

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 centralized 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 inferred 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].

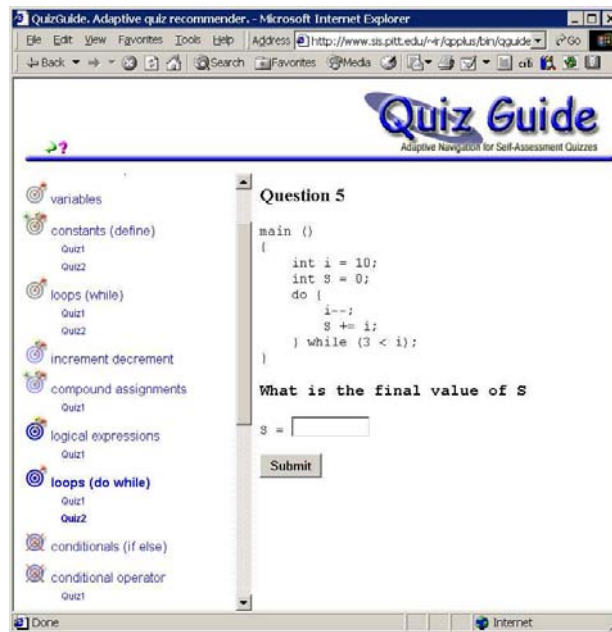


Fig. 1. Student interface of QuizGuide

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.

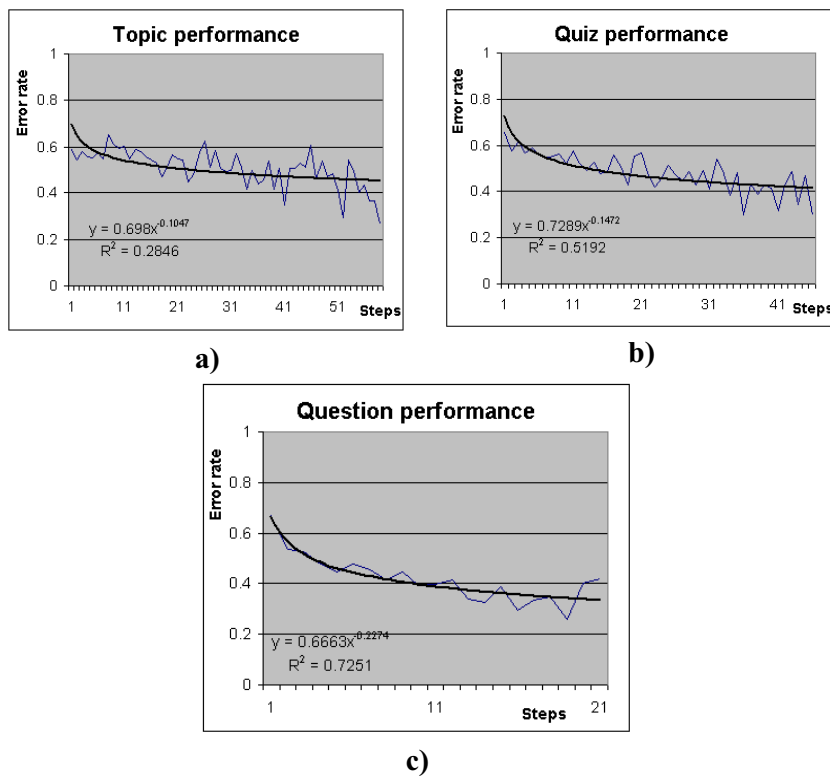


Fig. 2. Learning curves for the average topic (a), quiz (b) and question (c)

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.

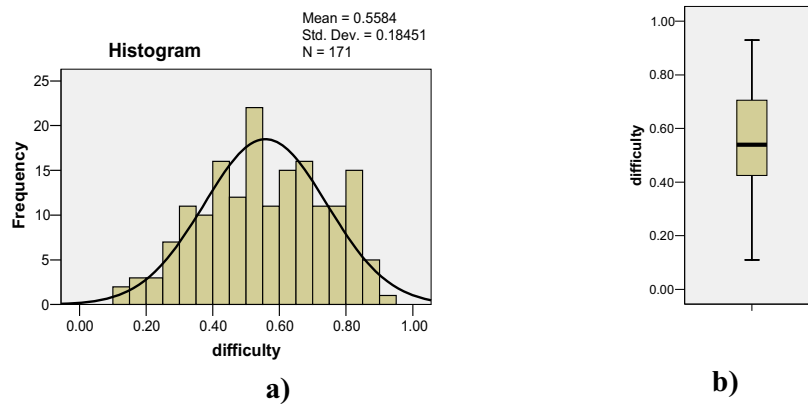


Fig. 3. The histogram (a) and the boxplot (b) of QuizGuide questions distributed according to their average difficulty.

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 than 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 than 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.

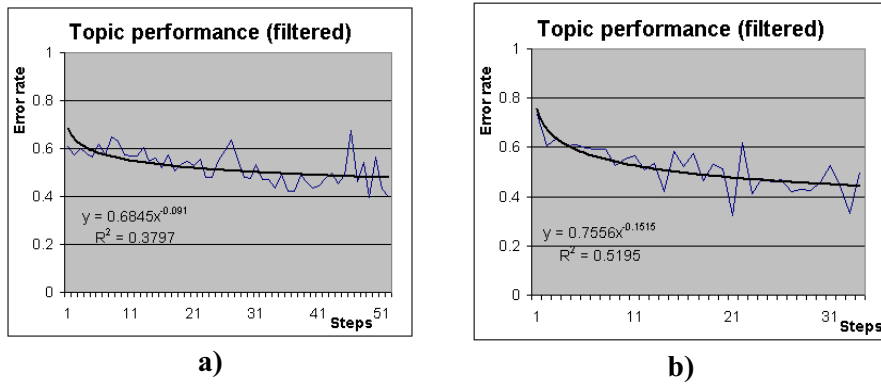


Fig. 4. Learning curves for the average topic, where 10% (a) and 50% (b) of “bad” questions are filtered out

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 i^{th} 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.

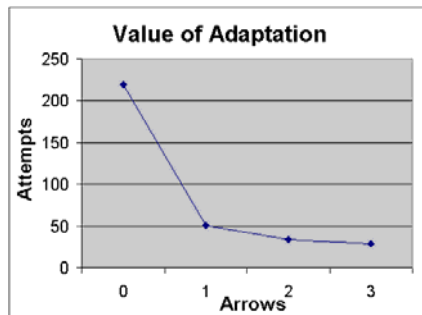


Fig. 5. Value of adaptation provided by QuizGuide

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 Nejd 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*.