Layered evaluation of adaptive learning systems

Peter Brusilovsky

School of Information Sciences, University of Pittsburgh, USA

Charalampos Karagiannidis

Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Department of Cultural Technology and Communication, Aegean University, Mytilini, Greece

Demetrios Sampson*

Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Department of Technology Education and Digital Systems, University of Piraeus, 150, Androutsou Street, Piraeus, GR 18534, Greece
E-mail: Sampson@unipi.gr E-mail: Sampson@iti.gr
*Corresponding author

Abstract: This paper suggests an alternative to the traditional 'as a whole' approach of evaluating adaptive learning systems (ALS), and adaptive systems, in general. We argue that the commonly recognised models of adaptive systems can be used as a basis for a *layered evaluation* that offers certain benefits to the developers of ALS. Therefore, we propose the *layered evaluation* framework, where the success of adaptation is addressed at two distinct layers:

- user modelling
- adaptation decision making.

We outline how layered evaluation can improve the current evaluation practice of ALS. To build a stronger case for a layered evaluation we re-visit the evaluation of the InterBook where the layered approach can provide a difference and provide an example of its use in KOD learning system.

Keywords: adaptive learning systems; adaptive hypermedia; user modelling; evaluation; layered evaluation.

Reference to this paper should be made as follows: Brusilovsky, P., Karagiannidis, C. and Sampson, D. (2004) 'Layered evaluation of adaptive learning systems', *Int. J. Cont. Engineering Education and Lifelong Learning*, Vol. 14, Nos. 4/5, pp.402–421.

Biographical notes: Dr. Peter Brusilovsky is an Assistant Professor at the School of Information Sciences, University of Pittsburgh, an Adjunct Research Scientist at the HCI Institute, School of Computer Science, Carnegie Mellon University, member of Intelligent Systems Program at the University of

Pittsburgh, and a Visiting Professor at the Free University of Bozen-Bolzano and National College of Ireland. He holds a PhD in Computer Science (1987) and a Diploma in Applied Mathematics (1983) from the Moscow State University. His main scientific interests are in the areas of Adaptive Hypermedia and Adaptive Web-based Systems, Intelligent Tutoring Systems and Shells, Student and User Modelling, Teaching programming to novices, Psychology of Programming, Human-Computer Interaction, and Artificial Intelligence. He is the co-author of more than 120 publications in scientific articles, journals and conferences.

Dr. Charalampos Karagiannidis is an Associate Researcher at the Advanced eServices for the Knowledge Society Research Unit (ASK) at the Informatics and Telematics Institute (ITI) of the Center of Research and Technology Hellas (CERTH), and an Assistant Professor on e-Learning at the Department of Cultural Technology and Communication of the University of Aegean, Greece. He holds a PhD in Electronic Engineering (1998) from the University of Kent at Canterbury, UK, an MSc in Information Technology (1992) from the University of London, UK, and a BSc in Mathematics (1991) from the Aristotelian University of Thessaloniki, Greece. His main scientific interests are in the areas of learning technologies and human-computer interaction, mainly related to personalised learning and personalised interaction. He is the co-author of more than 50 publications in scientific articles, journals and conferences, with at least 80 citations.

Dr. Demetrios G. Sampson is the Head of the Advanced eServices for the Knowledge Society Research Unit (ASK) at the Informatics and Telematics Institute (ITI) of the Center of Research and Technology Hellas (CERTH) and an Assistant Professor on e-Learning at the Department of Technology Education and Digital Systems of the University of Piraeus. He holds a Diploma in Electrical Engineering (1989) from Demokritus University of Thrace and PhD in Electronic System Engineering (1995) from University of Essex, UK. His main scientific interests are in the areas of Technology Enhanced Learning, Semantic and Context-based Knowledge Systems and Web Engineering. He is the co-author of more than 130 publications in scientific articles, journals and conferences with at least 92 citations.

1 Introduction and background

Adaptive web (Brusilovsky and Maybury, 2002) has attracted considerable attention due to its potential to provide personalised applications and services for the citizens of the knowledge society. For example, an online newspaper will deliver news that are most relevant to the user reading interests (Billsus et al., 2002), while a mobile tourist guide will present the information that is adapted to the user interest and location (Cheverst et al., 2002).

Adaptive learning systems (ALS) (Sampson et al., 2002c; Brusilovsky and Peylo, 2003) constitute one of the main areas where adaptive web technologies are used. The aim is to provide personalised applications and services, which overcome the 'no significant difference' effect that has been reported in the educational technology literature (Russell, 1999), making traditional web-based educational systems useful to individual learners. An adaptive web-based educational system typically collects data about the student working with the system, creates a student model (Brusilovsky, 1999)

and uses it to adapt the presentation of the course material, navigation through it, and its annotation, to the student. Student models can also be used to form a matching group of students for different kinds of collaboration, as well as to identify the students progressing too slow or too fast and act accordingly (e.g., show additional explanations, or present more advanced material) (Devedzic, 2003; Sampson et al., 2002c).

Adaptive hypermedia along with some other research fields has contributed significantly towards establishing a technological ground for adaptive web. Over the last ten years, the field of adaptive hypermedia has accumulated a large set of already reported technologies that can be used for building a variety of adaptive web systems (Brusilovsky, 2001; Kobsa et al., 2001). Given the large set of existing techniques and a practical orientation of most adaptive web projects, evaluation of adaptive systems and techniques is becoming more important than inventing new techniques with questionable benefits. To guide further research and practical work on adaptive learning systems, it is becoming essentially important for each new system or technology to be properly evaluated. It is essential to understand whether a particular technique works as expected, in which contexts or application areas it works, and what is the scale or benefits it can produce in exchange for its adaptivity complexity.

Evaluation is widely considered as an important and challenging research issue in the area of ALS, and adaptive systems, in general. In fact, the lack of evaluation data, as well as the difficulty in their generalisation, when available, and the resulting difficulty in the re-use of successful design practices, constitutes, among others, one of the main barriers for ALS to become mainstream technology (Hook, 1997).

A recent study of evaluation practice in the field of user (student) modelling (Chin, 2001) identified and reviewed thirty two papers that have addressed, in some way, empirical evaluation for adaptive hypermedia/hypertext, student modelling, plan recognition, mixed-initiative interaction and user interfaces/help systems. As it is shown in Table 1 that summarises the results of the study, the approaches used to evaluate adaptive systems depend on the type of the system being evaluated. Evaluation criteria may include both quantitative and qualitative measures, such as: task completion time, number of visited nodes, accuracy of tasks, how well the user remembers the structure of the information space, (Höök and Svensson, 1998); user's indication of utility, ease of use, naturalness, etc. (Maybury and Wahlster, 1998); number of navigation steps, number of repetitions of previously studied concepts, number of transitions from one concept to another concept, or from an index to a concept (Eklund and Brusilovsky, 1998).

Table 1 Traditional ways of evaluation of adaptive systems

Adaptive system or technology	*Traditional evaluation approach*
Adaptive hypermedia/hypertext	Measure of recall and precision, similarity/relevance metrics, comparison of the system with and without adaptation
Plan recognition	Percentage of actual plans recognised in a test corpus of plans; frequency and accuracy of predicted actions, comparison with an expert human recogniser
Mixed-initiative interaction	Comparing system responses choices with human choices, efficiency of the dialogue needed to achieve an information transfer task with either human-human dialogues or with theoretically minimum dialogues
User interfaces/help system	Subjective user satisfaction, task completion speed, error rate – quality of task achievement

At the same time, all of these approaches are similar in one aspect – they tend to evaluate an adaptive system 'as a whole', focusing on an 'end value' delivered by the system such as the overall user's performance or the user's satisfaction. Furthermore, the current practice in the evaluation of adaptive applications usually adopts a 'with or without' approach, where experiments are conducted between two groups of users, one working with the adaptive application, the other with its 'non-adaptive version' – assuming, of course, that an adaptive application can be easily decomposed into its 'adaptive' and 'non-adaptive' components (Höök, 2000).

Evaluating a system as a whole can be acceptable in the field where no acceptable component model of a system can be identified. However, it is not the case for adaptive systems. A number of useful models of adaptive systems have been suggested by the leading researchers in the field. These models recognise that adaptive systems are similar to each other at some level of consideration. For example, one of the first and most well-known models for adaptation (Benyon and Murray, 1993) recognises that adaptive applications include a user model, a domain model and an interaction model, which, in turn, may involve a number of additional models/components. In general, all models of adaptive systems acknowledge that the development of adaptive applications involves several sub-components, which are necessary for supporting the complex representation and inference underlying adaptive behaviour.

While these models have contributed to better understanding of the field, they have failed so far to influence the evaluation practice. As it was clearly shown by the cited study (Chin, 2001), evaluation practice does not take into account the different phases, processes, and components of adaptive behaviour. This paper suggests that the commonly recognised models could provide a better service for the field of adaptive systems than just serving as a reference point – they can guide a *layered evaluation* process that we consider as an alternative to the traditional 'as a whole' approach. We argue that the traditional evaluation approach become stumbling point on the way to developing useful adaptation technologies. While it can be used to report a success, it is not able to guide the authors of an adaptive system in the development process. First, evaluating a system as a whole requires building the whole system before it can ever be evaluated. It shifts major evaluation to the later stages of system development where it can provide only a limited influence on the design process. Secondly, it does not provide useful information for the improvement of a system in case that the performance of the adaptive system was not found satisfactory. Since adaptive behaviour is evaluated as a whole, the reasons behind unsatisfactory adaptive behaviour are not evident, and the ways to improve the system are not clear. Finally, traditional evaluation provides no feedback about performance of different system components, thus successful design practices cannot be easily re-used across different applications and services.

As a solution to the listed problems, this paper presents a model-based evaluation approach that we call the *layered evaluation* framework. In this framework the success of adaptation is addressed at two distinct layers:

- the user modelling
- adaptation decision making.

We argue that layered evaluation is a good approach for the evaluation of adaptive applications and services. It provides useful information for their improvement, and can contribute towards the generalisation and re-use of evaluation results. We discuss the

benefits of the layered evaluation framework and outline how layered evaluation can improve the current evaluation practice of adaptive applications and services. In order to demonstrate the potential benefits of the proposed framework we re-examine the evaluation of *InterBook*, an adaptive hypermedia system aiming to provide adaptive web-based textbooks (Brusilovsky et al., 1998) and demonstrate the use of the framework for evaluation of *KOD* (knowledge-on-demand), an adaptive web-based learning environment (Sampson et al., 2002c).

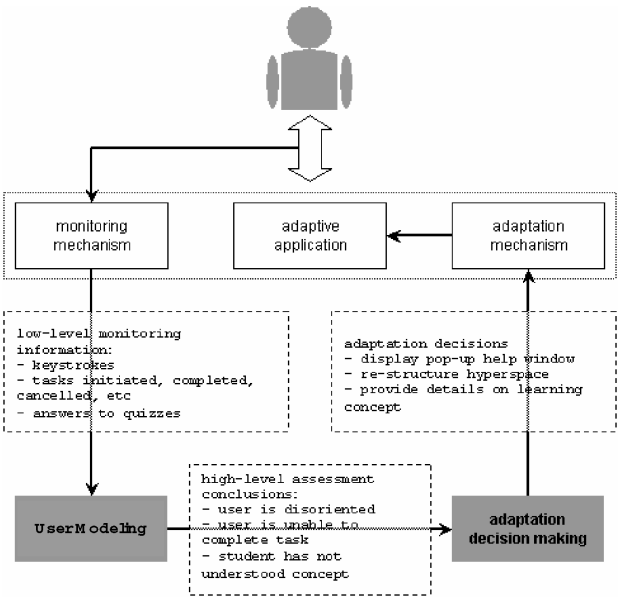
2 The model: abstract adaptation decomposition

The generic model of an adaptive system that is presented in this paper suggests recognising two main distinct high-level processes, or phases: *user modelling* and *adaptation decision making* (Figure 1). This follows a number of commonly acceptable models (Benyon and Murray, 1993; Brusilovsky, 1996; Totterdell and Rautenbach, 1990), yet it provides the minimal useful decomposition.

2.1 User modelling

The goal of the *User Modelling* phase is to reach high-level conclusions concerning the aspects of user-computer interaction that are considered significant for the particular application. For example, it 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. User modelling 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.

Figure 1 Adaptation decomposed



The user modelling (UM) process can take into account several aspects of user-computer interaction. These include the nature of the application, the tasks being performed, the educational material being presented, platform and network characteristics and so on. Nevertheless, in most existing systems, the UM process focuses entirely on long- or short-term user's characteristics. The result of the UM process are reflected in a *user model* (also called student model in the field of intelligent tutoring systems), which captures information concerning the user characteristics that are considered significant for a particular application. Adaptive hypermedia/hypertext applications, for example, usually take into account the user's goals, knowledge, background, experience and preferences (Brusilovsky, 1996).

2.2 Adaptation decision making

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

The logic of adaptation decision making is often captured into a set of *adaptation rules* that determine which adaptation constituent(s) should be selected, according to the results of the UM process. For example, in adaptive hypermedia/hypertext applications, these rules are responsible for adaptive – text and/or multimedia – presentation, and/or adaptive navigation support, including the sorting, hiding and annotation of links (Brusilovsky, 1996).

2.3 Relation between the two layers

The above processes are closely *interconnected*, since adaptation decision making takes input from the results of the interaction assessment. On the other hand, they are also *independent*, since the same user modelling outcomes may result in significantly different adaptation decisions. For example, the fact that the user is not working efficiently with an application may result in a specific adaptation, if the adaptation goal for this specific application is to speed up interaction; while, on the other hand, the same UM information could have been 'omitted' by a safety-critical application, where adaptations are initiated only when user modelling indicates a high user error rate. Nevertheless, the above mentioned processes are *both* important for the success of adaptation, in the sense that they should both be carefully designed and evaluated in order to ensure that adaptation is successful (Karagiannidis et al., 1997a, 1997b).

3 The evaluation framework: layered evaluation

As it was mentioned in Section 1, the current evaluation practice does not take into account the components of adaptation process presented in Section 2, but rather attempts to evaluate an adaptive system as a whole. When adaptation is found to be successful, one can reasonably conclude that both phases have been successful. When adaptation is found to be unsuccessful, however, it is not evident whether one, or both of the above

described phases has been unsuccessful. It could be the case that the adaptation decisions are reasonable, but they are based on poor user modelling or that the user modelling is good, but the adaptation decisions are inappropriate.

This section presents the *layered evaluation framework*, where the success of adaptation is evaluated at different layers, reflecting the main processes/phases of adaptation shown in Figure 1. We think that the proposed framework provides insight into the success of each of the phases of adaptation, thus facilitating the improvement of adaptive applications and services. It also contributes towards the generalisation and re-use of the evaluation results across different applications and services.

3.1 Layer 1: evaluation of user modelling

At this layer, only the UM process 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?".

For instance, in the case of adaptive hypermedia systems, following the classification described in (Brusilovsky, 1996), this layer addresses the following issues: Does the system detect the real user goals, as they are continuously changing? Is the user's actual knowledge of the subject being successfully captured? Are the user's interests detected by the system? Is the user's experience with respect to the hyperspace structure successfully reflected in the user model? Are the user's preferences successfully represented in the user model?

This phase can be evaluated, for example, through user tests, where experts can monitor users as they work with the system, comparing their expert opinion on the user's characteristics vs. the conclusions that are stored in the user model (Manouselis and Sampson, 2003). Additionally, the users can also themselves evaluate whether the conclusions drawn by the system at any particular instance reflect their real needs: "the system detected that my goal, at a particular instance, had been to know more about this subject; was this really the case?" This evaluation layer does *not* assume that the adaptation decision making component has already been developed. In a good spirit of modern interactive system design, it allows the developers of adaptive systems to start a solid evaluation of a system before it is fully developed.

The UM process evaluation can also provide details concerning the necessary granularity of the user model. For example, the requirements analysis phase may indicate that the user's knowledge should be classified into three categories: novice, intermediate and expert. The UM phase may indicate that this classification should be 'reduced', since the conclusions actually drawn always classify users as either being novices or experts; or, in the case that the evaluation indicates that more fine-grained conclusions can be drawn, the classification should be extended to include more levels of knowledge. This can inform the iterative design and the development phase of the adaptive system, and significantly improve them.

Given that the UM process has been evaluated separately and found satisfactory, its results can be generalised. The conclusions made by the UM process based on the low-level monitoring information can be re-used in similar contexts with different decision making modules (Manouselis and Sampson, 2003). This can facilitate the re-use of successful 'design practices', i.e. specific UM approaches.

3.2 Layer 2: evaluation of adaptation decision making

At this level, only the adaptation decision making is being evaluated. The question here can be stated as: "are the adaptation decisions valid and meaningful, for the given state of the user model?" For example, in adaptive hypermedia (Brusilovsky, 1996), one can try to evaluate "is the selected adaptive presentation technique appropriate for the given user goals?" or "does the selected adaptive navigation technique improve interaction, for specific user's interests, knowledge?"

This phase can, again, be evaluated through user testing, based on specific scenarios. For example, to evaluate a knowledge-based adaptation the user knowledge can be assessed by direct testing. To evaluate a goal-based adaptation, the user can be given a particular goal. The goal of evaluation then is to assess whether the specific adaptation is helpful given the known goal or level of knowledge. Alternatively, users and experts can evaluate whether specific adaptations contribute to the quality of interaction: "does the selected adaptation of the presentation of information improve the quality of the system, when the user is disoriented?". As in the previous case, this evaluation layer does *not* assume that the UM phase has already been developed thus allowing early evaluation.

Again, given that the decision making phase has been evaluated separately and found successful, we can generalise its results. We can argue that the design practice adopted in the particular application, as this is reflected in the adaptation logic can be re-used across similar applications, even with different UM processes (Karampiperis and Sampson, 2004).

4 Layered evaluation of adaptive link annotation in InterBook

In this section we attempt to demonstrate some benefits of layered evaluation on a practical case. From one side, we want to provide some insights on how a layered evaluation of adaptive hypermedia systems can be performed. From another side, we want to show that using the layered evaluation framework could help to interpret empirical data and guide further studies. Following the example provided by Specht and Kobsa (Paramythis et al., 2001), we re-visit and re-process the data of one of our older studies (Brusilovsky and Eklund, 1998) from the new prospect. The study under consideration attempted to evaluate adaptive annotation is InterBook, an adaptive hypermedia system and a shell targeted to the development of adaptive web-based textbooks (Brusilovsky et al., 1998). In subsection 4.1, we briefly describe InterBook's adaptive annotation in terms of the adaptation decomposition shown in Figure 1, and then we revisit our evaluation in the context of the layered evaluation framework.

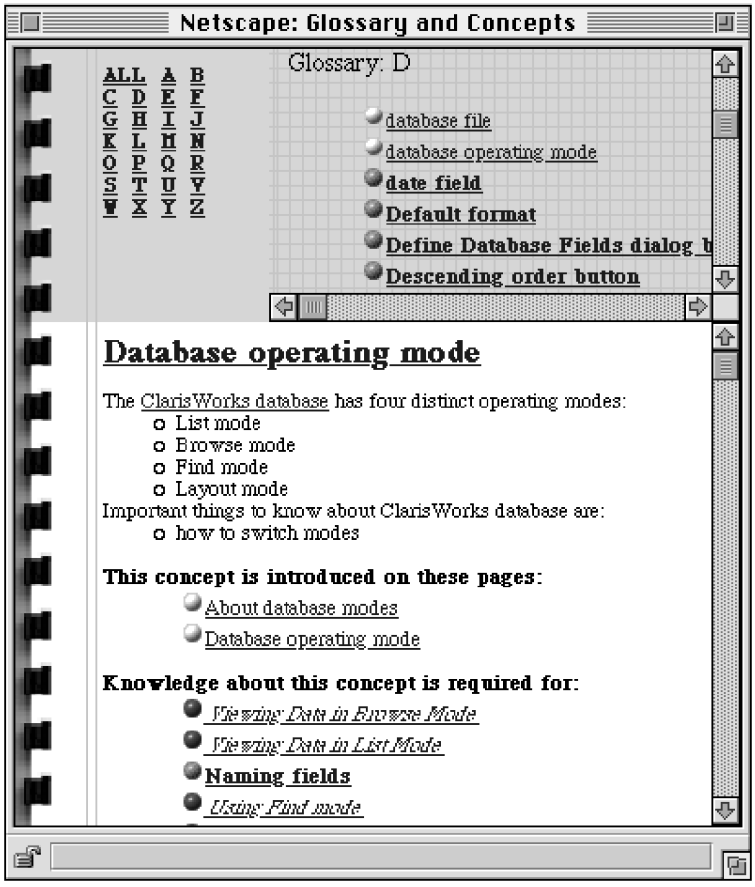
4.1 Adaptive link annotation in InterBook: the technology

Adaptive link annotation is a popular adaptation technology in the area of adaptive hypermedia systems. Its goal is to help users in selecting the most relevant links in the process of navigation. Together with other technologies, such as link sorting, it belongs to the group of adaptive navigation support technologies (Brusilovsky, 1996). The idea of adaptive annotation technology is to augment the links with annotations – some additional information that can tell the user more about the current state of the nodes behind the annotated links. These annotations are provided in the form of visual cues:

different icons (Brusilovsky et al., 1998; Passardiere and Dufresne, 1992), colours (Brusilovsky and Pesin, 1998), font sizes (Hohl et al., 1996), or font types (Brusilovsky et al., 1998). These annotations are adaptive, i.e. they depend on the current state of the user model: different users may see different annotations, and for the same user annotations may change over time reflecting the changes in the user model.

InterBook uses a concept-based approach to adaptive annotations that takes into consideration user's knowledge of the *domain concepts* that designate elementary pieces of knowledge about the domain. All concepts are made visible to the users via *the Glossary*: a description of each concept is individually accessible as a *glossary page* (Figure 2).

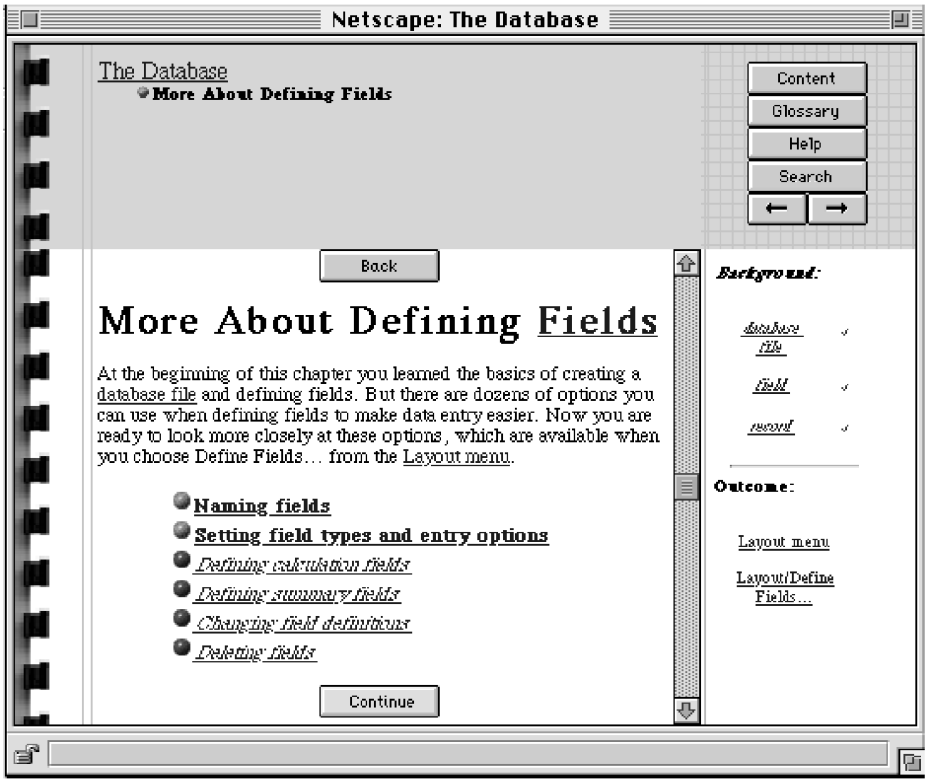
Figure 2 Glossary window of InterBook showing a glossary entry for the concept 'database operating mode'



An author of an electronic textbook can specify a list of relevant outcome and prerequisite concepts for every section of the book. A concept is listed as an outcome concept if some part of this section presents the piece of knowledge designated by the concept. A concept is listed as a prerequisite concept if a student has to know this concept to understand the content of the section (Figure 3).

InterBook visualises relationships between concepts and pages by generating links between glossary pages and textbook sections. Links are provided from each book section to the corresponding glossary pages for each involved background or outcome concept (Figure 3). Similarly, for each glossary page that describes a concept, InterBook provides links to all book pages that can be used to learn this concept or depend on the concept (Figure 2).

Figure 3 A section of an electronic textbook with related concepts in InterBook



The goal of adaptive link annotation in InterBook is to inform the user about the current educational status of all links to book and glossary pages. All links to book pages are consistently annotated with *bullets* of different colour, and font of different type. Red bullet and italic font tell the user that the page behind this link is not ready to be learned (not recommended), green bullet and bold font tell that the page is interesting and ready to be learned (recommended), while white bullet and regular font tell that the page has no new concepts. A check mark is added for already visited nodes. All links to glossary pages are annotated with *checkmarks* of different size. The size of the checkmark that annotates a link to this concept page indicates the system's estimate of the user's knowledge of the concept. Several sizes of checkmark reflect several levels of knowledge recognised by the system.

Naturally, as the student's knowledge of the subject progresses, the annotations change: more checkmarks appear near the links to glossary pages, the checkmarks grow, the bullets that were originally red become green and then white.

4.2 Adaptive link annotation in InterBook decomposed

According to the proposed 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 UM process are gathered by watching the user browsing activity. The UM mechanism assumes that user reads all pages that are observed for some reasonable time. While it looks as a simplification because we do not know what the user is doing while the page is 'observed', some recent studies (Claypool et al., 2001) demonstrate that page reading time is a reliable predictor of user interest in page content. 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 its outcome concepts elsewhere.

The results of the user modelling 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.

It is important to stress that the user modelling and adaptation decision making in InterBook are reasonably independent. The interface between these processes is the student model – a vector that stores the status value for each concept page and each book page. The UM process produces and updates this data, and the adaptation decision making process uses this data for generating the adaptation effect. It is very easy to imagine that the same kind of data is collected by a different mechanism, even one based on a different knowledge model. For example, an adaptive educational system can use quizzes to determine the user's level of knowledge for a concept more reliably, or even to ask the user to provide a self-estimation of his or her knowledge. Likewise, using the same UM results, an adaptive system could make a number of different adaptation decisions. For example, the AHA! system (De Bra and L. Calvi, 1998) hides links to pages that are not ready to be learned, and ELM-ART II (Weber and Specht, 1997) adds a new link to the current page that provides the student with a simple way to navigate to the best of the ready and recommended pages. The independence of the two processes of adaptation enables us to evaluate and change these parts independently.

4.3 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. We think that this study can clearly demonstrate the need and the benefits of layered evaluation. The study was originally reported in (Brusilovsky and Eklund, 1998). Here we consider this study from a different prospect, in the light of the layered evaluation approach.

The study involved 25 undergraduate teacher education students in an educational computing elective at the University of Technology, Sydney. The students were exposed to two chapters of a textbook about *ClarisWorks* databases and spreadsheets, and used the *InterBook* system both *with* and *without* adaptive link annotation (the version without adaptive annotation had no checkmarks and all bullets were green regardless of the link status). 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. Following our earlier experiment with the ISIS-tutor system (Brusilovsky and Pesin, 1998) we hypothesised that adaptive link annotation will help the students to build a more efficient path through the knowledge space and to achieve better learning outcomes.

The experiment took place over a four-week period. In the first two-hour session, students were introduced to *InterBook* and its features were explained to them. They used the system for an hour, and answered a questionnaire about its features. This questionnaire showed that almost all students were familiar with what each of the buttons and annotations meant. They were then free to use the system at any time during the following week. In the second session, students were randomly divided into two groups of equal size, one group receiving link annotation, while the other group did not. They were allowed access to the chapter of the textbook on databases, which had been authored into InterBook, and they completed a questionnaire. Students had access to the database chapter for the following week. In the third session, students took a multiple-choice test on the database section of the textbook. InterBook navigation logs were analysed along with the test results and the questionnaire responses.

An interesting aspect of the study, and the reason that this study could be used to support the case for layered evaluation, is that *it brought no significant results*. In particular, while students seem to understand and like adaptive navigation support (ANS) features, it did not 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. These results were surprising: following our past experience with a similar system (Brusilovsky and Pesin, 1998), we have expected the users of an adaptive version to achieve better test results.

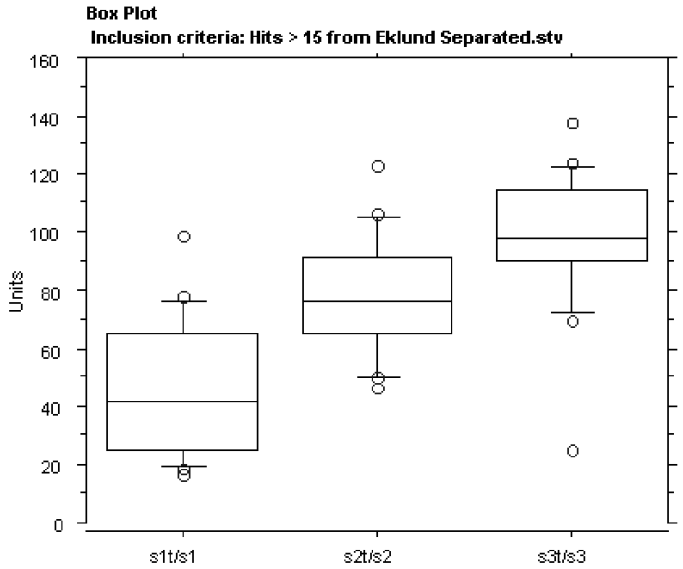
An analysis of the audit trails revealed at least one explanation of this result: about 80% of all navigation steps were made with continue and back buttons, or with hot words in text which were *not* annotated in the experimental version of *InterBook*. Only about 20% of step were made using *annotatable* links (i.e. links that were annotated in the ANS version). Moreover, only about a half of these clicks were made by the students of the ANS group who could see these links adaptively annotated. 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.

In any case, in our study 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 system

developers in such a situation is: *can we fine at least some differences between adaptive and non-adaptive versions*? It is exactly the question we have tried to answer in the original report of the study (Brusilovsky and Eklund, 1998). However, from the layered evaluation prospect of presented in this paper, we should have considered different questions*: Why does the adaptation not work? Was it the user modelling part where the system has performed poorly? Was it the adaptation decision making part where the adaptation decisions were not 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 user modelling 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 4). Since the pages were about the same size, the average reading time provides a reasonable estimation of page difficulty for students.

Figure 4 Average page visit time (seconds) for pages accessed though non-annotated links split by page category



s1t/s1 denotes 'nothing new' pages, s2t/s2 ready pages, and s3t/s3 'not ready' pages.

This data hints that the UM process, which predicts an educational status of 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.

It is important to note that what we observe in Figure 4 is the 'real value' of a page's pedagogical state. As we noted, in about 90% of cases the students navigated to learning material pages with non-annotated links and thus without any warning about the page state. If the students were able to see 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' behaviour. 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 reasonably fortunate not to have back and continue links annotated, since it enabled us to get reliable evidence that the user modelling component of the system works reasonably well.

In a situation where the UM 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 has not brought any significant results either. We think that it is a good example of how important is to have a good understanding of 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 (Weber and Specht, 1997), 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 (1999) 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. 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. 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 (Brusilovsky and Pesin, 1998), was not a good choice for teacher education students with little or no knowledge of the subject and background knowledge.

5 Layered evaluation in the adaptive web-based training platform KOD

To provide an example of planning layered evaluation of adaptive learning systems, we present below the KOD system. Our goal is to demonstrate how its adaptation can be decomposed and evaluated using the layered approach. The KOD (knowledge-on-demand) system is an adaptive learning environment providing personalised web-based content (Sampson et al., 2002a, 2002b). The KOD system is built upon the use of learning technology specifications. The IMS content packaging (CP) specification enables users (learning material authors, tutors, publishers, e-learning platform and service providers, etc.) to describe and structure learning objects using a common format (IMS Global Learning Consortium, 2001). Each content package is a single zip file which includes:

- the learning objects included in the package
- an XML file called manifest, which describes the sequence of learning objects included in the package.

This sequence is static (like a table-of-contents structure) with no capability of incorporating adaptation logic in a content package. As a result, learners access the same sequence of learning objects when navigating through a content package.

Recognising the limitations of these implementations in meeting the demanding needs of today's educational settings, the KOD system developed the knowledge packaging format (KPF), so that it can support the common content description and definition of adaptation logic (Sampson et al., 2002a). The difference of the KPs compared to the CPs is that the manifest includes in addition *adaptation rules* that determine which learning objects of the knowledge package should be selected for each learner, according to his/her particular profile.

The need for incorporating adaptation logic in content packages has been recognised worldwide. Indeed, the new version of Sharable content object reference model (SCORM 2004) proposes the use of IMS simple sequencing specification (SCORM sequencing and navigation v1.3) as a mechanism for defining sequencing rules inside a content package. In this paper we address the evaluation of adaptive learning systems, thus we focus on the adaptation logic itself rather than the technical approach used for describing the adaptation logic.

The demonstration and evaluation phase of the project involved the development and assessment of different knowledge packages (Sampson et al., 2002a). Following the 'traditional' evaluation methods of adaptive learning environments, the assessment of the KOD system would be conducted as follows:

- the KOD system would be installed in the demonstration sites
- one group of learners would work with the KOD (adaptive) system, i.e. accessing adaptive learning material through knowledge packages
- the same, or a different group of learners would work with the non-adaptive system, i.e. accessing learning material through content packages
- both groups would be then assessed according to pre-selected criteria (e.g. answer to quizzes) so as to evaluate whether adaptation was successful.

If the UM phase indicates that KOD is 'superior' (e.g. in terms of learning effectiveness), then the adaptation of the KOD system is considered successful. If, however, the KOD system is proved less effective, then we would not be able to identify the 'source' for this unsatisfactory result:

- it could be the case that the learner model of the KOD system is not appropriate; i.e. that the conclusions made by the KOD system for learners' background, preferences, etc., are not correct
- it could also be the case that the KOD learner model is satisfactory, but the (instructional design) rules included in the content package are not successful.

As it is evident from the description of KOD system, one of the key objectives is to be able to interchange learning material *together* with adaptation rules. As a result, we are particularly interested in evaluating these adaptation rules, so that they can be re-used. Therefore, we have adopted the layered evaluation approach, since it can ensure the effectiveness of adaptation rules, before they are interchanged and re-used.

6 Discussion and conclusions

This paper has argued that the evaluation practices for adaptive learning systems, and adaptive applications and services, in general, need to be informed and improved by adaptive systems models. It has outlined the *layered evaluation* framework, where the success of adaptation is addressed at two distinct layers: user modelling, and adaptation decision making. The paper has attempted to demonstrate the benefits of the layered evaluation framework by reconstructing one of our earlier studies that could greatly benefit from the suggested approach. It has also presented a specific example where an evaluation framework has been elaborated based on a specific adaptation model.

The paper argues that the proposed framework is a wise approach for the evaluation of ALS. Layered evaluation can provide useful information for their improvement, and can contribute towards the generalisation of evaluation results, and the re-use of successful design practices. Since user modelling is evaluated separately, the logic underlying this process can be re-used in similar contexts. Similarly, since adaptation decision making is evaluated separately, the underlying logic can be re-used in similar contexts.

The idea of evaluating separately the user modelling and the adaptation decision making phases has been implied in previous work, such as the work by Totterdell and Boyle (1990), where two types of evaluation are suggested for the user model (which, in this case, is also responsible for the adaptation decision making): "an assessment of the accuracy of the model's inferences about user difficulties; and an assessment of the effectiveness of the changes made at the interface" (Totterdell and Boyle, 1990). Moreover, the idea of decomposing adaptation into user modelling and adaptation decision making has also been expressed in the past – e.g. (Brusilovsky, 1996). The goal of this paper is to re-introduce this idea, to suggest an explicit layered evaluation framework, and to build a case for it by presenting a specific study where the benefits of layered evaluation are clearly visible.

It should be noted that the proposed framework is based on a 'first level' decomposition of adaptation, depicted in Figure 1. The components of this figure can be

further decomposed, therefore, layered evaluation could also be elaborated to address these sub-components. For example, adaptation could be decomposed into alternative components, e.g. into a 'higher level adaptor', i.e. the process within which the application assesses (at a meta level) whether adaptations have met their goal, and, if necessary, modifies (adapts) the 'lower level adaptor' – thus, in effect modifying how the system adapts (Totterdell and Rautenbach, 1990). A more detailed model of the adaptation process can provide a basis for a more fine-grain evaluation framework. Some good examples are provided by a three-component model in (Weibelzahl, 2001) and seven-component model in (Paramythis et al., 2001).

A finer-grain model-based framework offers the same benefits as the framework presented in this paper. It can further decrease the amount of work to be done before the evaluation become possible and increase the level of re-use of good design decision. At the same time, it become less universal and implicitly oriented on a specific category of adaptive systems. In this context, our two-component model presents a useful compromise between usefulness and generality. The position of this paper is that this first level decomposition of adaptation into the user modelling and adaptation decision making layers can provide significant insight into the success of adaptation, and valuable guidelines for the process of evaluation.

It should be also noted that layered evaluation only addresses the 'success of adaptation' which is the defining characteristic of adaptive applications and services. However, it can be argued that even if adaptation is successful, this does not necessarily mean that the adaptive application is considered acceptable by its end users. It has been long argued that adaptation is not a goal in itself, but rather a way of improving the effectiveness of ALS, or the usability of interactive applications, and that there may be many other ways for the same goal (Schneider-Hufschmidt et al., 1993). Consider, for example, the issue of system performance: adaptation poses additional computational load, since computational power needs to be assigned for the user modelling and the adaptation decision making. Given the (desired) portability and interoperability of modern applications and services, this load may be unacceptable for a specific application when provided through a slow machine, or over a network with a limited bandwidth.

Moreover, acceptability is widely recognised to be a complex issue, which is directly affected by organisational factors and context of use. There are still several additional issues that need to be taken into account when evaluating the overall acceptability of ALS (Höök, 2000). For example, it has been argued that adaptive applications bear the risk of the user feeling a loss of control of the application, or not trusting the application (Schneiderman and Maes, 1997). As another example, consider the privacy and security of the information stored for the user. As it is evident, adaptation presumes that information concerning the user's abilities, requirements, preferences, etc., is stored and analysed. The security of this information should be carefully taken into account, so that users feel comfortable with adaptive applications and services.

The proposed framework does not intend to replace current evaluation practices. It rather proposes a 'structured approach' to evaluation, where the main phases are evaluated separately. As such, the evaluation of each separate layer can make use of any existing evaluation technique, such as heuristic evaluation, user experiments, etc. Table 2 suggests how existing evaluation practices presented in (Chin, 2001) can be informed and improved by the layered evaluation framework.

Table 2 Empirical evaluation of adaptive systems

Domain	*Success of adaptation*		*Overall success of the system*
	User modelling	*Adaptation decision making*	
Adaptive hypermedia/hypertext	Evaluate the validity of the information maintained in the user model	Evaluate the success of the adaptation decisions for adaptive navigation and/or adaptive presentation	Measure of recall and precision Similarity/relevance metrics
Student modelling	Evaluate the validity of the information maintained in the student model	Evaluate the success of the adaptation decisions of the tutoring model	Comparison of the system with- and without-adaptation
Plan recognition	Percentage of actual plans recognised in a test corpus of plans	Frequency and accuracy of predicted next actions Comparison with an expert human plan recogniser Comparing system responses choices with human choices	Evaluate the overall success of the system
Mixed-initiative interaction	Evaluate the validity of the information maintained in the user model	Efficiency of the dialogue needed to achieve an information transfer task with either human-human dialogues or with theoretically minimum dialogues	Evaluate the overall success of the system
User interfaces/help systems	Evaluate the validity of the information maintained in the user model	Evaluate the success of the adaptation decisions	Subjective user satisfaction Task completion speed Error rate/quality of task achievement

Text in gray cells outlines the additions that need to be made to current evaluation practices (according to Chin, 2001) in the light of the layered evaluation framework.

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.iti.gr) through the Information Society Technologies Programme (IST).

References

Benyon, D. and Murray, D. (1993) 'Developing adaptive systems to fit individual aptitudes', *Proc. 2nd ACM International Conference on Intelligent User Interfaces*.

Billsus, D., Brunk, C.A., Evans, C., Gladish, B. and Pazzani, M. (2002) 'Adaptive interfaces for ubiquitous web access', *Communications of the ACM*, Vol. 45, No. 5, pp.34–38.

Brusilovsky, P. (1996) 'Methods and techniques of adaptive hypermedia', *User Modeling and User-Adapted Interaction*, Vol. 6, Nos. 2–3, pp.87–129.

Brusilovsky, P. (1999) 'Adaptive and intelligent technologies for web-based education', in Rollinger, C. and Peylo, C. (Eds.): *Künstliche Intelligenz*, No. 4, Special Issue on Intelligent Systems and Teleteaching, pp.19–25.

Brusilovsky, P. (2001) 'Adaptive hypermedia', *User Modeling and User Adapted Interaction*, Vol. 11, Nos. 1–2, pp.87–110.

Brusilovsky, P. and 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, Vol. 4, No. 4.

Brusilovsky, P. and Maybury, M.T. (2002) 'From adaptive hypermedia to adaptive web', *Communications of the ACM*, Vol. 45, No. 5, pp.31–33.

Brusilovsky, P. and Pesin, L. (1998) 'Adaptive navigation support in educational hypermedia: an evaluation of the ISIS-tutor', *Journal of Computing and Information Technology*, Vol. 6, No. 1.

Brusilovsky, P. and Peylo, C. (2003) 'Adaptive and intelligent web-based educational systems', *International Journal of Artificial Intelligence in Education*, Vol. 12, Nos. 2–4, pp.159–172.

Brusilovsky, P., Eklund, J. and Schwarz, E. (1998) 'Web-based education for all: a tool for developing adaptive courseware', *Proc. 7th International World Wide Web Conference*.

Cheverst, K., Mitchell, K. and Davies, N. (2002) 'The role of adaptive hypermedia in a context-aware tourist GUIDE', *Communications of the ACM*, Vol. 45, No. 5, pp.47–51.

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

Claypool, M.L.P., Wased, M. and Brown, D. (2001) 'Implicit interest indicators', *Proc. of 6th International Conference on Intelligent User Interfaces*, Santa Fe, NM, ACM Press, pp.33–40.

De Bra, P. and Calvi, L. (1998) 'AHA! an open adaptive hypermedia architecture', *The New Review of Hypermedia and Multimedia*, Vol. 4.

Devedzic, V.B. (2003) 'Key issues in next-generation web-based education', *IEEE Transactions on Systems, Man and Cybernetics Part C*, Vol. 33, No. 3, pp.339–349.

Eklund, J. and Brusilovsky, P. (1998) 'The value of adaptivity in hypermedia learning environments: a short review of empirical evidence', *Proc. Workshop on Adaptive Hypertext and Hypermedia, 9th ACM Conference on Hypertext and Hypermedia*.

Hohl, H., Böcker, H.D. and Gunzenhäuser, R. (1996) 'Hypadapter: an adaptive hypertext system for exploratory learning and programming', *User Modeling and User-Adapted Interaction*, Vol. 6, Nos. 2–3.

Hook, K. (1997) 'Evaluating the utility and usability of an adaptive hypermedia system', *Proc. 3rd ACM International Conference on Intelligent User Interfaces*.

Höök, K. (2000) 'Steps to take before intelligent user interfaces become real', *Interacting with Computers*, Special Issue on Intelligent Interface Technology, Vol. 12.

Höök, K. and Svensson, M. (1998) 'Evaluating adaptive navigation support', *Proc. IFIP Workshop on Personalized and Social Navigation in Information Space*.

IMS Global Learning Consortium (2001) *Content Packaging Specification*, Version 1.1.3.

Karagiannidis, C., Koumpis, A. and Stephanidis, C. (1997a) 'Adaptation in intelligent multimedia presentation systems as a decision making process', *Computer Standards and Interfaces*, Special Issue 'Towards a standard reference model for intelligent multimedia presentation systems', Vol. 18, Nos. 6–7.

Karagiannidis, C., Koumpis, A. and Stephanidis, C. (1997b) 'Modeling decisions in intelligent user interfaces', *International Journal of Intelligent Systems*, Vol. 12, No. 10.

Karampiperis, P. and Sampson, D. (2004) 'Adaptive learning object selection in intelligent learning systems', *Journal of Interactive Learning Research*, Special Issue on Computational Intelligence in Web-based Education.

Kobsa, A., Koenemann, J. and Pohl, W. (2001) 'Personalized hypermedia presentation techniques for improving online customer relationships', *The Knowledge Engineering Review*, Vol. 16, No. 2, pp.111–155.

Manouselis, N. and Sampson, D. (2003) 'Agent-based e-learning course recommendation: matching learner characteristics with content attributes', *International Journal of Computers and Applications (IJCA)*, Special Issue on Intelligence and Technology in Educational Applications, Vol. 25, No. 1.

Maybury, M. and Wahlster, W. (Eds.) (1998) *Evaluation, in: Readings in Intelligent User Interfaces*, Morgan Kaufmann.

Paramythis, A., Totter, A. and Stephanidis, C. (2001) 'A modular approach to the evaluation of adaptive user interfaces', *Proc. Workshop on Empirical Evaluation of Adaptive Systems* at the *8th International Conference on User Modeling*.

Passardiere, B.D.L. and Dufresne, A. (1992) 'Adaptive navigational tools for educational hypermedia', *Proc. 4th International Conference on Computers and Learning*.

Russell, T.L. (1999) *The No Significant Difference Phenomenon* as reported in 355 Research Reports, Summaries and Papers: A Comparative Research Annotated Bibliography on Technology for Distance Education, North Carolina State University, Office of Instructional Telecommunications.

Sampson, D., Karagiannidis, C. and Cardinali, F. (2002a) 'An architecture for web-based e-learning promoting re-usable adaptive educational e-content', *Educational Technology & Society Journal*, Vol. 5, No. 4.

Sampson, D., Karagiannidis, C. and Kinshuk (2002c) 'Personalised learning: educational, technological and standardisation perspectives', *Interactive Educational Multimedia*, Vol. 4.

Sampson, D., Karagiannidis, C., Schenone, A. and Cardinali, F. (2002b) 'Integrating a knowledge-on-demand personalised learning environment in e-learning and e-working settings', *Educational Technology & Society Journal*, Vol. 5, No. 2.

Schneider-Hufschmidt, M., Kuehme, T. and Malinowski, U. (Eds.) (1993) *Adaptive User Interfaces*, Elsevier.

Schneiderman, B. and Maes, P. (1997) 'Debate: direct manipulation vs. interface agents', *Proc. 3rd ACM International Conference on Intelligent User Interfaces*.

Specht, M. and Kobsa, A. (1999) 'Interaction of domain expertise and interface design in adaptive educational hypermedia', *Proc. 2nd Workshop on Adaptive Systems and User Modeling on the World Wide Web*.

Totterdell, P. and Boyle, E. (1990) 'The evaluation of adaptive systems', in Browne, D., Totterdell, P. and Norman, M. (Eds.): *Adaptive User Interfaces*, Academic Press.

Totterdell, P. and Rautenbach, P. (1990) 'Adaptation as a problem of design', in Browne, D., Totterdell, P. and Norman, M. (Eds.): *Adaptive User Interfaces*, Academic Press.

Weber, G. and Specht, M. (1997) 'User modeling and adaptive navigation support in WWW-based tutoring systems', *Proc. 6th International Conference on User Modeling*, pp.289–300.

Weibelzahl, S. (2001) 'Evaluation of adaptive systems', *Proc. 8th International Conference on User Modeling, UM 2001*, pp.292–294.