

The Benefits of Layered Evaluation of Adaptive Applications and Services

Peter Brusilovsky¹, Charalampos Karagiannidis² and Demetrios Sampson²

¹ School of Information Sciences, University of Pittsburgh
135 North Bellefield Avenue, Pittsburgh, PA 15260
Tel: +1-412-624-9404, Fax: +1-412-624-2788
peterb@mail.sis.pitt.edu

² Informatics and Telematics Institute (I.T.I.)
Centre for Research and Technology - Hellas (CE.R.T.H.)
1, Kyvernidou Street, Thessaloniki, GR-54639 Greece
Tel: +30-31-868324, 868785, 868580, int. 105, Tel: +30-31-868324, 868785, 868580, int. 213
karagian@iti.gr, sampson@iti.gr

Abstract. In this paper we present an empirical study of the InterBook system to demonstrate the benefits of the *layered evaluation* framework, where the success of adaptation is addressed at two distinct layers: (i) interaction assessment, and (ii) adaptation decision making.

1 Introduction to Layered Evaluation

Adaptivity, i.e. the automatic run-time, or use-time adaptation, lying at the heart of adaptive and intelligent applications and services, can be characterized by (the interaction of) two main distinct high-level phases, namely *interaction assessment* and *adaptation decision making*, as shown in Figure 1 [1].

In the interaction assessment phase, the aim is to reach high-level conclusions concerning the aspects of user-computer interaction that are considered significant for the particular application (i.e. it can be called the *Situation awareness process*). For example, assessment may detect that the user is unable to initiate and/or complete a task; the user is disoriented, and exhibits a high error rate; or, in the case of an educational application, that the user has not understood a particular concept; etc. Assessment is usually based on *low-level* information that is provided through a monitoring mechanism, including, for example, keystrokes, task initiation and completion, answers to quizzes, etc.

In the adaptation decision making phase, on the other hand, specific adaptations are selected, based on the results of the assessment phase, in order to *improve* selected aspects of interaction. Adaptation decisions may, for example, result in the presentation of a pop-up message helping the user complete a task; the re-structuring of the hyperspace helping the user navigate in it; or the provision of additional explanation for a specific concept, in the case of an educational application; etc.

The current evaluation practice does not take into account the different aspects of adaptation presented above, but rather attempts to evaluate adaptation as a whole.

That is, one should build the whole adaptive application (i.e. including the interaction assessment and adaptation decision making components), and then evaluate separately: (i) the whole adaptive application, and (ii) its non-adaptive part (i.e. the application without the interaction assessment and adaptation decision making components (that are already implemented)). When adaptation is found to be successful, one can reasonably conclude that both phases have been successful (except for the unlikely situation when adaptation is successful and both assessment and decision making phases are not — the minus times minus equals positive effect!). When adaptation is found to be unsuccessful, however, it is not evident whether one, or both of the above phases has been unsuccessful. It could be the case that the adaptation decisions are reasonable, but they are based on incorrect assessment results; or that the assessment results are correct, but the adaptation decisions are not meaningful.

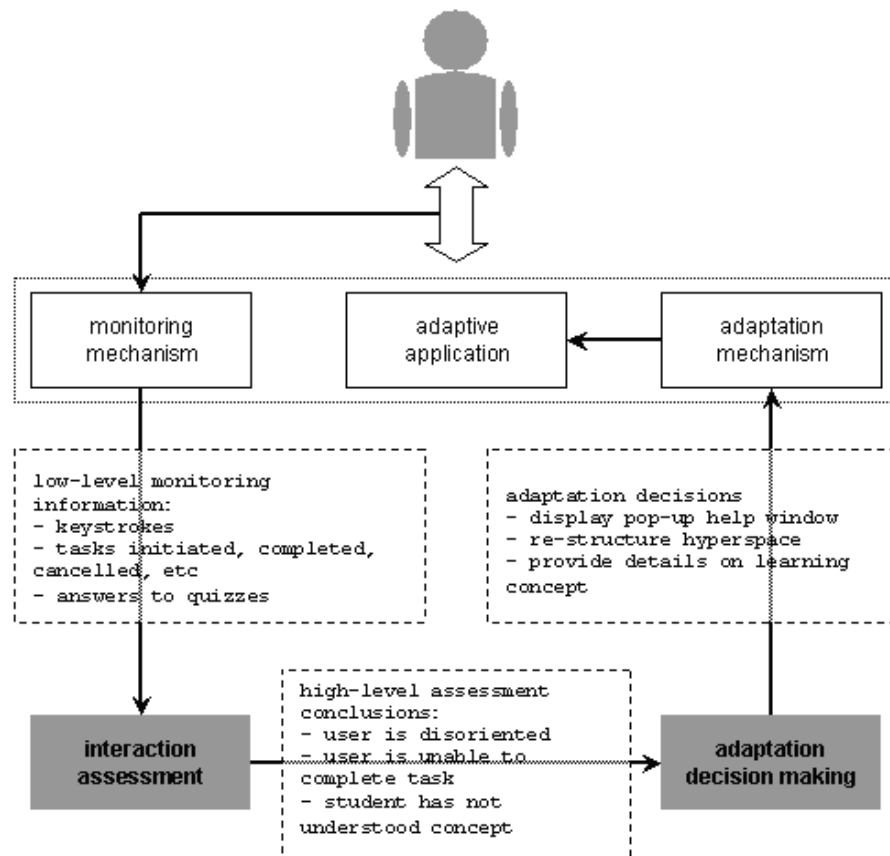


Fig. 1. Adaptation Decomposed. We consider that adaptation is characterized by (the interaction of) two high-level phases: *interaction assessment* and *adaptation decision making*

In this paper, we advocate the *layered evaluation framework* [2], where the success of adaptation is decomposed into, and evaluated at, different layers, reflecting the main phases of adaptation shown in Figure 1.

- In the interaction assessment layer, only the assessment phase is being evaluated. That is, the question here can be stated as: *Are the conclusions drawn by the system concerning the characteristics of the user-computer interaction valid?* or *Are the user's characteristics being successfully detected by the system and stored in the user model?*
- In this the adaptation decision making layer, only the adaptation decision making is being evaluated. That is, the question here can be stated as: *Are the adaptation decisions valid and meaningful, for selected assessment results?*

In the following section, we provide an example of layered evaluation of the InterBook system [3], to demonstrate the benefits of this approach. In the past we have run several studies where we explored the value of adaptive navigation support in InterBook [4]. The goal of this contribution is to revisit our evaluation in the context of the layered evaluation framework.

2 Layered Evaluation of Adaptive Link Annotation in InterBook

2.1 Adaptive Link Annotation in InterBook Decomposed

According to the layered evaluation framework, the process of adaptive annotation in InterBook can be split into two parts. The goal of the first part is to assess the user knowledge of the concepts and the educational states of book pages. The core part of the user model in InterBook represents levels of user's knowledge of every domain concept. The system distinguishes several levels of user knowledge of the concept. The first two levels that are important for adaptive annotation mechanism are *unknown* and *known*. The source data for the assessment process are gathered by watching the user browsing activity. The assessment mechanism assumes that user reads all pages that are observed for some reasonable time (this is a simplification because we do not know what the user is doing while the page is observed). When a ready-to-be-learned page is read, all unknown concepts from its outcome become known.

The concept knowledge is the key to the assessment of the educational status of book pages. A page that has at least one unknown prerequisite is considered *not ready to be learned*. A page that has no unknown prerequisites and at least one unknown outcome concept is considered *ready and recommended*. A page that has neither unknown outcomes nor unknown prerequisites is judged as *nothing new*. Note that a page can move to nothing-new status even if it has never been visited: the user can learn about its outcome concepts elsewhere.

The results of the assessment process, i.e. knowledge of concepts and educational states of book pages, are transferred to the second part of the adaptation process — the adaptation decision making. This process in InterBook aims to provide the least intrusive adaptation, by simply choosing different icons for links to the nodes with different status. As we have mentioned, a link to a *Nothing new* book page is

annotated with a white bullet, a link to a ready and recommended book page is annotated with a green bullet, and a link to a not ready to be learned book page is annotated with a red bullet. For the links to glossary pages, a link to an unknown concept is not annotated and a link to a known concept is annotated with a small checkmark. Larger checkmarks are used to annotate the links concept pages with knowledge state *“Better than known”*. This part is not discussed here in detail, since it was not a part of an experiment described later.

2.2 InterBook Evaluation Study Revisited

To support our case for the layered evaluation of adaptive systems, we are reconstructing here an earlier study of adaptive annotation in the InterBook system. We think that this study can clearly demonstrate the need and the benefits of layered evaluation. The study itself is reported in details in [4]. Here we consider this study with a different prospect, in the light of the layered evaluation approach. The goal of this experiment was to assess what impact, if any, user model-based link annotation would have on students’ learning and on their paths through the learning space. Contrary to our expectations, the study *brought no significant results*. In particular, while students seem to understand and like adaptive navigation support (ANS) features, it didn’t influence their performance on tests. A two-sample T-test showed that there was no significant difference at the 0.05 level in the test means for those with ANS and those without ANS.

An analysis of the audit trails revealed at least one explanation of this result. Most of all navigation steps were made with Continue and Back buttons, or with hot words in text which were *not* annotated in the experimental version of *InterBook*. Only a few of all clicks were made on annotatable links (i.e. links that were annotated in the ANS version). In a situation where adaptive annotations were used only in 1/10 of all navigation steps, it is hardly surprising that ANS has provided no significant difference.

We had a situation where the adaptation process as a whole has failed to achieve its goals. The question that is usually explored by the experimenters in such a situation is: *are there still any differences between adaptive and non-adaptive versions?* It is exactly the question we have tried to answer in the original report of the study [4]. However, from the prospect of a layered evaluation presented in this paper, different questions need to be considered such as: *Why does the adaptation not work? Was it the interaction assessment part where the system has performed poorly? Was it the adaptation decision making part where the adaptation decisions weren’t properly made? Or, maybe the system was far from perfection in both layers of adaptation?* A layered evaluation approach could provide answers to these questions and guidance for further work.

In our case we were not planning a layered evaluation in advance, however we made a wise decision to collect lots of data about student interaction (more than we were expected to use). In this situation it became possible to perform a limited layered evaluation *“post-factum”* by re-processing the data. The goal of our post-evaluation was to check how good is the assessment part of the system: i.e., how well it can predict the user knowledge level and the individual educational states of electronic

pages. We have decided to check whether the educational status of a page (i.e. ready, not ready, or nothing new) predicted by the system has any connection with their performance on the page. The parameter we have checked is the average time spent by a user on pages of each of the three possible types (these data could be obtained by re-processing InterBook log files. It turned out that the average time students spent on "nothing new", "not ready" and "ready" pages are very different. The average time spent on a not-ready page is much larger than the time for a ready page, which is close to the average time per hit. The average time spent on a "nothing new" page is much less than average time per hit (Figure 2). Note, that in general an average "time on page" may not be a good indicator of page "difficulty" since the pages may simply be of very different size. It was, however, a good indicator in our case since the material was "well-chunked" and all pages were very homogeneous in size.

This data shows that the interaction assessment process, which predicts an educational status of electronic pages, works quite well. A page classified as "nothing new" can be read much faster (or just passed over) because it has no new information, and a page classified as "not ready" is the most hard to understand because some background may be missed. This data gives us almost as much the students can give themselves by telling us "Yes, I do not think that I am ready to learn this page" or "Oh, I already know that!"

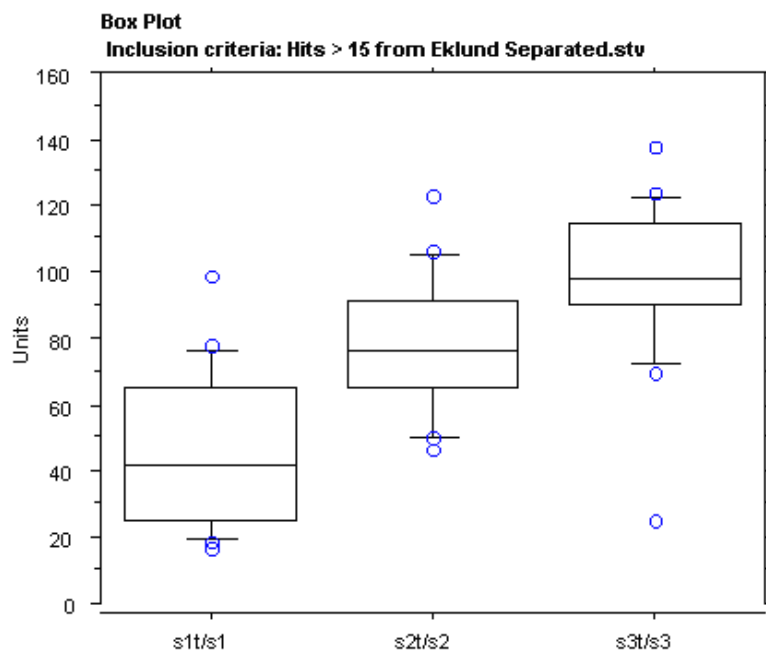


Fig. 2. Average time per page (sec.) for the case when the student navigate to a page using a non-annotated link. s1t/s1 is an average time for "nothing new" pages, s2t/s2 is an average time for "ready but new" pages, s3t/s3 is an average time for "not ready " pages.

It is important to note that in about 90% of cases the students navigated to a page of

learning material with a non-annotated link and thus without any warning about the page state. If the students can always use the adaptive annotations, we would not be able to measure the real value of the page state, since the very presence of adaptivity may change the students' behavior. For the student who navigates to a page using an adaptively annotated link, the time spent on the page is a function of both the *page status* and the influence of *being notified about that status*. For example, students who were warned that a visibly complicated page is not ready to be learned might leave this page without careful reading. In some sense, we were fortunate not to have back and continue links annotated, since it enabled us to get reliable evidence that the interaction assessment module of the system works reasonably well.

In a situation where the assessment part works, but the overall adaptation results are not satisfying, *the layered evaluation approach suggests that the problem is with the adaptation decision making*. That is, the decision to use adaptive link annotation to show page status was simply not an appropriate method of adaptation in the given context for the given student population. This conclusion was not made in our original study report, because we were not being guided by the layered evaluation approach at that time. As a result, we failed to do what we really had to do: try another method of adaptation for the given category of students, or find a category of students who can benefit from the existing adaptive annotation. Instead, we decided to blame the missing annotations of *next* and *previous* links and to repeat the experiment with some small modifications (such as having all links properly annotated). Needless to say that our new experiment hasn't brought any significant results either. We think that it is a good example of how important is to have a right approach to evaluation and a good understanding what is really being evaluated.

While we have failed to make a correct conclusion when originally processing the data of our experiment, the work of other researchers provides some good evidence that this conclusion is, indeed, correct. An evaluation of the ELM-ART system [5], has shown that adaptive link annotation is of use for students who have some previous experience that is relevant to the subject being learned from an adaptive hypermedia system. In turn, novices benefit more from direct guidance with the adaptive *next* link. Similarly, Specht and Kobsa have shown that adaptive link annotation, a technology with little guidance and restriction, is a good way to help students with high previous knowledge on the subject [6]. In turn, learners with low previous knowledge seem to profit from more guided and restrictive methods such as enabling/disabling links.

In our case, teacher education students in their majority had neither knowledge of ClarisWorks database, nor any experience that could be relevant to this subject. It is also known from the educational hypertext research that this kind of *total novice* hypertext users tend to follow a sequential way of navigation, and ignore links that can get them out of the linear path. So, indeed, adaptive link annotation, the technology that worked very well for computer science students with some good background knowledge in the ISIS-Tutor experiment [7], was not a good choice for teacher education students with little or no knowledge of the subject and background knowledge.

3 Discussion and Conclusions

The layered evaluation framework has been introduced in [2]. It is based on the separation of adaptation into the interaction assessment and adaptation decision making phases, which underlies several general models for adaptive systems which have been proposed in the literature. For example, Jameson's "General schema for processing in a user-adaptive system" [8], includes an "Upward inference" phase, where "generally relevant properties" are identified based on (raw) input, and concludes with a phase where decisions on relevant properties for adaptation are made; also, Benyon's overall architecture for intelligent interface technology [9] includes an inference and an adaptation mechanism, which are related to the above processes.

This paper has presented a *case* for layered evaluation by reconstructing one of our earlier studies that could greatly benefit from the suggested approach. We believe that the example presented in this paper demonstrates the benefits of layered evaluation. We really wish we were using the layered approach with full understanding during planning and performing our original study. It could have lead us to the correct conclusion right away. It could have pushed us to collect some more data about students (such as Web experience, level of education, etc) and possibly isolate a subgroup for which the selected method of adaptation may work. For this study, we could only reconstruct it and re-interpret its results with the layered approach at hand. We intend to use the layered approach for our future studies and we hope that the case described in this paper will convince other user modeling researchers to also adopt this approach.

Acknowledgements

Part of the work presented in this paper was partially financially supported by the European Commission under the IST No 12503 Project "KOD — Knowledge on Demand" (<http://www.kodweb.org>, <http://kod.itl.gr>) through the Information Society Technologies Programme (IST).

References

1. Karagiannidis C., Koumpis A., Stephanidis C. & Georgiou A.C., "Employing Queuing Modeling in Intelligent Multimedia User Interfaces" *International Journal of Human-Computer Interaction*, 10 (4) (1998) 297-326
2. Karagiannidis, C., Sampson, D.: Layered Evaluation of Adaptive Applications and Services. In: Brusilovsky, P., Stock, O., Strapparava, C. (eds.): *Adaptive Hypermedia and Adaptive Web-Based Systems*. Lecture Notes in Computer Science, Vol. 1892. Springer-Verlag, Berlin Heidelberg New York (2000) 343-346
3. Brusilovsky, P., Eklund, J., Schwarz, E.: Web-based education for all: A tool for developing adaptive courseware. In: Proc. 7th International World Wide Web Conference (1998)

4. Brusilovsky, P., Eklund, J.: 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) (1998)
5. Weber, G., Specht, M.: User modeling and adaptive navigation support in WWW-based tutoring systems. In: *Proc. 6th International Conference on User Modeling (1997)* 289-300
6. Specht, M., Kobsa, A.: Interaction of domain expertise and interface design in adaptive educational hypermedia. In: *2nd Workshop on Adaptive Systems and User Modeling on the World Wide Web (1999)*
7. Brusilovsky, P., Pesin, L.: Adaptive navigation support in educational hypermedia: An evaluation of the ISIS-Tutor. *Journal of Computing and Information Technology*, 6 (1) (1998)
8. Jameson, A.: User-Adaptive Systems: An Integrative Overview. Tutorial in *International Conference on Intelligent User Interfaces (2001)*
9. Benyon, D.: Editorial. *Interacting with Computers, Special Issue on Intelligent Interface Technology*. 12 (2000) 315-322