Modeling Students' Memory for Application in Adaptive Educational Systems

Radek Pelánek
Masaryk University Brno
pelanek@fi.muni.cz

ABSTRACT

Human memory has been thoroughly studied and modeled in psychology, but mainly in laboratory setting under simplified conditions. For application in practical adaptive educational systems we need simple and robust models which can cope with aspects like varied prior knowledge or multiple-choice questions. We discuss and evaluate several models of this type. We show that using the extensive data sets collected by online educational systems it is possible to build well calibrated models and get interesting insight, which can be used for improvement of adaptive educational systems.

1. INTRODUCTION

Development of intelligent tutoring system and other adaptive educational systems is often focused on teaching mathematics, physics, and similar domains. The related research in student modeling is thus concerned mainly with modeling skill acquisition. Another interesting area, where adaptability is very useful, is learning of facts [8], particularly in domains with varied prior knowledge like vocabulary, geography, or human anatomy. In this context, modeling of students' memory is important.

Principles of human memory and their consequences for education have been extensively studied in psychology, e.g., [2, 5, 9, 10]. Models developed in the psychological research are not, however, easily applicable in practical implementation of adaptive practice. The purpose of models described in psychological literature is to describe and explain mechanisms of human memory, e.g., the spacing effect [9]. Experiments are done using lab studies under controlled setting, in areas with little prior knowledge, e.g., learning of arbitrary word lists, nonsense syllables, obscure facts, or Japanese vocabulary.

In the context of development of adaptive educational systems, our goal is more pragmatic – we do not need to capture all details of human memory, we need a model which will work well in an adaptive system. A model needs to provide good input for other modules of an adaptive system (e.g., question selection or open learner model). The specific context of our work is an adaptive application `slepemapy.cz` for learning geography [8].

Although we can afford to model memory in a simplified manner, we have to deal with issues like varied prior knowledge, multiple-choice questions (with possibility of guessing), and no control on when students use the system. Compared to laboratory studies online educational systems can easily collect much more extensive data (millions of answers), so we can employ machine learning techniques to find fitting models. Specifically, in our work we use this approach to detect the dependence of memory activation on time from previous answer. The standard approach [9] is to make an assumption about the functional form of such dependence. We learn the function from the data and it turns out to be an S-shaped function which cannot be represented symbolically in a straightforward way. The results also show that there are large differences between learning of facts even in a seemingly compact domain like geography. These results may be useful for improving the behaviour of adaptive educational systems.

2. MODELING

Before we go into the description of models, let us clarify the context of considered models. In previous work [8] we described a modular architecture for an adaptive practice of facts based on three modules: estimation of prior knowledge, estimation of current knowledge, construction of questions. Here we focus on improving the estimation of current knowledge by taking timing between answer into account.

Specifically, we assume the following input: for each student and repeatedly answered fact (e.g., a country in the case of our application), we have an initial estimate of the student's knowledge of the fact and data about a sequence of student's answers. For each answer we consider the correctness of the answer, the type of question (either open question or multiple-choice question with a specified number of options), and time from previous answer (in seconds). For estimating initial activation we use a variant of the Elo rating system [4, 13] as specified in [8]. For purpose of this work this estimation is treated as a black box.

As an output a model provides estimated probability that the next answer will be correct. This output can be used for the adaptive construction of questions (in such a way that

they have appropriate difficulty) [7, 8]. Model parameters can be also used for presenting feedback to students in the form of an open learner model.

2.1 Basic Approach

Student models of learning [3] most commonly use either a binary skill (a typical model of this type is Bayesian Knowledge Tracing) or a continuous skill with probability of correct answer specified by the logistic function of the skill. For modeling memory it is natural to use a continuous skill since memory is build gradually – as opposed, for example, to understanding or insight in mathematics, which may undergo sudden transition from unlearned to learned state as assumed by Bayesian Knowledge Tracing [1]. Modeling based on the logistic function was also previously used for modeling memory [9]. In the following we use the notion of *memory activation* instead of skill.

All models that we consider have the following basic form. Based on the data we estimate memory activation m. Probability that the next answer will be correct is estimated using a logistic function: $P(m) = \frac{1}{1+e^{-m}}$. In the case of multiple-choice question with n options the probability of correct answer is given by the shifted logistic function: $P(m) = \frac{1}{n} + (1 - \frac{1}{n})\frac{1}{1+e^{-m}}$. Note that this functional form is a simplification, since it does not consider the possibility that a student answers correctly by ruling out distractors.

2.2 Computing Memory Activation

A basic model applicable under the outlined approach is a simplified, one-dimensional variant of Performance Factor Analysis (PFA) [11] (originally PFA was formulated in terms of skills and vectors, as it uses multiple knowledge components). In this model the memory activation is given by a linear combination of an initial activation and past successes and failures of a student: $m = \beta + \gamma s + \delta f$, where β is the initial activation, s and f are counts of previous successes and failures of the student, γ and δ are parameters that determine the change of the skill associated with correct and incorrect answers. The basic disadvantage of this simple approach is that it does not consider the time between attempts; in fact it even ignores the order of answers (it uses only the summary number of correct and incorrect answers).

ACT-R model [9, 12] of spacing effects can be considered as an extension of this basic model. In this model the memory activation is estimated as $m = \beta + \log(\sum b_i t_i^{-d_i})$, where the sum is over all previous attempts, values t_i are the ages of previous attempts, values b_i capture the influence of correctness of answers, d_i is the decay rate, which is computed by recursive equations [9]. The model also includes additional modifiers for treating time between sessions. The focus of the model is on modeling the decay rate to capture the spacing effect. Studies using this model [9, 12] did not take into account the probability of guessing and variable initial knowledge of different items (initial activation was either a global constant or a student parameter). In the current work we focus on these factors and for the moment omit modeling of spacing effects.

Another possible extension [8] of the basic PFA model is to combine it with some aspects of the Elo rating system [4, 13]; in the following we denote this version as PFAE (PFA Elo/Extended). The estimated memory activation is updated after each answer as follows:

$$m := \begin{cases} m + \gamma \cdot (1 - P(m)) & \text{if the answer was correct} \\ m + \delta \cdot P(m) & \text{if the answer was incorrect} \end{cases}$$

To include the timing information into this model, we can locally increase the memory activation for the purpose of prediction, i.e., instead of $P(m)$ to use $P(m + f(t))$, where t is the time (in seconds) from the last attempt and f is a *time effect function*. As m denotes memory activation, the value $f(t)$ corresponds to temporal increase in memory activation due to (short) time from previous exposure of an item.

It is natural to use as a time effect function some simple analytic function, but analysis of our data suggests that this approach does not work well. Figure 1 shows calibration analysis for two time effect functions: $f(t) = \frac{w}{t}$ (used in previous work [8]) and $f(t) = 1.6 - 0.1 \log(t)$ (the functional form is based on [9] and fitted to data). We see that neither of these functions leads to well calibrated predictions. Since we were not able to find a simple time effect function that would provide a good fit, we represent the function $f(t)$ as a staircase function with fixed bounds $\vec{b}$ and values $\vec{v}$ which we learn from the data:

$$f(t) = \begin{cases} v_i & \text{if } b_i \leq t < b_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

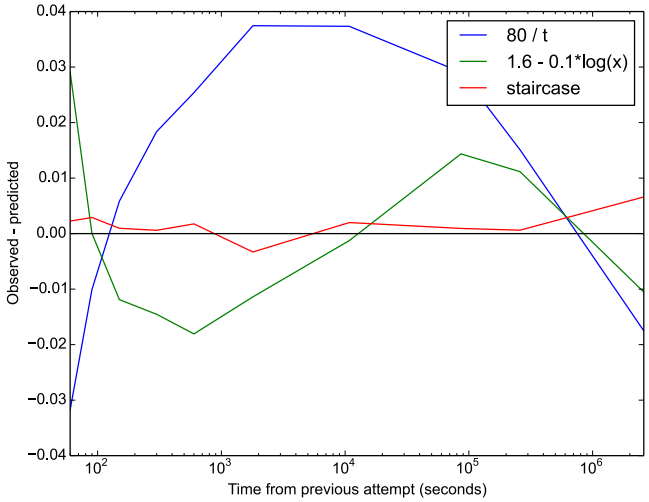

Figure 1: Calibration for the PFAE model with different time effect functions – the y axis shows difference between observed frequency of correct answers and average prediction.

3. EXPERIMENTS

We report experiments with the PFAE model with time effect function. For evaluation we used data from an online system for practicing geography [8] (`slepemapy.cz`). Data were filtered to include only students with at least 20 answers, items (places) with at least 40 answers, and we consider only sequences where a student answered at least 3 questions about an item. For experiments we divided the data into 10 sets, each containing 52,190 sequences of answers.

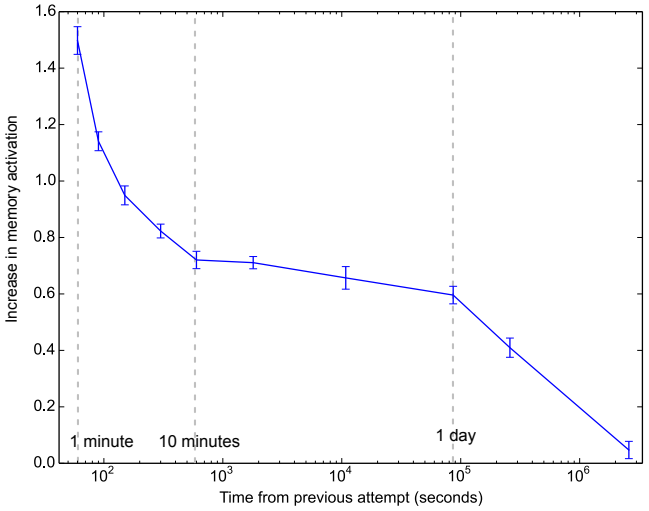

Figure 2: Time effect function – average from 10 independent data sets, error bars show standard deviations of parameter estimates.

3.1 Model Parameters

As the fixed bounds used in the staircase representation of time effect function we have chosen the following values: 0, 60, 90, 150, 300, 600, 1800, 10800, 86400, 259200, 2592000. These values were chosen to be easily interpretable (e.g., 30 minutes, 1 day) and at the same time to have reasonably even distribution of data into individual bins.

The model has the following parameters which we need to estimate from the data: update constants γ, δ and the vector $\vec{v}$ representing the time effect function. To estimate these parameters we use a gradient descent. To evaluate stability of parameter estimates we computed the parameter values for the 10 independent data sets. The results show that the obtained parameters are very stable: $\gamma = 2.290 \pm 0.042$, $\delta = -0.917 \pm 0.018$; values $\vec{v}$ for the representation of time effect function are depicted in Figure 2.

Since our data set is large and parameter estimates are stable, we can afford to do more detailed analysis. Figure 3 shows fitted time effect functions and γ, δ values when the parameters are fitted using only part of the data. Figure 3 A shows that there is quite large difference between parameter values for cases with high and low prior knowledge. This suggests possible improvement to the PFAE model – not just by including more parameters, but also by changing its functional form. However, prior knowledge is not the only factor that plays role. Figure 3 B shows fitted parameters for several types of places. In all of these cases the prior knowledge is low, yet there are still large differences between fitted parameters values. These parameters may contain useful information about students' learning in particular parts of the domain, e.g., data in Figure 3 B illustrate that it is easier to learn states of Germany than provinces of China.

In the case of countries we have enough data to perform parameter fitting for individual places. In this case we fix the time effect function (as learned on the whole data set and reported in Figure 2) and we learn only the γ, δ parameters on

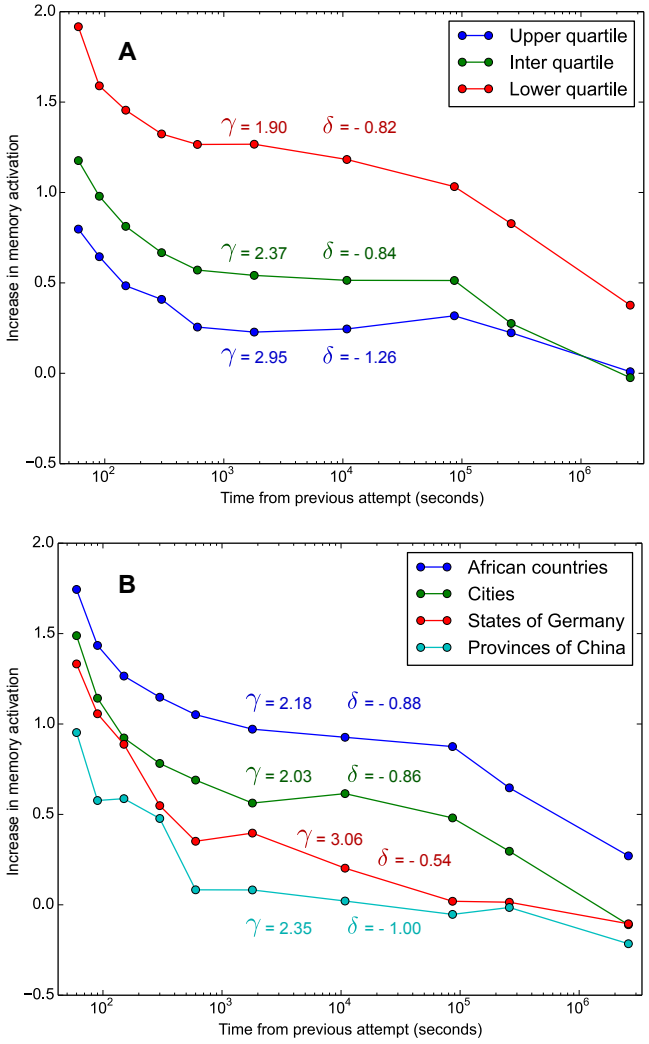

Figure 3: Time effect function and γ, δ parameters fitted to filtered data: A) by estimated prior knowledge, B) by the type of a place.

data for a single place. We use only places for which we have at least 1300 students answering at least 3 questions. The fitted parameter γ is has an interpretable meaning "how easy it is to remember a country". Examples of countries with high γ (>3.3): Western Sahara, Southern Sudan, Vietnam, Egypt, Somalia; countries with low γ (< 1.7): Bulgaria, Romania, Serbia, Moldova. Note that the reported results are clearly dependent on the origin of students using the system – in our case mostly Czech students.

3.2 Accuracy of Predictions

Table 1 show comparison of several model variants with respect to three common performance metrics [14]: root mean square error (RMSE), log-likelihood (LL), and area under the ROC curve (AUC). The results show averages from 10 runs on different training/testing sets. The results are consistent over the three metrics and show that the PFAE models brings quite large improvement over the PFA model. Differences between variants of the PFAE model due to the used time effect function are statistically significant, but other-

Table 1: Comparison of models with respect to three performance metrics.

model	time effect	RMSE	LL	AUC
PFA	–	0.3593	-106517	0.719
PFA	$80/t$	0.353	-103441	0.7195
PFAE	$80/t$	0.3377	-94454	0.757
PFAE	$1.6 - 0.1\log(t)$	0.3367	-93987	0.7591
PFAE	staircase	0.3363	-93642	0.7614

wise rather small. Individual predictions are actually highly correlated (correlation coefficient around 0.97).

4. DISCUSSION

We have evaluated several variants of a model of memory activation in the context of adaptive practice of facts. We proposed a model which incorporates the effect of time from previous answer by a general staircase function, which is learned from data (as opposed to assuming a specific symbolic form of the function). The model is better calibrated than other studied models and provides slightly better predictions. More importantly, the model is simple, parameters are easy to learn from data and robust. The learned function also provides interesting insight into students memory in the particular application – there is fast decrease in memory activation within the first 10 minutes, then the effect is nearly steady for 1 day, after that the activation decreases again.

By performing fine-grained analysis of the data, it is possible to use the model parameters to determine items that are easy or difficult to remember. Such results may be useful for improvement of educational systems, e.g., by offering mnemonics for difficult to remember facts, or by changing the adaptive selection of questions to prefer easy to remember facts at the beginning of a session. Specifically, results reported in Figure 3 suggest that different adaptive behaviour may be useful for learning African countries and provinces of China.

A possible limitation of this study is that the used data do not come from a properly designed and controlled experiment, but from an adaptive system which uses a student model to choose questions [8]. This may potentially cause a bias in the performed analysis. Although it seems unlikely that the reported results would be significantly influenced by this data source, feedback loops between student models and data collection deserve attention [6].

Another simplification of the current work is that we do not consider the feedback provided by the used system when a student answers incorrectly. This feedback clearly has impact on memory activation of the selected wrong answer. This raises a more general question: What is more important for the practical development of adaptive educational systems – proper treatment of principal issues (e.g., spacing effect) or incorporation of practical features into the model (e.g., effect of wrong answers)?

Acknowledgement

The author thanks Vít Stanislav and Jan Papoušek for their work on the `slepemapy.cz` project and for their assistance with the data.

5. REFERENCES

[1] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[2] Peter F Delaney, Peter PJL Verkoeijen, and Arie Spirgel. Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of learning and motivation*, 53:63–147, 2010.

[3] Michel C Desmarais and Ryan SJ d Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2):9–38, 2012.

[4] Arpad E Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.

[5] Jeffrey D Karpicke and Henry L Roediger. Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57(2):151–162, 2007.

[6] Juraj Nižnan, Jan Papoušek, and Radek Pelánek. Exploring the role of small differences in predictive accuracy using simulated data. 2015. Submitted.

[7] Jan Papoušek and Radek Pelánek. Impact of adaptive educational system behaviour on student motivation. In *Proc. of Artificial Intelligence in Education (AIED)*, 2015.

[8] Jan Papoušek, Radek Pelánek, and Vít Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining (EDM)*, pages 6–13, 2014.

[9] Philip I Pavlik and John R Anderson. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29(4):559–586, 2005.

[10] Philip I Pavlik and John R Anderson. Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14(2):101, 2008.

[11] Philip I. Pavlik, Hao Cen, and Kenneth R. Koedinger. Performance factors analysis-a new alternative to knowledge tracing. In *Proc. of Artificial Intelligence in Education (AIED)*, volume 200 of *Frontiers in Artificial Intelligence and Applications*, pages 531–538. IOS Press, 2009.

[12] Philip Pavlik Jr, Thomas Bolster, Sue-Mei Wu, Ken Koedinger, and Brian Macwhinney. Using optimally selected drill practice to train basic facts. In *Intelligent Tutoring Systems*, pages 593–602. Springer, 2008.

[13] Radek Pelánek. Time decay functions and Elo system in student modeling. In *Educational Data Mining (EDM)*, pages 21–27, 2014.

[14] Radek Pelánek. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 2015. To appear.