Exploring the Applications of User-Expertise Assessment for Intelligent Interfaces

Michel C. Desmarais *Jiming Liu*
Centre de recherche informatique de Montréal
1801 ave. McGill College, bureau 800
Montréal, Québec, Canada H3A 2N4
Tel: 514-398-1234 Fax: 514-398-1244
e-mail: desmarais@crim.ca jiming@crim.ca

ABSTRACT

An adaptive user interface relies, to a large extent, upon an adequate user model (e.g., a representation of user-expertise). However, building a user model may be a tedious and time consuming task that will render such an interface unattractive to developers. We thus need an effective means of inferring the user model at low cost. In this paper, we describe a technique for automatically inferring a fine-grain model of a user's knowledge state based on a small number of observations. With this approach, the domain of knowledge to be evaluated is represented as a network of nodes (knowledge units—KU) and links (implications) induced from empirical user profiles. The user knowledge state is specified as a set of weights attached to the knowledge units that indicate the likelihood of mastery. These weights are updated every time a knowledge unit is reassigned a new weight (e.g., by a question-and-answer process). The updating scheme is based on the Dempster-Shafer algorithm. A User Knowledge Assessment Tool (UKAT) that employs this technique has been implemented. By way of simulations, we explore an entropy-based method of choosing questions, and compare the results with a random sampling method. The experimental results show that the proposed knowledge assessment and questioning methods are useful and efficient in inferring detailed models of user-expertise, but the entropy-based method can induce a bias in some circumstances.

KEYWORDS

User-expertise assessment, probabilistic reasoning, evidence aggregation, entropy, intelligent interfaces, adaptive training systems, knowledge spaces.

[Published in "InterCHI'93, Bridges between worlds", Amsterdam, 24–29 April, pp. 308–313.

INTRODUCTION

Knowledge assessment is a fundamental component of user modeling. It has been used to adapt the level of technical details of on-line documentation and error messages; it is also an essential ingredient found in practically every intelligent tutoring systems, coaches, and consultants (see for e.g., [12, 15]). Its widespread use in commercial applications, however, remains very limited, or non-existent. We suspect that, among other things, one reason for this is the unavailability of simple and efficient knowledge assessment modeling techniques that non-specialists of AI/cognitive-modeling can use while developing their applications. There exist a number of approaches, from simple and not-that-useful, to sophisticated and costly, but the tradeoff between complexity, development-cost, and usefulness rarely allows the application of expertise modeling techniques outside the experimental laboratory.

The most simple approach to this problem consists of categorizing the user onto a novice-expert scale. Although this approach is feasible with moderate development cost, it provides only very coarse information. Another technique is based on the idea that a knowledge domain can be defined as a set of knowledge units (KU). An individual's knowledge state is defined as a subset of it. This model thus constitutes a *fine grain* assessment to the extent that it provides information about each individual KU. Moreover, the set of KUs can be structured with various types of relations [8]. One such relation is the *precedence* relation, which indicates the order in which KUs are learned. A number of researchers have used this type of relation to infer user knowledge state (e.g., [1, 10, 14]). The main problem lies in building this structure of relations among KUs, which can be a very tedious task when the number of KUs is more than a few tens, unless the task can be automated.

This paper presents a technique for automatically constructing a structure of precedence relations among KUs from a small number of subject's knowledge states. The structure is thereafter used in conjunction with the Dempster-Shafer evidential reasoning method [9, 13] for assessing someone's knowledge state. This provides an entirely algorithmic approach to knowledge assessment, relieving the developer

from the burden of building such tools.

In the remainder of the paper, we provide an overview of the knowledge structure induction and user knowledge inference techniques, and describe a series of empirical tests on an implemented module with two approaches to knowledge inference: a first approach in which KUs are chosen at random (a situation similar to that of a non-obstructive user-modeling facility, where KUs are observed but not chosen) and a second one in which KUs are subject to explicit choices based on entropy-minimization.

OVERVIEW

In the current modeling approach, the domain knowledge is represented as a *knowledge structure* [7], whose nodes are the fine-grain knowledge units (KU). An individual's knowledge about the domain, i.e., a knowledge state, is modeled by a collection of numerical attribute values attached to the nodes. Each value indicates the likelihood (i.e., probability) of a user's knowing a specific KU. In the network, KUs are connected by implication (precedence) relations. An implication relation is in fact a gradation constraint which expresses whether a certain concept has to be understood before another difficult one, or whether a certain skill is acquired prior to an advanced one.

Empirical Construction of Knowledge Structures

In contrast to other work that also adopted similar approaches (see in particular [1]), the knowledge structure in the current study is induced entirely from empirical data composed of samples of knowledge states. Because the knowledge structure induction process is automatic, it allows a much larger number of KUs to be included than other approaches.

The basic idea behind the empirical construction is that if there is an implication relation $A \Rightarrow B$, then ideally we would 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, an implication relation between the two KUs is asserted. Two weights are associated with the relation (according to the two directions of the inference, i.e. *modus tollens* vs. *modus ponens*). They correspond to the relation's conditional probabilities $P(B|A)$ and $P(\neg A|\neg B)$. The weights express the degree of certainty in that relation. This method is inspired by previous work [3, 4]. A more detailed treatment of the knowledge structure construction method is given in [11].

Aggregation of Evidence

Once a knowledge structure is obtained, it can be used as a basis for knowledge assessment. The knowledge state of a user is built and updated as soon as some observations are made (e.g., questions are answered). Each observation can be viewed as a piece of evidence. This new information may be propagated to other nodes in compliance with the gradation constraints (inference structure).

In the present work, we applied the Dempster-Shafer (DS) evidential reasoning to recursively propagate evidential supports (whether confirming or disconfirming) throughout the knowledge structure. Different degrees of support, gathered from different sources of evidence, are combined to yield a new weight using the Dempster's rule of combination [9].

Selecting Questions

Observations about a user's knowledge state can be driven by a user monitoring module that reports evidence of known KUs[1], or they can be made through a sequence of question-and-answer sessions. In the first approach, evidence of known and unknown KUs is not under control and can thus be considered a random process. However, in the question-and-answer sessions, the evidence gathering process can be controlled during the knowledge assessment through an entropy-driven selection method. This approach, advocated in a number of previous studies (e.g., [6, 7]), applies the rule of minimum entropy and chooses the most *informative* questions. In the entropy-driven method, the expected information yield of each individual question over all the possible answers/outcomes (i.e., the entropy before and after the question is answered) is computed and weighted by the likelihood of each outcome. The information yield of a single question is thus given by the sum of differences between initial and updated entropy:

$$\Delta H = \sum \left[H(KU_i) - \left(p_k H'_k(KU_i) + p_{\neg k} H'_{\neg k}(KU_i) \right) \right] \quad (1)$$

where $p_k H'_k(KU_i)$ is the entropy of the updated knowledge structure given the user knows KU_i (weighted by the likelyhood factor p_k that KU_i is known), and where $p_{\neg k} H'_{\neg k}(KU_i)$ is the converse.

The question that has the maximum expected information yield is chosen as the most informative one. Both the random and entropy-driven methods of selecting a question are tested in this study.

IMPLEMENTATION

We have implemented the knowledge-structure induction procedure and the evidential reasoning scheme described in the previous section. The resulting knowledge assessment engine, UKAT (User Knowledge Assessment Tool), is composed of a set of C library routines. The observation gathering module of UKAT allows the questions to be selected either randomly or non-randomly. The algorithmic details of the modules can be found in [11]. The UKAT interface was developed in X-windows using the Motif widget package. In addition to presenting questions, acquiring answers (either text or choices), and building fine-grain models, it also permits users to browse and select the previously

[1] In [2] for instance, observation of text-editing methods by a plan recognition module indicated which commands were mastered by the user.

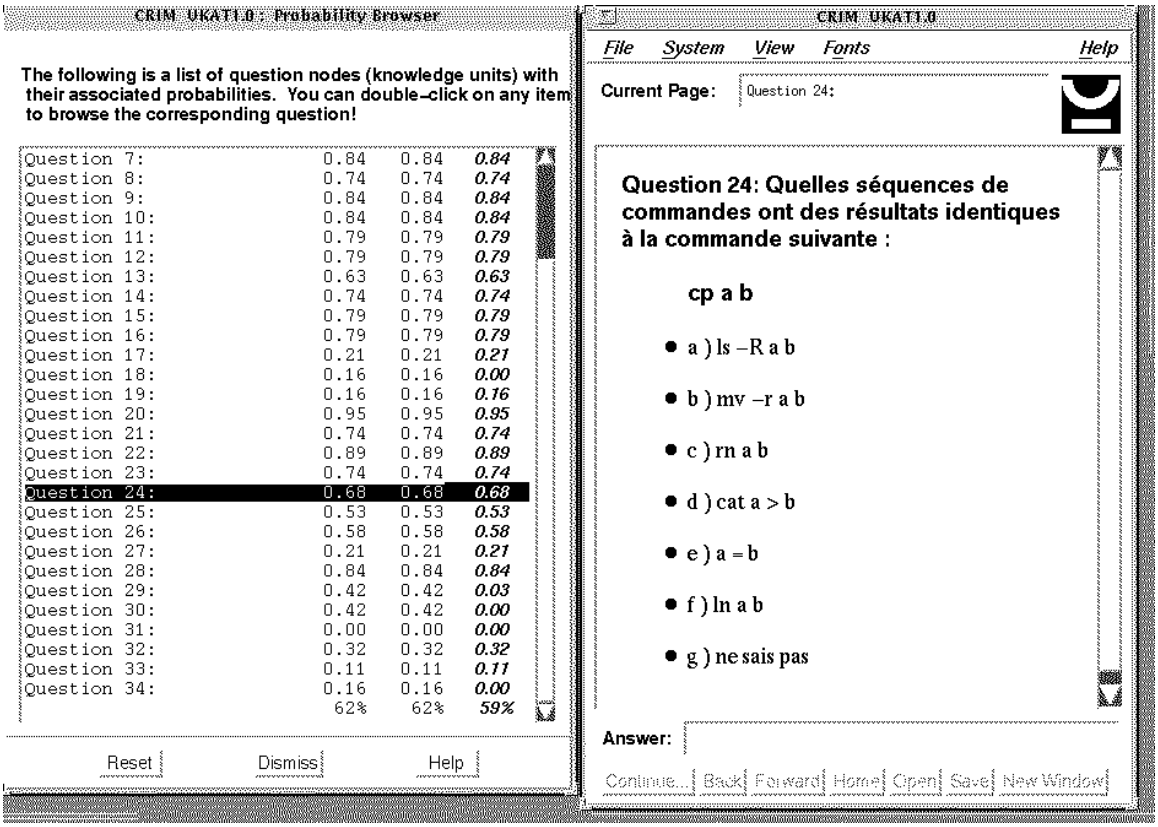

Figure 1: The interface of UKAT – A User Knowledge Assessment Tool

(un)answered questions and to look up scores (see Figure 1). Hence, the interface can easily be extended into a drill-and-practice training system.

EXPERIMENTAL RESULTS

In order to validate our modeling technique, we have chosen the WordPerfect[TM2] text editor as a domain of which users' knowledge states are to be evaluated. This domain knowledge is composed of 192 identified KUs (knowledge units) that correspond the mastery of WordPerfect commands (see [5] for details).

Empirical Data

The implication relations (i.e., gradation constraints) among the KUs are generated with the knowledge structure construction method, applied to a number of empirically obtained subjects' knowledge states. We used 47 sample knowledge states; each of them is the result of a test covering all 192 KUs. All 192×192 pairs of KUs were tested for implication relations, using the statistical criterion mentioned earlier (where $P(B \mid A) > 0.5$ and $P(\neg A \mid \neg B) > 0.5$ with $\alpha < 0.05$). As a result, 2,368 implications were included in the knowledge structure.

In order to test the derived knowledge structure, we performed a series of user modeling simulations. Each simulation run consisted of selecting (in either a random or an entropy-driven fashion) a proportion of a subject's knowledge state and propagating evidence to update the probability that a certain KU was mastered by the subject. Prior to the inferencing, all the nodes of the knowledge structure were assigned initial beliefs (i.e., initial probabilities). The results of simulations, based on 26-subject test data, are described below.

Building Fine-Grain User Models: The Knowledge Unit Level Assessment

The first level of testing corresponds to the residual errors in individual KUs assessment. Each KU is associated a weight that indicates the probability of knowledge and which is updated after each new observation (evidence). The difference between those weights and the actual knowledge state represents the residual errors.

The basic measure of performance is the global standard error of estimate:

$$\sigma_{node} = \sqrt{\frac{\sum_{i=1}^{26} \sum_{j=1}^{192} \left(x_{\text{obs}_{ij}} - x_{\text{est}_{ij}}\right)^2}{N_s \times N_k}} \qquad (2)$$

where N_k is the number of knowledge units (192). N_s is the number of subjects used for the test (26). $x_{\text{obs}_{ij}}$ is 1 if the corresponding KU_j is *known* and 0 otherwise, and $x_{\text{est}_{ij}}$ is the estimated probability of mastery.

In addition to the standard error of estimate which is an indicator of the dispersion around the estimate based on the second moment, we also investigated the bias around the estimate, which is given by the first moment:

$$\frac{\sum_{i=1}^{26} \sum_{j=1}^{192} \left(x_{\text{obs}_{ij}} - x_{\text{est}_{ij}}\right)}{N_s \times N_k} \qquad (3)$$

[2]WordPerfect is a trademark for WordPerfect Inc.

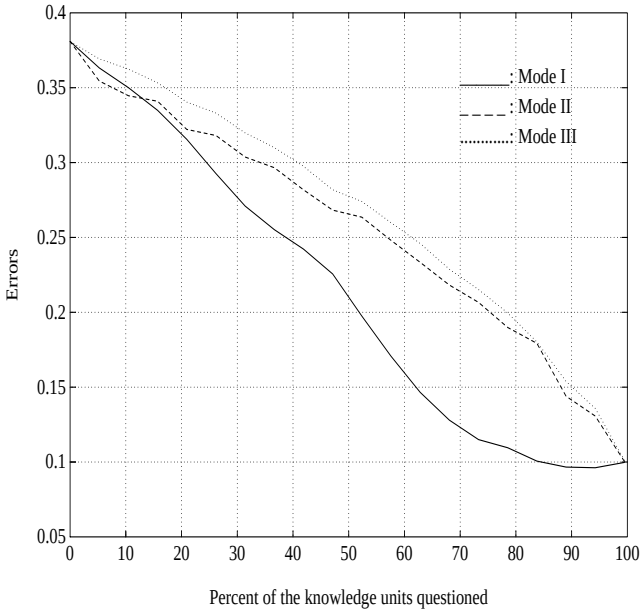

Figure 2: The residual errors in estimating probability of knowledge of individual KUs as measured by the *standard errors of estimate*. It shows the uncertainty convergence in three diffferent operating modes. The solid line, the dashed line, and the dotted line correspond to mode I (entropy-driven evidence gathering), II (random evidence gathering), and III (no inference condition), respectively.

The results of the system's performance in three different simulation modes are displayed in Figures 2 and 3. These show, respectively, the standard error scores and the bias over 192 KUs and 26 subjects. The three simulation modes are:

(I) *inferences based on the entropy-driven question selection*: nodes were given their initial probabilities and, when the chosen question is asked (based on entropy minimization), they are assigned 0.9 for a successful answer and 0.1 otherwise, and inference propagation is performed around the node according to the Dempster-Shafer algorithm;

(II) *inferences based on random sampling of the questions*: same as (I) but questions are chosen at random; and,

(III) *no inference condition*: same as (II) but no inference propagation is performed.

Note that we have assigned 0.9 and 0.1 weights for successful and unsuccessful answers respectively to reflect the

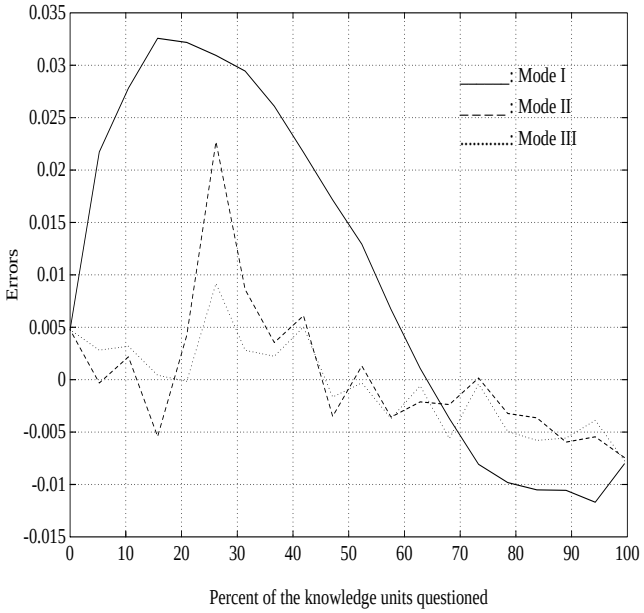

Figure 3: Absolute errors in estimating probability of knowledge of individual KUs, as measured by the *sum of the differences* between the observed KU mastery and its estimated probability. The results show a clear bias for the entropy-driven method (mode I).

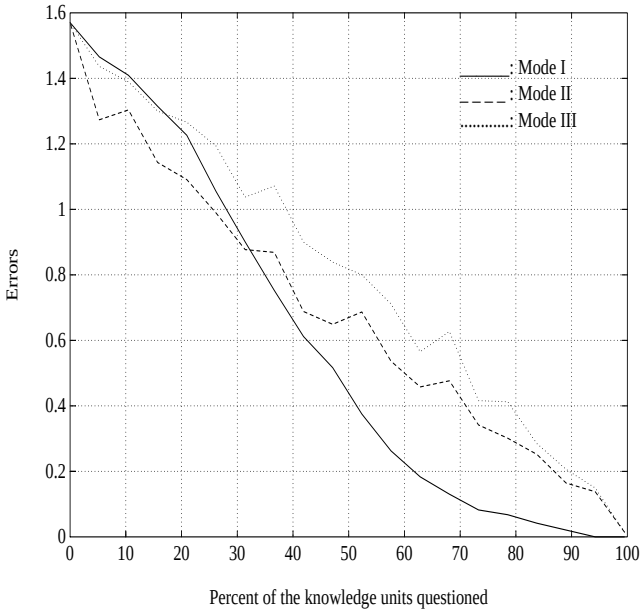

Figure 4: The residual errors in estimating global knowledge states as measured by the *standard error* scores of the three operating modes.

residual uncertainties associated with such a process (e.g. good answers by chance and bad errors by mistake). Consequently, the expected score at 100% observation is below the perfect score since the nodes' weights are matched against 1 and 0 and not against 0.9 and 0.1.

The results from Figure 2 clearly indicate that the entropy-driven approach (mode I) is more efficient in reducing the standard error of estimate. However, Figure 3 shows that below 60%, it has a tendency to underestimate the probability of knowledge. This tendency is slightly reversed above 70%. A plausible explanation for this behavior is that the system asks questions that are likely to be failed at the beginning, leaving more successful questions for the end. Although this order does not introduce any error in the standard error score, it does introduce a bias when we simply sum up the differences between observed and estimated values. This phenomenon is further discussed later.

As far as mode II is concerned, Figure 2 indicates that it is consistently better than mode III, but less efficient in reducing the errors than mode I, especially when the amount of observation is greater than 20%. It is not subject to a bias as mode I is, however.

Estimating User Overall Scores: The Knowledge State Level Assessment

We have studied the performance of the system at estimating whether or not each individual KU is known by a subject. Another useful application of our technique is to guess an individual's *overall score*, as it is often required in adaptive training systems.

The performance scores are given by a similar measure as KU level assessment, namely the standard error of estimate: $\sqrt{\frac{\sum (x_{\text{obs}} - x_{\text{est}})^2}{N_s}}$. However, x_{obs} and x_{est} represent, respectively, $\sum_{i=1}^{192} x_{\text{obs}_i}$ and $\sum_{i=1}^{192} x_{\text{est}_i}$, where x_{est_i} is 0 if $KU_i < 0.5$ and 1 otherwise.

Figure 4 show the residual errors of the performance. A bias similar to the one in Figure 3 was also found in this experiment, although those results are not shown here for brevity. Figure 4 indicates that both modes I and II are consistently better than mode III in guessing the scores (except mode I with less than 16% observation). However, the advantage of mode I over mode II is only manifest after 40% sampling.

Discussion and Further Results

As conjectured by previous researchers [7], the above results have provided an empirical demonstration that the minimum entropy inference is effective in reducing the inferences' standard error, or uncertainty about a knowledge state. However, an interesting finding of this investigation is that the minimum-entropy approach induces a bias at the knowledge state level score and thus may not be appropriate for the purpose of inferring users' *global* scores, where this bias would manifest itself.

The explanation we have given for this bias lies in the fact that if the system asks the questions that will likely be failed at

the beginning, and those that will likely be answered correctly at the end, a bias will be found in the global knowledge state. This does not necessarily introduce more errors in the estimations of individual KU probabilities, i.e., the global knowledge state entropy. The cause for this behavior may stem from the fact that given the average knowledge state is 46% of all KUs, it is likely that there are more nodes closer to 0 than to 1, and thus a greater reduction in entropy will be achieved by bringing these nodes closer to 0.

Let us look at two typical individual cases in which the minimum-entropy *searching* behavior can best be illustrated. Figures 5 and 6 display scores of two subjects — one with the actual score lower than the average and the other with the actual score higher than the average. In Figure 5 (lower than average score), the entropy method is shown to do much better than the other two methods, since lowering weights is the right direction for reducing uncertainties of the knowledge state, whereas in Figure 6 (higher than average), the entropy method starts lowering the weights only after observing more than 16% of the total KUs.

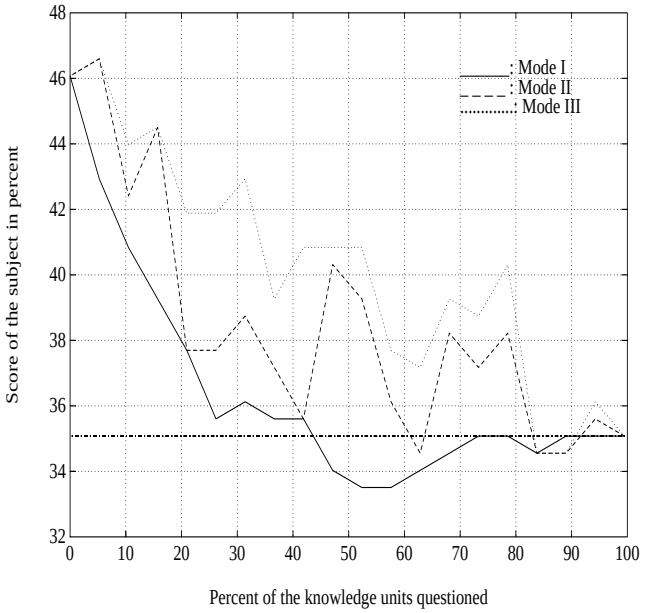

Figure 5: The estimated *scores of knowledge state* for subject 1. The true subject's score is 35% (shown at 100% of KUs asked) and the average of all subjects is 46% (shown at 0%).

CONCLUSION

The results presented in this paper indicate that in general, the approach to knowledge assessment is efficient in inferring new information and deriving the knowledge states of users, given either random or non-random observations. Two important applications of this technique in the context of adaptive user interfaces have been addressed. One involves building fine-grain user models; the other is guessing users' overall scores or levels of expertise. While both random and minimum-entropy-based observations are useful and effective, the latter appears to affect the order of questions asked and consequently the average knowledge state scores. In other words, the entropy method is the best for reducing uncertainties, but may generate a bias in assessing global user knowledge at the beginning. This result has rejected the previously held, rather intuitive, belief that the entropy minimization technique is a monotonically informative solution to the assessment/diagnostic problems.

ACKNOWLEDGEMENTS

We are grateful to Leslie Daigle and Jeanne-Estelle Thebault for their helpful comments on previous drafts of this paper. Jean-François D'Arcy is the author of the interface to the UKAT software.

REFERENCES

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

[2] Michel C. Desmarais. *Architecture et fondements empiriques d'un système d'aide assistée par ordinateur pour l'édition de texte*. PhD thesis, Université de Montéal, Département de psychologie, 1990.

[3] Michel C. Desmarais, Luc Giroux, and Serge Larochelle. Fondements méthodologiques et empiriques d'un système consultant actif pour l'édition de texte : le projet EdCoach. *Technologies de l'information et société*, 4(1):61–74, 1992.

[4] Michel C. Desmarais, Luc Giroux, Serge Larochelle, and Serge Leclerc. Assessing the structure of knowledge in a procedural domain. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, pages 475–481, 14–17 August, Montréal 1988.

[5] Michel C. Desmarais and Jimming Liu. Experimental results on user knowledge assessment with an evidential reasoning methodology. In *International Workshop on intelligent user Interfaces*, pages 223–226, 1993.

[6] Jean-Claude Falmagne and Jean-Paul Doignon. A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, 41:1–23, 1988.

[7] Jean-Claude Falmagne, Jean-Paul Doignon, Mathieu Koppen, Michael Villano, and Leila Johannesen. Introduction to knowledge spaces: how to build, test and search them. *Psychological Review*, 97(2):201–224, 1990.

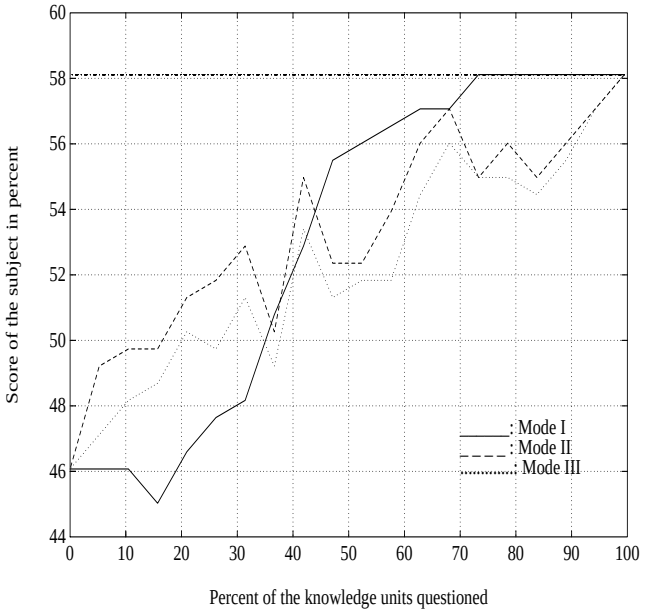

Figure 6: The estimated *scores of knowledge state* for subject 15. The true subject's score is 58% (shown at 100% of KUs asked) and the average of all subjects is 46% (shown at 0%).

[8] I.P. Goldstein. The genetic graph: a representation for the evolution of procedural knowledge. In Sleeman and Brown [15], pages 51–77.

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

[10] A. Kobsa. Modeling the user's conceptual knowledge in BGP-MS, a user modeling shell system. *International Journal of Man-Machine Studies*, 6:193–208, 1990.

[11] Jiming Liu and Michel C. Desmarais. Knowledge assessment based on the Dempster-Shafer belief propagation theory. Technical Report CRIM–92/09–06, Centre de recherche informatique de Montréal, 1992.

[12] J. Self. Student models: what use are they? In P. Ercoli and R. Lewis, editors, *Artificial intelligence tools in education*. North-Holland, Amsterdam, 1988.

[13] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N. J., 1976.

[14] D. Sleeman. UMFE, a user modeling front end subsystem. *International Journal of Man-Machine Studies*, 23:71–88, 1985.

[15] D. Sleeman and J.S. Brown, editors. *Intelligent Tutoring Systems*. Academic Press, London, 1982.