Adaptive training based upon computerized knowledge assessment

MICHEL C. DESMARAIS JIMING LIU DAVID MALUF
Centre de recherche informatique de Montréal
1801 ave. McGill College, bureau 800
Montréal, Québec, Canada H3A 2N4
e-mail: desmarais@crim.ca

February 24, 1994

Introduction

One of the most praised qualities of computer assisted training (CAT) is its ability to deliver individualized training. In CAT, the pace of individual learning is not determined by the learning group's average, or maximum, or minimum learning pace. It adapts to the individual's capabilities. Even more, the content can be tailored to the individual's background knowledge and personal training needs.

Current state-of-the art CAT systems support individualization of the training process by allowing free or flexible navigation within the training content, as opposed to a rigid linear content presentation controlled by the computer. However, this flexibility imposes an additional burden on both the instructional designer and the learner. For example, the instructor must select the appropriate sequence of exercises according to the successes and failures of the learner. Or, the learner must explore the whole training content and decide what is appropriate.

However, dynamic modeling of a user's current knowledge state can alleviate this problem. With such capabilities, it becomes possible to specify the learning material's prerequisites and appropriate level of expertise in such a way as to avoid presenting too advanced or trivial content to a given user. Moreover learning material can be tailored in a very specific manner, targeting only the necessary content for a given learner's knowledge state.

We present a method for the dynamic modeling of a user's knowledge state. The method induces, from empircal data, the learning order of knowledge units (KU—they can represent mastery of concepts or skills). This ordering is thereafter used to infer an individual's knowledge state (see Falmagne *et al.*, 1990, for the cognitive foundations of this method). Through a sampling of a person's knowledge state, the system assesses the probability that this person knows any given KU.

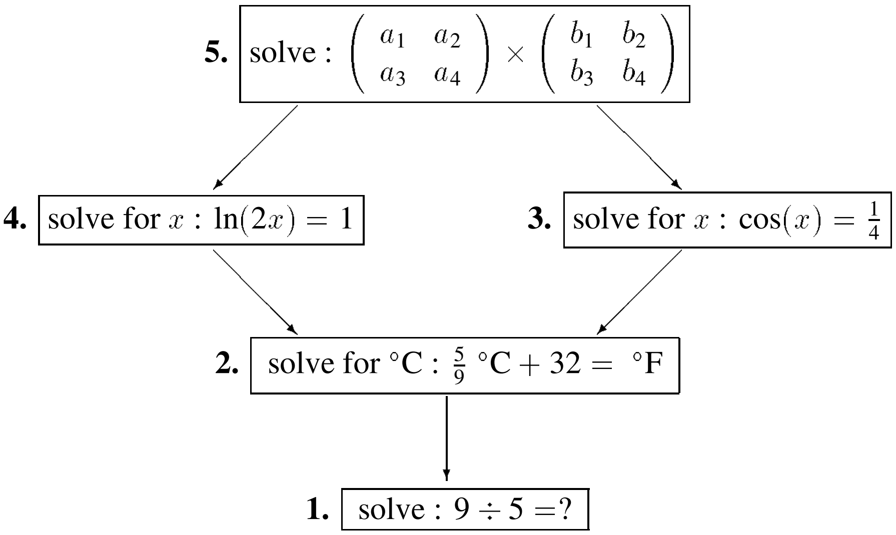

Figure 1: Inference network among abilities to solve five mathematical problems.

The induction of order among knowledge units

The current method of knowledge assessment is based upon the fact that people learn KU in a more or less predetermined order. This ordering is not linear nor totally strict, but the order imposes constraints on the possible knowledge states. In the current study, we model the different possible knowledge states by a *partial order*. Although this formalism could not fully represent all possible knowledge states (Degreef et al., 1986; Falmagne et al., 1990), it captures a large part of the constraints on the ordering among KU and can be used for the purpose of automatic knowledge assessment. An example of such ordering is given in figure 1. This example states that among the 32 combinations, only 7 knowledge states are plausible: $\{\{\}, \{1\}, \{1, 2\}, \{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 3, 4\}, \{1, 2, 3, 4, 5\}\}$.

The induction method is based on a pairwise comparison of every KU in search of an implication relation. It produces a partial order of KU and it has the great advantage of working with very small amount of data cases.

The basic idea behind the empirical construction is that if there is an implication relation $A \Rightarrow B$, then we would ideally never expect to find an individual who knows A but not B. This translates into two conditions: $P(B|A) \approx 1$ and $P(\neg A|\neg B) \approx 1$. These conditions are verified by computing the lower bound of a $[1 - \alpha_{\text{error}}]$ confidence interval around the measured conditional probabilities. If the confidence intervals are above a predefined threshold, p, an implication relation between the two KUs is asserted. Two weights are associated with the relation[1] which correspond to the relation's conditional probabilities $P(B|A)$ and $P(\neg A|\neg B)$. In fact, these weights express the degree of certainty in that relation.

[1] According to the two directions of the inference, i.e. *modus tollens* vs. *modus ponens*.

Knowledge assessment with evidence propagation

Once the structure of implication among KU is obtained, it is possible to use this network with an evidencial propagation method to assess an individual's knowledge state. The current study is based upon the Dempster-Shafer evidence propagation scheme (Gordon and Shortliffe, 1984). Other propagation scheme can be used such as, for example, the Bayesian scheme which was used by DeRosis (1992) in a similar fashion for the purpose of user modeling. However, informal experiments within our laboratory has shown relatively similar performance between the Dempster-Shafer and Bayesian methods of evidence propagation.

A guided tour of the UKAT adaptive questionnaire

Examples of the utility of such a method are numerous. UKAT is an application of this method for an adaptive questionnaire. Every question represents a KU. For each question the user succeeds or fails, the probability of correctly answering every other question is updated automatically based upon the empirically derived ordering of questions. Twenty subjects were asked to complete the questionnaire and the data was thereafter used to induce the ordering.

Figure 2 is a picture of the interface to UKAT. Questions are displayed on the right side. The user can simply click on the chosen answer. The left side illustrates the knowledge assessment status and includes some relevant information about each question. For example, the status corresponds to an incorrect answer to question 1. From left to right, the first column indicates the question's number. The next three columns contain indices about entropy, which will be ignored here. The column entitled "initial" indicates the estimate of each question's probability of success based on the sample data. This provides an indication of the degree of difficulty of each question. The "last" and "current" columns display respectively the last and current estimates of the probability of success. Some of the columns' averages are displayed at the bottom. They represent global scores of initial, previous ("last"), and updated ("current") scores.

To illustrate the behaviour of UKAT, let us define a number of scenarios that are summarized in figure 3. The graph contains six scenarios. In each scenario four questions are asked and the system updates the user's global score accordingly. The initial assessment is based upon the 19 subjects' average of 61% and the six lines represent the evolution of the different global score averages:

1. S-Dif: (Success-Difficult) consecutive successes to most difficult questions;

2. S-Ent: (Success-Entropy) consecutive successes to entropy-driven questions;

3. S-Eas: (Success-Easy) consecutive successes to easiest questions;

4. F-Dif: (Failure-Difficult) consecutive failures to most difficult questions.

5. F-Ent: (Failure-Entropy) consecutive failures to entropy-driven questions;

6. F-Eas: (Failure-Easy) consecutive failures to easiest questions;

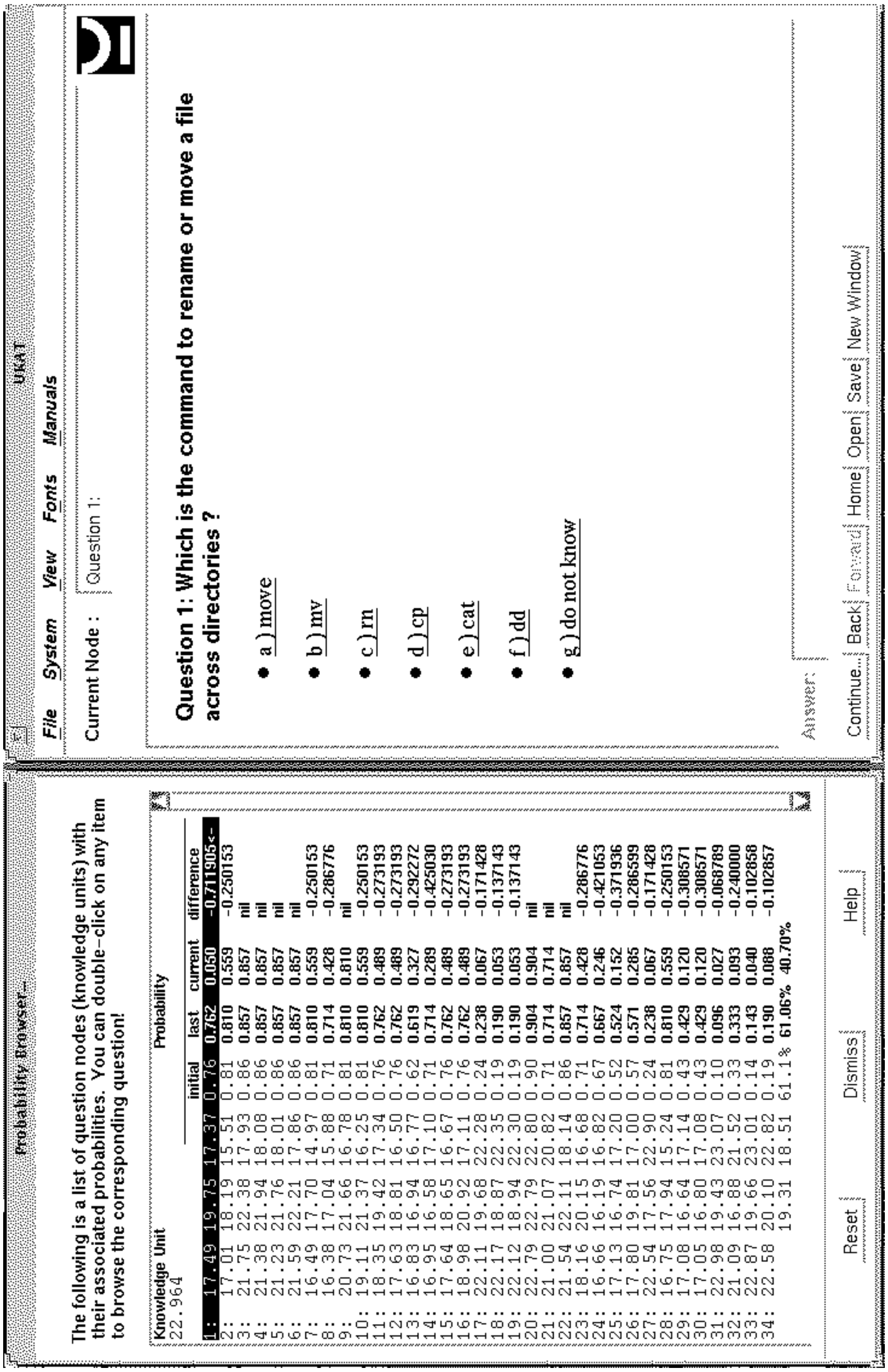

Figure 2: Interface to UKAT.

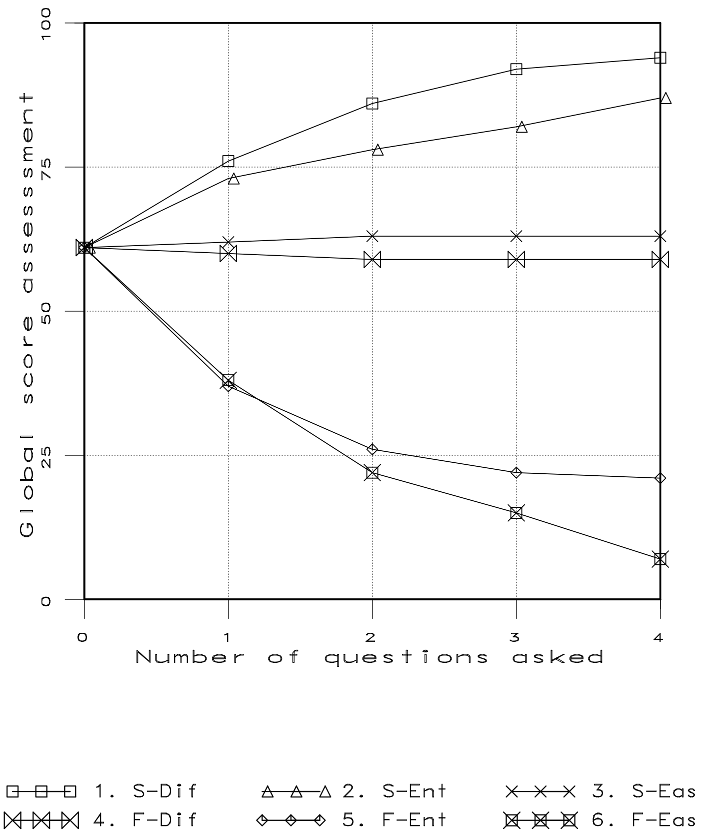

Figure 3: Global score assessment evolution with four questions and according to a number of scenarios: S-Dif: (Success-Difficult) consecutive successes to most difficult questions; S-Ent: (Success-Entropy) consecutive successes to entropy-driven questions; S-Eas: (Success-Easy) consecutive successes to easiest questions; F-Dif: (Failure-Difficult) consecutive failures to most difficult questions. F-Ent: (Failure-Entropy) consecutive failures to entropy-driven questions; F-Eas: (Failure-Easy) consecutive failures to easiest questions;

As expected, successes to difficult questions (scenario 1) rapidly raise the global score from an initial assessment of 61% close to 100% (94%) after only four questions asked. Conversely, the score is brought down very low (7%) with failures to easy questions (scenario 6). However, successes with easy questions (scenario 3), and failures with difficult questions (scenario 4) bring little information and hardly change the initial assessment. This behaviour is consistent with intuitive expectations about such systems. A closer look also reveals that failures at an easy question will lower the probability of success of more difficult questions, and vice-versa. This is shown in figure 2 where failure to answer correctly question 1 affects the more difficult questions but generally leaves questions of the same degree of difficulty intact.

Since it is more likely that a user will fail difficult questions and succeed easy ones, entropy-driven questions can be used to choose the most informative questions. Figure 3 reveals that consecutive successes or consecutive failures to entropy-driven questioning do affect the score significantly after four questions (with resulting scores of 87% and 21% respectively), though not as much scenarios 1 and 6.

Validation results

The previous section provides an overview of the system's validity from an intuitive perspective through scenarios. A more thourough validation was performed to systematically validate the results. Figure 4 illustrates the validation results for the same data set consisting of data on the 34 questions about the UNIX command language and utilities. All validations are based around an entropy-driven simulation of questions asked and comparison of inferences with actual test data. It is composed of three validations: a high-risk inference network ("Hig-rsk"), a medium-risk network ("Med-rsk") and a low-risk network ("Low-rsk"). The three networks corresponds to the different alpha errors used to induce the network from empirical data (0.99, 0.50, and 0.20 respectively—the threshold p was kept constant at 0.50; the alpha 0.99 corresponds to keeping all relations above p). The high-risk network contains the greatest number of relations (436) and performs relatively more inferences than the others, but it also produces the greatest number of errors. On the contrary, the low-risk network contains only 86 relations but it makes much fewer mistakes. The medium-risk network contains 343 relations.

The solid line ("Obser") serves as a comparison point for each of the three validations. It stands for the "no-inference" condition. The graph represents the standard error score as a function of the number of questions asked. For each subject, the difference between the probability of success of a question is compared with the actual test result of this subject (the second moment is taken which represents standard error score). This number is averaged over all questions and all subject. When no questions are asked, the standard error is approximately 0.38. As questions are answered, the network updates the questions probabilities of successes and new standard error score is computed for each inferred probability. The observed questions are assumed to be correctly assessed and thus the standard error score is 0 when all questions are answered.

As figure 4 illustrates, the high-risk and low-risk networks generally perform slightly worse than the medium-risk network. However, the low-risk network tends to perform better than the others after 20 questions. The medium-risk network has reduced the standard error score by

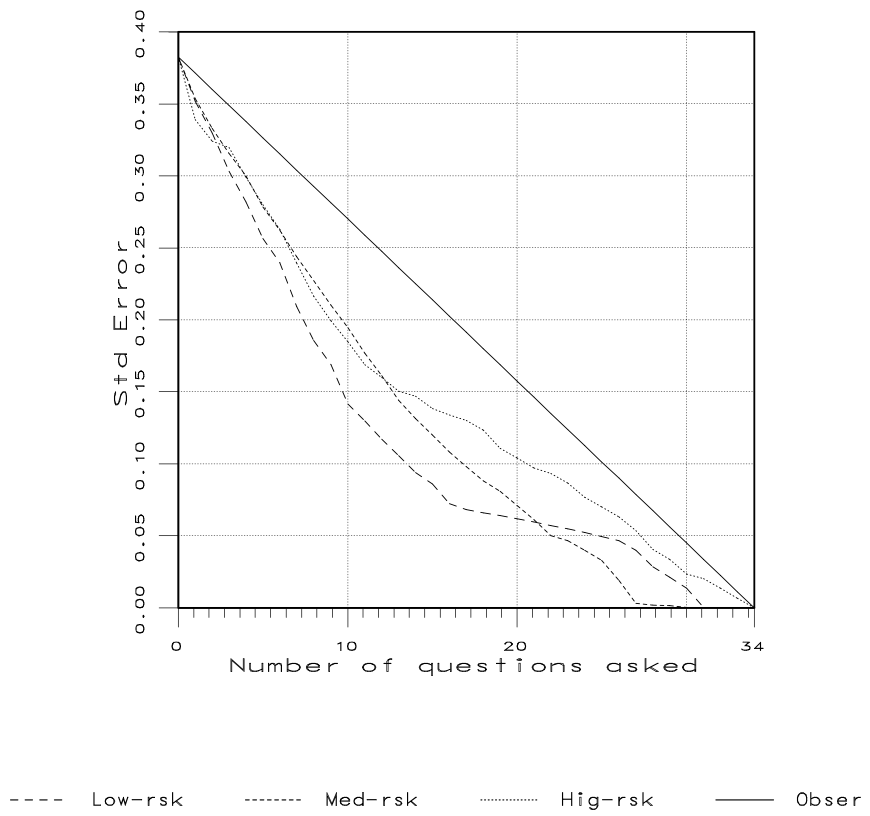

Figure 4: Standard error score of knowledge assessment as a function of the number of questions asked. The three lines entitled “low-rsk”, “med-risk”, and “hig-rsk” represent the performance of three networks that differ in the Alpha parameter used in their construction. The Aplha parameter affects the number and the exactness of network inferences.

almost half after only 10 questions, which is quite encouraging[2].

Systematic validations of the method for assessing an individual's knowledge state were performed in a number of different experiments on the knowledge of WordPerfect™ text-editing commands (Desmarais and Liu, 1993; Desmarais et al., 1993). These results showed that, for example, a close to perfect knowledge assessment was inferred after sampling 70% of a subject's knowledge state, and that sampling 50% of the knowledge state would reduce the standard error score of estimates to about half of the error score without inferences. Thus, the method was successful in reducing the number of questions that needs to be asked to assess a subject's score.

Other applications of such tool for expertise assessment are adaptive hypermedia structure, in which only nodes appropriate to a given level of expertise should be accessible to a user of the corresponding level. We are also currently working in conjunction with Hydro-Quebec to build a large scale questionnaire composed of hundreds of questions. UKAT will be used to minimize the number of questions asked to an individual and to design automatically a specific training content that addresses precisely the skills that must be taught to a given worker without a need to cover material already known.

Conclusion

In order to increase the effectiveness and practical utility of different types of adaptive tutoring systems, it is essential to lower the difficulty of building such systems. The proposed method for knowledge assessment constitutes a promising approach since it replaces the tedious task of designing expertise-tailored didactic content and didactic strategies, by a flexible and automatic approach. This approach relies on the segmentation of the content into KU and on a data collection process to gather the information upon which the automatic method can be applied.

References

de Rosis, F., Pizzutilo, S., Russo, A., Berry, D. C., and Molina, F. J. N. (1992). Modeling the user knowledge by belief networks. *User Modeling and User-Adapted Interaction*, 2(4).

Degreef, E., Doignon, J.-P., Ducamp, A., and Falmagne, J.-C. (1986). Languages for the assessment of knowledge. *Journal of Mathematical Psychology*, 30:243–256.

Desmarais, M. C., Giroux, L., and Larochelle, S. (1993). An advice-giving interface based on plan-recognition and user knowledge assessment. *International Journal of Man-Machine Studies*. In press.

Desmarais, M. C. and Liu, J. (1993). Exploring the applications of user-expertise assessment for intelligent interfaces. In *InterCHI'93, Bridges between worlds*, pages 308–313.

[2]However, the validation was performed upon the same 19 subjects used for network construction. Thus it must contain a positive bias simply because the method can also exploit intra-data noise. We performed Monte Carlo simulations to verify the extent of this bias and the results suggest that the performance after 10 questions would be approximately a third of the standard error reduction instead of half as observed here.

Falmagne, J.-C., Doignon, J.-P., Koppen, M., Villano, M., and Johannesen, L. (1990). Introduction to knowledge spaces: how to build, test and search them. *Psychological Review*, 97(2):201–224.

Gordon, J. and Shortliffe, E. H. (1984). The Dempster-Shafer theory of evidence. In Buchanan, B. G. and Shortliffe, E. H., editors, *Rule-Based Expert Systems*. Addison-Wesley, Reading, M. A.