An Efficient Student Model Based on Student Performance and Metadata

Arndt Faulhaber and Erica Melis[1]

Abstract.
This paper describes a new student model technology that combines evidences and knowledge about pedagogical and domain structure. Its structure is generated from the metadata available in the content representation of the adaptive web-based learning platform ACTIVEMATH (or other contents). The evidences are processed with Item Response Theory and Transferable Belief Model uncertainty methodologies. We summarize evaluation results for this student model.

1 Introduction

The main goal of user modeling has been to facilitate adaptivity by estimating values of user variables. In the context of adaptive learning environments, the student model estimates dynamic variables, such as competencies and affective state. In the web-based learning environment for mathematics ACTIVEMATH [9], the variable *competency* determines, e.g., the difficulty of exercises the system adaptively selects for the student. Furthermore, when a course is generated for an individual student, the course planner may add missing prerequisites, if the student is insufficiently competent.

For web-based learning systems, collaborative authoring and modifications of content are more frequent than for traditional intelligent tutoring systems, which rely on a fix topic map. For such systems, the structure of the student model has to change according to alterations in the ontologies caused by content changes.

Manually constructed Bayesian Networks (BNs) and even automatically constructed BNs [2, 6, 1] cannot adapt in real-time to such changes, since they rely on previously observed data to infer their probability tables.

In order to cope with the potentially modified (implicit) ontology we devised the *semantics-aware student model* (SLM). It dynamically extracts relations and metadata from the ontology and makes use of their semantics. These, together with data from exercise interactions, enable SLM to estimate students' competencies.

The second and more conceptual contribution is a new processing of the evidences for competencies by a combination of *Item Response Theory* (IRT) [8] and *Transferable Belief Model* (TBM) [15], two mechanisms to reason about uncertainty. Finally, the new student model scales and efficiently and accurately computes values, which is important for real world applications.

Following, we describe how SLM interprets student exercise interactions in order to estimate the student's competencies. We start with a description of essential metadata, continue by detailing how the metadata is interpreted to allow evidences to be quantified by IRT and thereupon update beliefs in the TBM . Finally, the procedure of competency estimation from current beliefs within TBM is detailed and the utilization of semantic relations within the domain ontology for enhancing estimations. A summary of evaluations then briefly outlines the model's performance by comparing it to another student model, XLM [7], which also uses TBM, and by determining the gain of employing semantic relations. Eventually, we discuss related work and conclusions.

2 The New Student Model

The structure of SLM consists of nodes, each for a single rule/concept to estimate competencies for. Inter-node relations are dynamically extracted from the domain ontologies. Incoming evidences from students' exercise performance are processed by IRT for interpretation and to quantify proficiency probabilities, which are then taken as beliefs into TBM to compute, combine and update competency estimations (represented by the nodes). Propagation along the relations adds additional information for the competency estimation.

SLM automatically creates a node for each con-

[1] Saarland University / DFKI, Germany, email: Arndt.Faulhaber@dfki.de, Erica.Melis@dfki.de

cept/rule k included in the learning content such as the concept 'definition of fraction' or the rule 'addition of fractions with unlike denominators'. SLM stores each associated competency value $m(k, p)$ within the node, where a competency is defined as a pair (k, p), in which p is a cognitive process, such as *apply an algorithm* or *model* a mathematical problem that is applied to k.

For each competency, beliefs about the competency values are computed separately from recent evidences. In this computation, relations between exercises and concepts/rules as well as exercise competency metadata determine to which nodes evidences are attributed.

The combination of IRT and TBM was chosen in order to set SLM on a reliable theoretic foundation, avoid ad hoc mechanisms and to make it as simple, flexible and extensible as possible.

2.1 Metadata and Ontology

The content of learning systems such as ActiveMath consists of learning objects including concepts, rules, and auxiliary learning objects such as explanations, examples and exercises that are related to certain concepts/rules by the *for* relation [16]. The learning objects are enriched with administrative, domain-specific and educational metadata [4]. The educational metadata define, among other things, relations between learning objects such as a *prerequisite* relation. Together with their metadata, the learning objects induce an implicit ontology.

The metadata related to competencies can, e.g., be aligned with the PISA [10] specification for mathematics 'competencies' including *think, argue, model, solve, represent, language* and *tools*. The exercise metadata also include a *competency level* (CL) – a measure of the level of expertise needed to correctly solve an exercise. In PISA, values range from *elementary* (CL_I), *simple conceptual* (CL_{II}), *multi-step* (CL_{III}), to *complex* (CL_{IV}).

Additionally and conforming to the metadata standards IMS and LOM, exercises are assigned a difficulty metadata. The difficulty value is one of {*very easy*, *easy*, *medium*, *difficult*, *very difficult*}. These values represent the difficulty for a student having the same competency level as specified for the exercise. I.e., it should be easier to solve for more advanced students and harder for less proficient students.

2.2 Beliefs

The choice of the belief model heavily influences a student model. Belief models include Bayesian Networks [12], fuzzy logic [17], the Dempster-Shafer theory of evidence (DST) [14], and the TBM. TBM is an interpretation of DST and has been developed by Smets [15].

TBM works with a set of hypotheses to which *belief masses* are attributed. These are derived from the interpretation of evidences. Belief masses of different evidences are combined by the *Dempster-Shafer rule of combination*.

In SLM, we model beliefs about a competency value, called *competency mastery*[2], as hypotheses in TBM. We define a set of atomic hypotheses, $H(m_j), 0 \leq j \leq 34$ that encode "Student has mastery $m_j = j * 0.03$". For instance, $H(0.12)$ means the student has mastery 12% and so forth. That means we consider hypotheses in 3% steps starting at 0% and going up to 102%. This optimal granularity was determined by simulation [3].

Since SLM regards a sequence of the most recent evidences, **Dynamic** Bayesian Networks would be required for similar modeling with Bayesian Networks. For such a model the conditional probability tables need to be manually constructed (extra effort) or learned from previous student interactions, for which - in our case - the amount of available training data is insufficient. Additionally, the notion of *ignorance* within TBM serves a more skeptical interpretation of evidences in the way that no further assumptions are made except for what has been directly observed (e.g. A is attributed $P(A) = 40\%$ TBM assumes nothing about the other 60%, BNs would quantify $P(\neg A) = 60\%$).

The interpretation of evidences and the resulting assignment of belief masses to these hypotheses is described in the following.

2.3 Evidences and Interpretation

The most important information for the student model is the student's performance in an exercise (step). We henceforth refer to exercise steps simply as exercise. It is called *achievement* in ActiveMath and has a value $ach \in [0, 1]$. A value of 1 indicates a correct answer, 0 a wrong one. Other values refer to partially correct input.

Evidences can be represented as a tuple $e = (s_{id}, ex_{id}, K, P, CL, d, ach)$, where s_{id} is the student's ID, ex_{id} is the exercise's ID, K is a set of concepts/rules trained by the exercise, P is a set of cognitive processes required to solve the exercise, CL is the competency level of the exercise, d is the difficulty of the exercise, ach is the student's achievement.

For each $k \in K$ and $p \in P$, e is converted to an *evidence* object that is attached to a triple $\tau = (s_{id}, k, p)$. One event can trigger the creation of multiple evidences, one for each triple.

To enable SLM to adapt to changes in the student's knowledge only the 6 latest evidences per triple τ are kept. Thus, old evidence does not influence the estimation of the current mastery. Additionally, indirect evidence is attached to each τ.

[2] In the following, in case the context disambiguates we just call it mastery.

Indirect evidence is computed by a propagation algorithm that sends evidences up and down along prerequisite relations as described in §2.6. Based on the set of available evidences E for a triple τ, SLM estimates the mastery $m(\tau|E) \in [0, 1]$.

In order to interpret evidences, IRT provides a simple model to relate exercise difficulties to proficiencies in terms of probabilities for correct answers. Here, these probabilities serve to update beliefs, stored, combined and computed by TBM. IRT is a data-driven statistical model that has been employed for adaptive testing etc. for some decades and proved to be a reliable technique.

In IRT the probability $Pr_i(X \mid \theta)$ for a correct response X to item (question) i based on the test person's proficiency θ is modeled as follows:

$$Pr_i(X = correct \mid \theta) = c_i + \frac{1 - c_i}{1 + e^{-D*a_i*(\theta - b_i)}}, \quad (1)$$

where a_i is the item discrimination factor, b_i the item location, c_i the item guessing probability and D a constant. The resulting function is called *item characteristic curve* (ICC) and has the form of a sigmoid function. The guessing parameter is modeled as a lower bound probability for a correct response and the discrimination factor determines the curvature of the function. In our context, θ is m and b_i is derived from (CL, d) as indicated in Fig. 1. The parameters a_i and c_i can be fitted by analyzing previous performance statistics and D is chosen to scale the item characteristic curve (see §3).

In the absence of previous performance statistics, the parameter b_i for equation (1) is derived from the content metadata: competency level and difficulty are combined (both convey information about the absolute difficulty of an exercise) into an item location b_i used in IRT.

Let $\Lambda = (CL_I, CL_{II}, CL_{III}, CL_{IV})$ and $\Delta =$ (very easy, easy, medium, difficult, very difficult) be ordered sets of competency levels and difficulties with elements Λ_q and Δ_r respectively, both with indices starting at 0. To quantify b_i (with associated Λ_q and Δ_r for item i), we define:

$$b_i(\Lambda_q, \Delta_r) = 0.2 * q + 0.1 * r. \quad (2)$$

The interpretation of competency level and difficulty imposes an equidistant placement on the item location scale. The pairs (Λ_q, Δ_r) are mapped to item locations $b_i \in [0, 1]$ in correspondence with the value space of the mastery (see Figure 1).

Moreover, we have to map the achievement of the student to the dichotomous variable X in the dichotomous IRT model. This mapping is defined for achievement *ach* as $X = correct$, if $ach \geq 0.5$ and $X = \neg correct$ otherwise.

On the basis of IRT, SLM derives belief masses as follows: belief mass $mass(H(m))$ is assigned to each atomic hypothesis $H(m)$ equivalent to the probability derived from equation (1) $mass(H(m)) = Pr_i(X = correct \mid \theta)$, if $X = correct$, and $mass(H(m)) = 1 - Pr_i(X = correct \mid \theta)$, if $X = \neg correct$, with $m = \theta$.

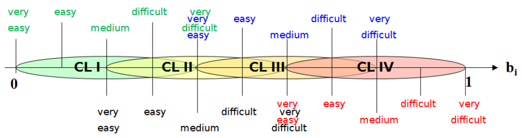

Figure 1. Mapping: competency level and difficulty to item location b_i.

2.4 Calculating the Mastery

The mastery is computed at two levels: (a) the competency mastery at the level of competencies, and (b) *concept mastery* as aggregation of all competency masteries of a specific concept.

To compute a mastery $m(\tau|E)$, SLM derives belief masses from each $e \in E$ with IRT and combines the masses using TBM. Then, it chooses the hypothesis H_{max} with the maximal belief mass. The mastery associated with H_{max} is the current estimation or 100% if the mastery is estimated above 100%. Currently, if multiple hypotheses have the same maximal mass, the mean of their associated masteries is calculated and taken as the estimation.

SLM computes the concept mastery for a specific concept by the weighted sum of the competency masteries for which an estimation is available. The weights are determined by the amount of evidence available for the single competencies, though it would be desirable to make use of relations (probably a partial order) between the cognitive processes.

2.5 Information Radius

Deriving masses for all atomic hypotheses as described may lead to the following undesirable situations: a) The mastery estimation upon first evidence is either 0% or 100% regardless of the associated difficulty, because the ICC is increasing strictly monotonic and hence, maximal mass is either attributed to $H(0)$ or $H(1.02)$ (depending on the achievement). Therefore, the adaptive system may confront the student with an overly hard (or easy) exercise in the next step. b) If for any available positive evidence[3] a negative evidence exists with the same difficulty and vice versa, the unweighted combination of all evidences leads to the same belief mass for all hypotheses. This results in a mastery of 51%, regardless of the difficulties of the exercises from which the evidences originate.

We solve a) by assigning belief masses only to hypotheses that are close to the item location, namely for hypotheses with $\mid \theta - b_i \mid < \delta$. For instance, if $b_i = 0.1$ and $\delta = 0.1$, SLM computes

[3] A positive evidence is derived, iff $ach \geq 0.5$, else a negative evidence is derived.

masses for $H(m), 0 \leq m \leq 0.2$. The first problem has thus been addressed, since an estimation based on a student's performance for exercises with maximal item location b_i cannot exceed mastery $b_i \pm \delta$. We call δ the *information radius*.

Introducing δ also solves b) because either the different hypotheses receive different masses (if the difficulties vary), or not all hypotheses receive a mass because of the restriction introduced by δ, resulting in a mastery close to the mean difficulty. E.g., if for two evidences (one positive, one negative) $b_i = 0.3$ and $\delta = 0.2$, we get $mass(H(m_k)) = mass(H(m_l)) \forall k, l : 0.1 < m_k, m_l < 0.5$. The estimated mastery is 0.3 instead of 0.51.

Actually, SLM uses a variable value for δ, which increases with more available evidence. A simulation using virtual students showed, that this approach further reduces the estimation error [3].

2.6 Propagation

Indirect evidence informs an educated guess about the mastery in the absence of direct evidence. It is disregarded as soon as enough (a configurable parameter) direct evidence is available. SLM instruments concept and rule inter-item-prerequisite relations to *propagate* estimations. The rationale is that students will not be able to solve exercises for which they do not understand the prerequisites (to a certain degree) and, conversely, if they can solve an exercise they most likely understand its prerequisites. For instance, if the student has difficulties adding two fractions with the same denominator, it is rather unlikely that she will be able to add fractions with unlike denominators.

3 Evaluation

We quantitatively compare the new student model with the XLM (with propagation) wrt. the accuracy of predicting the correctness of student answers and computing performance. We evaluate the prediction accuracy by replaying real usage data from log-files.

This data has been collected from two evaluations conducted at the university of Edinburgh. One was performed in April 2007 with 42 students and the second in May 2007 with 46 students. All students had to solve calculus exercises.[4]

Our approach to compare the student models is to predict the students' result in the next exercise based on the current competency estimation of the student model. This is similar to [5], however, differs in that our student model is supposed to learn after each input instead of prior (off-line) training. The idea is to compute the probabilities for a correct/incorrect answer for the next exercise and to choose the most probable outcome. After each prediction, the evidence is passed to the student model to recalculate the mastery estimation.

The prediction of the next result is based on the student model estimate m and the item location b_i of the exercise. The ICC, with values $c_i = 0$, discrimination factor $a_i = 1$ and $D = 10$, then determines the probability for a correct answer.[5]

The prediction for the next achievement in exercise i_{t+1} at time t is simply 1 if $b_{i_{t+1}} \leq m_t$, where m_t denotes the mastery of the required competency. Otherwise, the predicted achievement is 0.

The results for predicting the next outcome of an exercise are shown in Figure 2. The dimensions are the number of direct evidences processed by the student model at the time of prediction and the prediction accuracy. On average, propagation increases the accuracy of SLM significantly from 65.8% ($\sigma = 1.4\%$) to 71.1% ($\sigma = 1.2\%$). XLM's average accuracy is at 68.5% ($\sigma = 1.3\%$).

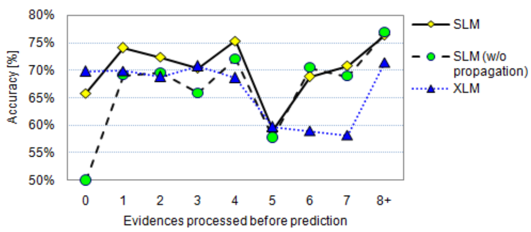

Figure 2. Prediction accuracy.

Processing speed is essential because updating an inspectable student model as well as adaptation decisions have to be made in real-time. The average time to process an exercise step event is 9.74 seconds for XLM and 0.03 seconds for SLM, while reaching similar estimation accuracy.

4 Related Work

Currently, most student modeling research uses (Dynamic) Bayesian Networks (see [2, 5, 11]). For instance, Desmarais et al. [2] introduce an efficient algorithm to construct learner models based on BNs by exploiting constraints imposed by item to item structures. For the construction it relies on previously observed system usage.

Using belief masses on a per competency basis saves additional modeling effort and the estimation of conditional probabilities needed when employing Bayesian Networks. It allows to distribute evidence to multiple competencies in case the assignment is unclear and to weigh evidences based on additional parameters.

Recently, IRT has found its way into some student models. For instance, Johns et al. [5] combine

[4] Since most exercises were fill in the blanks, guessing should have a minimal effect.

[5] Lacking a priori knowledge about the discrimination properties of the exercises, we chose a constant value of $a_i = 1$. D fits the probabilities to the interval [0, 1] corresponding to the mastery values.

it with a Dynamic Bayesian Network to take students' engagement into account, during their use of the tutoring system *Wayang Outpost*. Again, the student model is pre-trained with log data.

A student model that is also based on TBM is XLM [7]. It uses TBM to estimate the competency level of a student and indirectly derives the mastery by combining the belief masses attributed to each competency level. For propagation, it uses a static topic map. Comparing the SLM with the related XLM shows that SLM clearly outperforms the XLM in terms of computing performance, while delivering a slightly better prediction accuracy (see §3). In contrast to XLM, SLM supports parameters for item discrimination, guessing, and continuous difficulty values, which can be calibrated using data-mining techniques (to overcome the ad hoc derivation of difficulty from content metadata and personal bias of the author).

5 Conclusion and Future Work

Our contribution has been to build a student model, whose structure is automatically generated and that dynamically adapts to changes in the implicit ontology. Its performance is adequate for real-time applications regarding both: estimation accuracy and processing speed. The clear separation between student model and domain structure reduces modeling and coordination effort.

Furthermore, the design of SLM allows for taking into account the discrimination properties of an exercise and a continuous value for difficulty. A proof of the benefit of using these properties still has to be given, since the amount of data was not sufficient to estimate these parameters. By combining IRT and TBM in a relatively simple way, we gain flexibility to vary the interpretation of evidences (e.g. modifying information radius or propagation parameters).

The use of indirect evidence serves two goals. For one, it provides information for an initial estimation, in case no direct evidence is available. Additionally, it improves the estimation accuracy for insufficient direct evidence hence relaxing the sparse data problem. The propagation mechanism is based on the assumption that the student's mastery of a concept influences the mastery of closely related concepts. A more sophisticated mechanism would differentiate between different relations and takes into account conditions for a propagation.

A promising way to enhance the accuracy of the model's estimations is mining log files to adjust the IRT-parameters of exercises. Hence, we plan to calibrate the student model automatically based on previous interactions with students.

Acknowledgements

This article results from the ATuF (ME 1136/5-1) and LeActiveMath (FP6-507826) projects, funded by the German National Science Foundation (DFG) and the European Union, respectively.

REFERENCES

[1] I. Arroyo and B.P. Woolf, 'Inferring learning and attitudes from a bayesian network of log file data', in *Proceedings of the* 12^{th} *International Conference on Artificial Intelligence in Education*, pp. 33–40, (2005).

[2] M.C. Desmarais and M. Gagnon, 'Bayesian student models based on item to item knowledge structures', in *EC-TEL 2006*, eds., W. Nejdl and K. Tochtermann, volume 4227 of *LNCS*, pp. 111–124. Springer-Verlag, (2006).

[3] A. Faulhaber, *Building a new Learner Model for ActiveMath Combining Transferable Belief Model and Item Response Theory*, Master's thesis, Saarland University, 2007.

[4] G. Goguadze, C. Ullrich, E. Melis, J. Siekmann, C. Groß, and R. Morales, 'Leactivemath structure and metadata model'. Deliverable D6 for LeActiveMath project, 2004.

[5] J. Johns and B.P. Woolf, 'A dynamic mixture model to detect student motivation and proficiency', in *Proceedings of the* 21^{st} *National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA, (2006).

[6] A. Jonsson, J. Johns, H. Mehranian, I. Arroyo, B.P. Woolf, A. Barto, D. Fisher, and S. Mahadevan, 'Evaluating the feasibility of learning student models from data', in *AAAI Workshop on Educational Data Mining*, ed., J. Beck, pp. 1–6, (2005).

[7] N. Van Labeke, P Brna, and R. Morales, 'Opening up the interpretation process in an open learner model', *International Journal of Artificial Intelligence in Education*, **17**(3), 305–338, (2007).

[8] F.M. Lord, *Applications of item response theory to practical testing problems*, Hillsdale, NJ: Erlbaum, 1980.

[9] E. Melis, G. Goguadze, M. Homik, P. Libbrecht, C. Ullrich, and S. Winterstein, 'Semantic-aware components and services of activemath', *British Journal of Educational Technology*, **37**(3), 405–423, (may 2006).

[10] *Learning for Tomorrows World – First Results from PISA 2003*, ed., OECD, Organization for Economic Co-operation and Development (OECD) Publishing, 2004.

[11] Z.A. Pardos, N.T. Heffernan, B. Anderson, and C.L. Heffernan, 'Using fine-grained skill models to fit student performance with bayesian networks', in *Educational Datamining Workshop, at ITS-2006*, pp. 5–12, (2006).

[12] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kauffman, 1988.

[13] G. Rasch. Probabilistic models for some intelligence and attainment tests. Copenhagen, Paedagogiske Institute, 1960.

[14] G. Shafer, *A mathematical theory of evidence*, Princeton University Press, Princeton, NJ, 1976.

[15] P. Smets and R. Kennes, 'The transferable belief model', *Artificial Intelligence*, **66**, 191–234, (1994).

[16] C. Ullrich, 'The learning-resource-type is dead, long live the learning- resource-type!', *Learning Objects and Learning Designs*, **1**(1), 7–15, (2005).

[17] L. Zadeh, 'The role of fuzzy logic in the management of uncertainty in expert systems', *Fuzzy Sets and Systems*, **11**, 199–227, (1983).