

The layered evaluation approach (Brusilovsky, Karagiannidis, & Sampson, 2001; Karagiannidis & Sampson, 2000) suggests to separate the *interaction assessment* and the *adaptation decision*. Both layers should be evaluated separately in order to be able to interpret the evaluation results properly. If an adaptation is found to be unsuccessful, the reason is not evident: either the system has chosen the wrong adaptation decision, or the decision was based on wrong assessment results.

Based on these first ideas on layered evaluation, two more frameworks have been introduced that slice the monolithic adaptive system into several layers (respectively stages) that can then be evaluated separately or in combinations (Paramythis, Totter, & Stephanidis, 2001; Weibelzahl, 2001). Recently, these frameworks have been merged, and some validating evidence has been presented (Paramythis & Weibelzahl, submitted). According to this new proposal there are five stages that might be evaluated separately: *collection of input data*, *interpretation of data*, *modeling the current state of the world*, *deciding upon adaptation*, and *applying adaptation*.

In addition, utility-based evaluation of adaptive systems (Herder, 2003) offers a perspective of how to reintegrate the different layers again.

Magoulas et al. (2003) introduced an integration of the layered evaluation approach and heuristic evaluation. Based on existing heuristics that have been used in human-computer interaction (Nielsen, 1994; Chen & Ford, 1998) the authors propose a set of refined heuristics and criteria for every layer. For instance the *acquisition of input data* is evaluated by a heuristic called *error prevention*. It is conducted by checking for typical error prevention techniques (e.g., *data inputs are case-blind whenever possible* or *when learners navigate between multiple windows, their answers are not lost*). In summary, the approach guides the diagnosis of design problems at an early design stage and can thus be seen as a complement to the other frameworks.

---

[2] available at http://bmj.com/advice/

[3] available at http://www.apa.org/journals/amp/amp548594.html

The layered evaluation approach might also be extended by dicing rather than slicing the interaction. Groups of users or even single users might be observed across the layers. Thus, the focus is shifted from the whole sample on one layer to a subset of the sample across layers. For example, the evaluation of an adaptive online course could analyze learners with high and low reading speed separately in order to demonstrate that the inference mechanism works better for one group than for the other. In summary, this perspective might identify sets of (unmodeled but controlled) user characteristics that require a refinement of the user model or at least shape the evaluation results.

It has also been proposed to facilitate evaluation processes through separating design perspectives (Tobar, 2003). The framework integrates abstract levels, modeling issues, traditional concerns, and goal conditions into a so-called extended abstract categorization map which guides the evaluation process. Thus, it addresses in particular the problem of defining adequate evaluation criteria.

This diversity of frameworks and approaches might look a little bit confusing at first glance, but in fact it is a mirror of the current state of the art.

4. Evaluation Criteria

The frameworks and approaches described above provide some guidance concerning adequate criteria for evaluation at each layer. However, the evaluation of the effectiveness and efficiency of a system requires a precise specification of the modeling goals in the first place, as this is a prerequisite for the definition of the criteria. The criteria might be derived from the abstract system goals for instance by using the Goal-Question-Metric method (GQM) (van Solingen & Berghout, 1999), which allows to systematically define metrics for a set of quality dimensions in products, processes and resources. Tobar (2003) presented a framework that supports the selection of criteria by separating design perspectives (see above).

Weibelzahl (2003) also provides an extended list of criteria that have been found in current evaluation studies. For adaptive learning systems obviously the most important and commonly applied criterion is learning gain. However, other general criteria such as learner satisfaction, development of communication or problem solving skills, learner motivation, etc might have to be considered, too. The layered evaluation approach would also suggest evaluating system factors such as the reliability and validity of the input data, the precision of the student model, or the appropriateness of the adaptation decision.

The diversity of these criteria currently inhibits a comparison of different modeling approaches. Future research should aim at establishing a set of commonly accepted criteria and assessment methods that can be used independent of the actual user model and inference mechanism in order to explore the strength and weaknesses of the different modeling approaches across populations, domains, and context factors. While current evaluation studies usually yield a single data point in the problem space, common criteria would allow integrating the results of different studies to a broader picture. Utility-based evaluation (Herder, 2003) offers a way how such a comparison across systems could be achieved.

References

Altman, D., Gore, S., Gardner, M., & Pocock, S. (1983). Statistical guidelines for contributors to medical journals. *British Medical Journal*, *286*, 1489–1493.

Bares, W. H., & Lester, J. C. (1997). Cinematographic user models for automated realtime camera control in dynamic 3D environments. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (pp. 215–226). Vienna, New York: Springer.

Beck, J., Stern, M., & Woolf, B. P. (1997). Using the student model to control problem difficulty. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (pp. 277–288). Vienna, New York: Springer.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schultz, K., Simel, D., & Stroup, D. (1996). Improving the quality of reporting randomized trials (the CONSORT statement). *Journal of the American Medical Association*, *276*(8), 637–639.

Billsus, D., & Pazzani, M. J. (1999). A hybrid user model for news story classification. In J. Kay (Ed.), *User modeling: Proceedings of the Seventh International Conference, UM99* (pp. 98–108). Vienna, New York: Springer.

Brusilovsky, P., & Eklund, J. (1998). A study of user-model based link annotation in educational hypermedia. *Journal of Universal Computer Science, special issue on assessment issues for educational software*, *4*(4), 429–448.

Brusilovsky, P., Karagiannidis, C., & Sampson, D. G. (2001). The benefits of layered evaluation of adaptive applications and services. In S. Weibelzahl, D. N. Chin, & G. Weber (Eds.), *Empirical Evaluation of Adaptive Systems. Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001* (pp. 1–8). Sonthofen, Germany.

Brusilovsky, P., & Pesin, L. (1998). Adaptive navigation support in educational hypermedia: An evaluation of the ISIS-tutor. *Journal of Computing and Information Technology*, *6*(1), 27–38.

Chen, S. Y., & Ford, N. (1998). Modelling user navigation behaviours in a hypermedia based learning system: An individual differences approach. *International Journal of Knowledge Organization*, *25*(3), 67–78.

Chin, D. N. (2001). Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, *11*(1-2), 181–194.

Chiu, B. C., Webb, G. I., & Kuzmycz, M. (1997). A comparison of first-order and zerothorder induction for Input-Output Agent Modelling. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (pp. 347–358). Vienna, New York: Springer.

Eklund, J., & Brusilovsky, P. (1998). The value of adaptivity in hypermedia learning environments: A short review of empirical evidence. In P. Brusilovsky & P. de Bra (Eds.), *Proceedings of Second Adaptive Hypertext and Hypermedia Workshop at the Ninth ACM International Hypertext Conference Hypertext'98, Pittsburgh, PA, June 20, 1998* (pp. 13–19). Eindhoven: Eindhoven University of Technology.

Encarnação, L. M., & Stoev, S. L. (1999). Application-independent intelligent user support system exploiting action-sequence based user modeling. In J. Kay (Ed.), *User modeling: Proceedings of the Seventh International Conference, UM99* (pp. 245–254). Vienna, New York: Springer.

Fischer, G., & Ye, Y. (2001). Personalizing delivered information in a software reuse environment. In M. Bauer, J. Vassileva, & P. Gmytrasiewicz (Eds.), *User modeling: Proceedings of the Eighth International Conference, UM2001* (pp. 178–187). Berlin: Springer.

Herder, E. (2003). Utility-based evaluation of adaptive systems. In S. Weibelzahl & A. Paramythis (Eds.), Proceedings of the Second Workshop on Empirical Evaluation of

Adaptive Systems, held at the 9th International Conference on User Modeling UM2003 (pp. 25–30). Pittsburgh.

Höök, K. (2000). Steps to take before intelligent user interfaces become real. *Interacting With Computers*, *12*(4), 409–426.

Karagiannidis, C., & Sampson, D. G. (2000). Layered evaluation of adaptive applications and services. In P. Brusilovsky & C. S. O. Stock (Eds.), *Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH2000, Trento, Italy* (pp. 343–346). Berlin: Springer.

Kitchenham, B., Pfleeger, S. L., Pichard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., & Rosenberg, J. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, *28*(8), 721–733.

Krogsæter, M., Oppermann, R., & Thomas, C. G. (1994). A user interface integrating adaptability and adaptivity. In R. Oppermann (Ed.), *Adaptive user support* (pp. 97–125). Hillsdale: Lawrence Erlbaum.

Lang, T., & Secic, M. (1997). How to report statistics in medicine: Annotated guidelines for authors, editors and reviewers. Philadelphia, PA: American College of Physicians.

Magnini, B., & Strapparava, C. (2001). Improving user modeling with content-based techniques. In M. Bauer, J. Vassileva, & P. Gmytrasiewicz (Eds.), *User modeling: Proceedings of the Eighth International Conference, UM2001* (pp. 74–83). Berlin: Springer.

Magoulas, G. D., Chen, S. Y., & Papanikolaou, K. A. (2003). Integrating layered and heuristic evaluation for adaptive learning environments. In S.Weibelzahl & A. Paramythis (Eds.), *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003* (p. 5-14). Pittsburgh.

Masthoff, J. (2002). The evaluation of adaptive systems. In N. V. Patel (Ed.), *Adaptive evolutionary information systems.* Hershey, PA: Idea Group Publishing.

Mitrovic, A., & Martin, B. (2002). Evaluating the effects of open student models on learning. In P. de Bra, P. Brusilovsky, & R. Conejo (Eds.), *Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, Málaga, Spain, AH2002* (pp. 296–305). Berlin: Springer.

Nielsen, J. (1994). Heuristic evaluation. Usability inspection methods. New York: Wiley.

Paramythis, A., Totter, A., & Stephanidis, C. (2001). A modular approach to the evaluation of adaptive user interfaces. In S. Weibelzahl, D. N. Chin, & G. Weber (Eds.), *Empirical Evaluation of Adaptive Systems. Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001* (pp. 9–24). Freiburg.

Paramythis, A. & Weibelzahl, S. (submitted). A Decomposition Model for the Layered Evaluation of Interactive Adaptive Systems.

Strachan, L., Anderson, J., Sneesby, M., & Evans, M. (1997). Pragmatic user modeling in a commercial software system. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (pp. 189–200). Vienna, New York: Springer.

Totterdell, P. & Boyle, E. (1990). The Evaluation of Adaptive Systems. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive User Interfaces* (pp. 161-194). London: Academic Press.

Tobar, C. M. (2003). Yet another evaluation framework. In S. Weibelzahl & A. Paramythis (Eds.), Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003 (pp. 15–24). Pittsburgh.

van Solingen, R., & Berghout, E. (1999). The goal/question/metric method: A practical guide for quality improvement of software development. London: McGraw-Hill.

Weibelzahl, S. (2001). Evaluation of adaptive systems. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva (Eds.), *User Modeling: Proceedings of the Eighth International Conference, UM2001* (pp. 292–294). Berlin: Springer.

Weibelzahl, S. (2003). *Evaluation of adaptive systems.* Doctoral dissertation, University of Trier, Trier.

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

Weibelzahl, S., Lippitsch, S., & Weber, G. (2002). Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems. *Künstliche Intelligenz*, *3/02*, 17–20.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Yancey, J. (1996). Ten rules for reading clinical research reports. *American Journal of Orthodontics and Dentofacial Orthopedics*, *109*(5), 558–564.

Introduction to the First Adaptive System Evaluation Challenge

Abstract. The Fourth Workshop on Evaluation of Adaptive Systems launched the first "evaluation challenge". The challenge concerns an adaptive system, which recommends sequences of music clips to groups of users. Focusing on a real world problem, the challenge aimed to foster the development of innovative evaluation designs as well as encourage controversial discussion during the workshop. Participation in the challenge entailed proposing an empirical evaluation design purposely created to answer specific design questions regarding the system's modeling component. This chapter describes the task and participation requirements given to the participants. The following two chapters contain the two submissions received.

1. Introduction

The Evaluation Challenge is about an adaptive system with a need for evaluation. Focusing on a real world problem, the challenge aimed to foster the development of innovative evaluation designs as well as encourage controversial discussion during the workshop. Potential participants were required to submit a short proposal on how this system can be evaluated. The proposals received were presented during the workshop, and are included in these proceedings. Workshop participants discussed the presented proposals in the context of a dedicated workshop track. The subsequent sections contain the detailed system description and express evaluation requirements that proposals were asked to address.

2. The System

2.1. Description

The system to be evaluated is a recommender system. The recommendation domain is music clips. Specifically, the system recommends sequences of music clips, either for individual users, or groups of users (the later being the most typical use case of the system, and the one addressed by this challenge). Recommendations are based on models of individual user preferences in relation to individual clips. The primary goal of the adaptive component of the system is to recommend a sequence of clips that will achieve reasonable levels of satisfaction for all members of the group, throughout the sequence. The rest of this section presents certain aspects of the system in more detail.

2.1.1. The domain

The system's application domain is music clips. The following information is available to the system for each clip: (a) performing artist(s); (b) compilation(s) in which the clip has appeared; (c) music genre(s); (d) recording company. This information is available for all clips in the system's database. The system has no way of further analyzing clips or locating / deriving additional information about them.

2.1.2. User modeling

The system is capable of modeling the preferences of each individual for any music clip. Primary input for the models is provided in the form of ratings from 1 to 10 for each music clip for each individual. The system uses that input to perform a hybrid of content-based and collaborative filtering-based modeling. The content-based part identifies generalizable "patterns" in the user's preferences (e.g., preference for a particular music genre). The collaborative filtering-based part performs dynamic user clustering based on explicit clip preferences and inferred general preferences, and follows traditional approaches in enriching individual user models and maintaining aggregate virtual models for clusters.

Recently, the system has been extended to also model how happy each individual is as a consequence of having seen the clips so far. This is the main aspect of the system that the challenge addresses. It is the intention that the happiness of individuals be used as input for an improved selection algorithm (see next section for an overview of the system's recommendation algorithms).

Based on literature, the following assumptions have been made to underlie the happiness modeling

- A1. Mood impacts evaluative judgment: when people are in a good mood, they evaluate more positively.
- A2. People's affective forecasting can change their actual emotional experience: if you expect to like something, then you might end up liking it more than if you did not have any expectations (this is called assimilation).
- A3. Emotional reactions become less intense with time: happiness wears of.
- A4. The difference between a rating of 9 and 10 might feel higher than the difference between a 5 and a 6.
- A5. Mood cannot be of unbounded intensity.
- A6. Actual feelings experienced differ from those reported retrospectively.

Initially, when no clips have been viewed yet, the happiness of each individual is modeled as zero: $Happiness(<>) = 0$. The happiness of an individual who has viewed item i after already having viewed a sequence *items* is modeled as a function of their happiness with sequence *items*, and the impact on their happiness of new item i. There are three proposals for this modeling (with $0 \leq \delta \leq 1$):

1. $\text{Happiness}(items+<i>) = \delta * \text{Happiness}(items) + \text{Impact}(i)$
2. $\text{Happiness}(items+<i>) = (\delta * \text{Happiness}(items) + \text{Impact}(i)) / (1 + \delta)$
3. $\text{Happiness}(items+<i>) = \delta * \text{Happiness}(items) + \text{Impact}(i, \delta * \text{Happiness}(items))$
   with $\text{Impact}(i, s) = \text{Impact}(i) + (s - \text{Impact}(i)) * \varepsilon$, for all s and $0 \leq \varepsilon \leq 1$

Where for item i: $\text{Impact}(i) = \begin{cases} (\text{Rating}(i) - 5.5)^2, & \text{if Rating}(i) \geq 5.5 \\ -(\text{Rating}(i) - 5.5)^2, & \text{if Rating}(i) < 5.5 \end{cases}$

and Rating(*i*) is the inferred rating for a given user for item *i*.

Multiplying Happiness(*items*) with δ is done because of assumption A3. The quadratic definition of Impact(*i*) is because of assumption A4. Dividing by (1+ δ) in proposal 2 is to use some kind of average, rather than summation, and is partly done because of assumption A5. The use of Impact(i, δ * Happiness(items)) in proposal 3 is because of assumptions A1 and A2. Epsilon (ε) in proposal 3 models the extent to which the mood a user is already in influences that user's evaluative judgment (assumption A1): with $\varepsilon = 0$ there is no such influence, with $\varepsilon = 1$ the influence is so immense that the new clip cannot have any effect on the user's mood.

2.1.3.Recommendation algorithm

The system uses a selection algorithm to determine what item to show next (i.e., what is the next most suitable item to place in the clip sequence), based on the preferences of the individuals in the group. At the moment, it uses a Multiplicative Utilitarian Selection Algorithm (MUSA), which basically multiplies the (inferred) preference ratings of individuals, to arrive at the preference of the group as a whole.

The aim of the system is to keep the group as whole happy, ensuring that everybody in the group remains reasonably happy throughout the sequence. Based on the recent additions for happiness modeling, as described in the previous section, a new selection algorithm is being developed. In broad terms, this algorithm is based on MUSA, but excludes items that might bring an individual's happiness below a certain threshold from being added to the list.

2.2. Evaluation Goals

The evaluation aims to provide answers to at least one of the following questions:

- Which of the three proposals for modeling happiness (see previous section) succeed in making relatively valid predictions (i.e., when they predict that a clip will make an individual unhappy this is indeed the case, and vice versa, independently from the level of un-/happiness)?
- Which of the three proposals is best at predicting inter-individual differences in happiness (i.e., managing to determine that clip C would make user U1 happier than user U2)?
- Which of the three proposals achieves the highest modeling precision (i.e., manages to more precisely predict the user's happiness after having watched a clip / series of clips)?

There is only one restriction that the evaluation has to observe, namely it has to be of an empirical nature, involving end users. It is also perfectly acceptable if your evaluation design comprises more than one components, with only one of them being of empirical nature.

3. The Challenge

Challenge entries are, in essence, proposals of how the described system should be evaluated. Specifically, entries were required to explicitly:

- Propose how the described system, and in particular the happiness modeling and its underlying assumptions, can be evaluated (see the specific questions in the previous section). This should include a full description of one or more empirical designs: sampling, setting and material used, treatments (if applicable), dependent and independent variables, and analysis.
- Discuss what the main difficulties for this evaluation are, and how you propose to overcome them.

For the purpose of this challenge it can be assumed that the preference modeling has been shown by a previous, dedicated study, to be quite accurate.

4. Summary

The challenge has been widely advertised and care has been taken to keep the barrier for participation as low as possible. Nevertheless, only two challenge entries were received. Both entries were reviewed by the workshop's program committee and accepted for presentation at the workshop. Revised versions of the entries are included in these proceedings.

The workshop participants discussed the merits and shortcomings of the entries. An amalgamation of the proposals received, coupled with the comments made by workshop participants during the respective discussions, will form the basis for the evaluation of a real-world system that is very close to the one described for this challenge (the evaluation will be overseen by members of the workshop's organizing committee). The proposers will be included as co-authors on any publications that are based on the evaluation proposal.

Evaluating an Adaptive Music-Clip Recommender System

Tingshao Zhu and Russ Greiner

University of Alberta, Edmonton, Alberta T6G 2E8, Canada,
{tszhu, greiner}@cs.ualberta.ca

Abstract. In this paper, we propose an experiment design to address three evaluation goals of "First Adaptive System Evaluation Challenge" [1], and demonstrate how to achieve each of these goals.

1 Introduction

The system to be evaluated is a recommender system, specifically suggest music clip, either for individual users, or groups of users. Recently, the system has also been extended to model how happy each individual is as a consequence of having seen the clips so far, and the challenge is how to evaluate the user models (i.e., modelling the preferences of each individual) that have been used to recommend music clips.

Primary input for the models is provided in the form of ratings from 1 to 10 for each music clip for each individual. There are three proposals for modelling the happiness (i.e., M^1, M^2, and M^3). The system uses a selection model to determine which item to show next (i.e., what is the next most suitable item to place in the clip sequence), based on the individual's preferences of the viewed clips. The evaluation aims to provide answers to the following questions:

Relatively Valid Prediction
Which of the three proposals for modelling happiness succeed in making relatively valid predictions (i.e., when they predict that a clip will make an individual unhappy this is indeed the case, and vice versa, independently from the level of un-/happiness)?

Inter-individual Difference
Which of the three proposals is best at predicting inter-individual differences in happiness (i.e., managing to determine that clip C would make user U_1 happier than user U_2)?

Precision
Which of the three proposals achieves the highest modelling precision (i.e., manages to more precisely predict the user's happiness after having watched a clip/series of clips)?

[1] http://www.easy-hub.org/hub/workshops/um2005/challenge.html

2 Experiment Design

To evaluate the performance of the three happiness models, we can conduct an user study. The participants are given access to the recommender system, browsing the music clips by their own choices and ask for recommendations any time they want.

Everytime a subject is presented with a recommended clip, s/he is required to give an evaluation based on her/his happiness, from 1 to 10 (i.e., $ScoreEvaluated$). Here, since *the input is provided in the form of ratings from 1 to 10 for each music clip for each individual*, we thus use the primary input of the user's rating for the suggested clip as its $ScoreEvaluated$. The higher score, the happier the subject.

Whenever a subject asks for recommendation, the system will first randomly pick out one model from M_1, M_2, and M_3; calculate each clip how happy the user would be (i.e., $ScoreComputed$) if that clip were presented next (given the visited clips so far); then normalize $ScoreComputed$ to $[1, 10]$; and finally rank all the remaining clips based on their $ScoreComputed$.

The system will randomly choose one of the following two policies to select a clip as recommendation:

1. If there is only one clip with the highest score (i.e., $ScoreComputed = 10$), output it, or randomly output one clip if there are multiple clips that result in maximum happiness; otherwise follow the second option.
2. Choose one clip randomly, and output it.

For each evaluation, we will record the following information:

$$< UserID, ClipSeq, ModelID, ScoreComputed, ClipSuggested, ScoreEvaluated >$$

UserID
The identification of the subject.

ClipSeq
The music clip that the subject has visited so far.

ModelID
Which model has been chosen to generate the recommendation.

ScoreComputed
A continuous value from 1 to 10, denotes the happiness score of the suggested clip according to the selected model.

ClipSuggested
The music clip that presented to the subject as recommendation.

ScoreEvaluated
The primary input to show the subject's happiness for the suggested clip.

3 Evaluation Goals

Here, we will demonstrate how the proposed experiment design is able to address each of these evaluation goals.

3.1 Relatively Valid Prediction

Alternatively, we want to find which model will generate the least difference (statistically significant) between *ScoreComputed* and *ScoreEvaluated*. It is expected that the model which generates less differences will be more promising for making relatively valid predictions. Note that the primary goal focuses on un-/happy, we will test these suggested clips with extreme *ScoreComputed*, either happy ($ScoreComputed = 10$) or unhappy ($ScoreComputed = 1$). Since *ScoreComputed* can approach *ScoreEvaluated* in either way, we only care about the absolute value of the difference, that is, $|ScoreComputed - ScoreEvaluated|$.

For each subject s_i, we collect the suggested clips with $ScoreComputed = 10$ or $ScoreComputed = 1$, then compute the mean of the differences between *ScoreComputed* and *ScoreEvaluated* for each of three models (i.e., M^1, M^2, and M^3).

Table 1. Subject s_i's ($|ScoreComputed - ScoreEvaluated|$) for all three models

	M^1	M^2	M^3
	5	0.6	2.4
	4	0.55	1.5
		0.45	0.9
Mean	4.5	0.53	1.6

For example, in Table 1, s_i has asked 8 times for recommendations, in which the suggested clips had *ScoreComputed* either 10 or 1. Among these 8 recommendations, the system has selected M^1 2 times, 3 times for both M^2 and M^3. For each suggested clip, we compute the absolute difference between its *ScoreComputed* and *ScoreEvaluated*. Then we calculate the average difference for each model which is shown in Table 1. After we have computed the average difference for each model of each subject, we can build a happiness difference matrix as shown in Table 2.

Table 2. Happiness Differences

Subject	M^1	M^2	M^3
⋮			
s_i	4.5	0.53	1.6
⋮			

At first, we run Friedman test [2] on Table 2 using $k = 3$. The null hypothesis states that there is no significant difference among the three models.

If no significant difference can be detected (i.e., $p > 0.05$), we can conclude that there is no significant difference among these three model for making relatively valid predictions.

If there does exist significant difference (i.e., $p \leq 0.05$), we can then run Wilcoxon test [3] on any pair of models to identify the best model(s) for making relatively valid predictions. For example, after we have concluded that there exists significant difference among the three models on Table 2, we then run Wilcoxon test to verify two hypotheses: $M^2 \leq M^1$, $M^2 \leq M^3$. If both result in $p \leq 0.05$, then we can make conclusion that M^2 is the best model for making relatively valid predictions.

3.2 Inter-individual Difference

The intuition here is that the model that can predict the most inter-individual difference in happiness is the model that can produce the maximum number of significant differences among the subjects.

At first, for each model ($M \in \{M^1, M^2, M^3\}$), we compare each pair of subjects (*e.g.*, s_i and s_j), to detect whether M can identify significant difference between them. To do so, we just collect all the suggested clips for each subject, and compute the difference :

$$ScoreComputed - ScoreEvaluated$$

For example, Table 3 summarizes the difference produced by M for subject s_i and s_j.

Table 3. Inter-Individual Difference

M	s_i	s_j
	⋮	⋮
	0.15	-4.4
	-0.08	5
		6.2

We run Mann-Whitney test on Table 3 to detect whether there exists significant difference between s_i and s_j (i.e., Yes if $p \leq 0.05$, otherwise No), then construct a matrix to present the difference between any pair of the subjects for model M.

[2] Friedman is a statistical measure of two-way analysis of variance by ranks, with k repeated (or correlated) measures.

[3] Wilcoxon test is a nonparametric test that can be used for 2 repeated (or correlated) measures.

$$\begin{pmatrix} & s_1 & s_2 & \dots & s_i & \dots \\ s_1 & & Y & \dots & N & \dots \\ s_2 & & & \dots & Y & \dots \\ \vdots & & & & & \\ s_i & & & \dots & & \dots \\ \vdots & & & & & \end{pmatrix}$$

Note that the Mann-Whitney test that we run is non-directional, so the matrix is symmetric. We define the inter-difference score of M as :

$$\text{Inter-Difference}(M) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} sgn(s_i, s_j)$$

where

$$sgn(s_i, s_j) = \begin{cases} 1 \text{ there eixsts significant difference between } s_i \text{ and } s_j; \\ 0 \; otherwise. \end{cases}$$

The model that has the highest Inter-Difference Score will be the best model to predict inter-individual differences.

3.3 Precision

We follow the process that described in Section 3.1, the only difference here is that we use all the evaluations, not only these $ScoreComputed = 10/1$. The model that achieves the highest precision will be the model that produce the least difference between $ScoreComputed$ and $ScoreEvaluated$.

Addressing Problems in the First Adaptive System Evaluation Challenge

David N. Chin

University of Hawaii
Department of Information & Computer Sciences
1680 East West Rd, POST 317
Honolulu, HI 96822 USA
chin@hawaii.edu

Abstract. The adaptive evaluation challenge system has several problems with assumptions including ignoring other emotions, ignoring other influences on happiness and assuming ratings of 5.5 have neutral happiness impact. The recommended evaluation combines established affective response to music surveys and heart rate monitoring to measure happiness. First, ratings are recalibrated by asking users whether music clips affect their emotions positively, neutrally, or negatively. Clips that affect emotions other than happiness are filtered out and a pilot study is used to determine how many positive/negative clips are needed to maximize/minimize Happiness. Finally a custom series of clips are designed for each user in the study following standard DJ music selection practices. The measured/reported happiness of the users is compared to the proposed Happiness models to determine the best model. Ethics of the study are also discussed.

1 Introduction

The First Adaptive System Evaluation Challenge presents an evaluation challenge adaptive system (ECAS for short) that adaptively selects a sequence of music clips to play to a group of users to maximize the overall Happiness of the group. It uses precompiled ratings for each music clip by each user for selection and hypothesizes that any particular user's Happiness can be modeled as positively impacted by higher rated music clips and negatively by lower rated music clips. The goals of the adaptive system evaluation challenge are to determine which of three models best predict Happiness with the models varying in assumptions including whether Happiness is bounded and whether current mood affects the user's reaction to a music clip.

2 Problems

There are several problems with the assumptions of the ECAS. First, the ECAS assumes that users will feel happy when listening to music that they rate highly. In

fact, music engenders a wide range of emotional responses in listeners other than happiness, include opposing emotions such as sadness, dislike, anger, hopelessness, fear, remorse, fears confirmed, shame, and reproach [1]. Also, emotional response to music is very individualistic and can be colored by the person's idiosyncratic associations with the music. For example, I feel very sad whenever I hear what most people would consider a happy song, Israel "Iz" Kamakawiwo'ole's very beautiful "Somewhere over the Rainbow, What a Wonderful World" because of the singer's unfortunate death from overweight shortly after the song's debut (a tragic waste of a wonderful talent from a treatable eating disorder). Others might rate a happy music clip highly, yet feel sad/angry because the song is associated with a former girl/boyfriend, deceased loved one or a negative event.

Another problem with the ECAS is the assumption that Happiness is based purely on current Happiness and the Impact of the music clips. In fact, response depends on many other factors. For example, if a music clip is associated with the user's spouse, then response to the clip may depend on the current status of the relationship. Even the quality of the music playback system may annoy or even enrage some users who consider the poor sound quality an insult to their favorite music. For the ECAS, group dynamics, which change constantly, can influence the response of users to the music clips. Also, previous music clips influence response. For example, a well-liked slow-paced clip may engender negative responses if users are bored by a long series of slow clips. Likewise, a fast-paced clip may be viewed as irritating if users are over simulated by a series of previous fast-paced clips. This is why classical symphonies typically have movements that vary in tempo and mood and why DJs work in sets of 2-6 fast songs of similar genre and vary the music type to suit the audience [2].

Also, the ECAS assumes that a 5.5 music clip rating (halfway in the 1-10 ratings) is neutral in happiness for all users. In fact users are known to vary in how they bunch ratings: some tend to rate very few items low in the scale whereas others tend to rate very few items high in the scale *even when told where neutral is in the scale.* As a result, clips with a 6 or 7 rating for one user may actually have negative impacts on Happiness while clips with a 4 or 5 rating may have positive impacts on another user.

3 Measurement

A major problem is how to measure happiness. Standard questionnaires can be used. It is best to use previously validated questionnaires specifically designed for music response such as Asmus' 9-Affective Dimensions (9-AD) [3] test or Bartel's CART-M test (Cognitive-Affective Response Test) [4]. It may be possible to use shortened forms of these questionnaires containing only those questions related to happiness. Another possibility is to use physiological measurements such as heart rate, which has been found in some studies to increase due to emotional response to music [5]. Of course there is still the matter of whether the emotional response is happiness or some other emotion and whether the emotional response is positive or negative. Nevertheless, a physiological measure provides a different perspective than

questionnaires, which suffer from unreliable introspection and differences in scale among users (i.e., is one user's happiness +2 equivalent to another user's +2?). Physiological measures do not suffer from introspection problems and are easily normalized and many such as heart rate are easily measured with minimal and inexpensive equipment. Using only physiological measures would *not* be recommended. However when all measures agree, then that gives extra confidence in the results and when measures disagree, the results can be labeled as suspect. The advantage of using both types measures is that each covers the weaknesses of the other.

4 Evaluation

This evaluation addresses the Adaptive Challenge evaluation goals, namely determining which of three different Happiness models as represented by three parameterized equations best predicts Happiness for a variety of users.

4.1 Recalibration

First recalibrate the ratings to users' actual neutral points by asking users to re-rate the music clips as to whether hearing the clip would *usually* make the user more happy, less happy, or neutral. The questionnaire should also ask about other emotions. The results should be used to filter out those clips that engender emotions other than happiness in the user, because such music clips will confound the Happiness measurement. It may be useful to add a multi-point scale to happiness (and the other emotions) that can be used to infer the relationship between ECAS ratings and happiness (is it linear?). However this would lengthen the recalibration considerably and it is unclear whether users' self-ratings of happiness correlate with their actual responses. By eliminating music clips that affect other emotions, any changes in heart rate can be assumed to be due to positive or negative Happiness. The direction can be assumed to be the same as the user originally rated the music clip. For example a larger increase in heart rate to a positively rated music clip is assumed to have a higher Happiness rating and a larger increase in heart rate to a negatively rated music clip is assumed to have a lower Happiness rating.

The recalibration should follow standard experimental practices with a rest period between music clips that is long enough for the user's emotions to settle back to a neutral base. A pilot study should be done to determine the optimal length of the rest period, which should be long enough to allow settling for the strongest music clips. Here, the physiological measures are very helpful. Since heart rate lags the emotional stimulation, once heart rate settles down, then one can safely assume that the user's emotional state has also settled.

4.2 Pilot Study

Now we are ready to test the Happiness equations. Rather than attempt to study multiple users in a group where individuals can influence each other with respect to their enjoyment of a music clip, the study should evaluate a single individual at a time. To maintain emotional detachment, the study should be carried out in a neutral room with no distractions (e.g. posters on the wall, windows, window savers on nearby computers, reading material on a desk, experimenters walking around, etc.). Because the purpose of the ECAS is to select a sequence of music clips that are played continuously one after another, the same format should be used for the experiment. The user should fill out an emotion questionnaire that is shorted to only ask about Happiness at the very start and end of each music sequence and periodically during the middle of the music sequence. The optimal period between questionnaires can be determined in the pilot study by giving very frequent questionnaires and noting how long it takes between significant changes in the questionnaire answers. Heart rate should be measured continuously.

The order of music clips must be carefully designed for each user. It is probably best to follow standard DJ practices [2] to group related music clips and vary the tempos. A pilot study should be performed to determine how many positive music clips of what Ratings in a row tend to max out Happiness (assumption A5) and likewise for negative clips. The pilot study should have at least several participants so as to determine the variability among users. A random selection of music clips probably will not work well to move Happiness to the extremes possible with just music, which will best distinguish Happiness equation 3 from 1 and 2, so a series of clips should be selected to get to those extremes for each user.

4.3 Full study

The full study will present a small number of users (one at a time) with personalized sequences of music clips designed to maximize and minimize their Happiness as well as randomly selected sequences. Users should be presented with long enough sequences to achieve the max/min as determined by the pilot study. The selection of music clips should be based on the individual's recalibrated ratings and follow standard DJ practices [2] even though the actual ECAS may not be able to do so because the purpose of this study is not to test the ECAS in situ, but to gather data to determine which (if any) of the Happiness equations is correct. Users should be continuously monitored for heart rate and should be asked to fill out questionnaires about their Happiness periodically throughout the session.

4.4 Analysis

The first step in analysis is to verify that the change in Happiness before and after hearing a music clip is correlated with the Rating of the clip. The Pearson chi-square test can be used to compute the correlation between the categorical variables of delta Happiness (as derived from the surveys and heart rate) and Impact. Multiple users

should be tested to determine the variability among users. Next the three different Happiness equations can be tested to see which equation fits best with the data after optimal selection of the equation parameters based on the data. A simple F test can be used to determine fit. It may be that none of the equations really fit the data, in which case the assumptions may need to be reconsidered.

4.5 Caveats

Note that this experiment does not really test whether the ECAS will work well as advertised. First, the experiment only looks at one person at a time, thus ignoring group dynamics, which adds another layer of confusion to the Happiness equations. Second, this experiment only deals with the single emotion of Happiness and music that engenders other emotions are not modeled at all. Third, the experiment does not actually use the adaptive music-clip selection algorithm. Further experiments would be needed to determine whether the ECAS will work as advertised as opposed to the stated goals of the Adaptive System Evaluation Challenge, which all relate to evaluating the Happiness models and not to evaluating the actual effectiveness of the adaptation. Fourth, there are critical problems in the database of the ECAS that really should be addressed before trying to evaluate it. For example, the tempo of the music clips, an essential element for DJ selection [2] is missing from the database. Also, the context of the group is completely ignored. For example, if the group is at a dance party versus a funeral, a completely different selection of music would make the group happy. These problems should probably be addressed before trying to deploy or evaluate the full ECAS.

5 Ethics

Because asking users to listen to music that is suspected to cause negative happiness in the user is akin to torture (e.g., General Manuel Noriega was blasted with rock music to persuade him to leave his sanctuary), there are sensitive ethical questions about this study. At worse, this study will have to avoid negative Impact clips and use only positive and neutral clips.

References

1. Vink, A.: Music and Emotion. Living apart together: a relationship between music psychology and music therapy. In: Nordic Journal of Music Therapy 10(2) (2001) 144–158
2. DJU: Music Selection 101. At: http://dju.prodj.com/courses/music/c2.shtml
3. Asmus, E. P.: The development of a multidimensional instrument for the measurement of affective responses to music. In: Psychology of Music 13 (1985). 19-30
4. Bartel, L. R.: The Development of the Cognitive Affective Response Test. In: Psychomusicology 11 (1992) 15-26
5. Dainow, E.: Physical Effects and Motor Responses to Music. In: Journal of Research in Music Education 25(3) (1977) 211-221