Pardos, Z. A., Heffernan, N. T. (submitted) Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In *Proceedings of the International Conference on User Modeling, Adaptation and Personalization*. Big Island of Hawaii.

Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing

Zachary A. Pardos[1], Neil T. Heffernan

Worcester Polytechnic Institute
zpardos@wpi.edu, nth@wpi.edu

Abstract. The field of intelligent tutoring systems has been using the well known knowledge tracing model, popularized by Corbett and Anderson (1995) to track individual users' knowledge for 15 years. Surprisingly, models currently in use do not allow for individual learning rates nor individualized estimates of student background knowledge. Corbett and Anderson, in their original articles, were interested in trying to add individualization to their model which they accomplished but with mixed results. Since their original work, the field has not made significant progress towards individualization of knowledge tracing models in fitting data. In this work, we introduce an elegant way of formulating the individualization problem entirely within a Bayesian networks framework that learns individualized as well as skill specific parameters in a single step. With this new model we are able to show a reliable improvement in knowledge estimation and data prediction. The novelty of the model is the ability to learn model parameters with the assumption that different students have different initial background knowledge probabilities. We evaluate the individualized model and standard knowledge tracing model with empirical results of predicting real student data. The implication of this work is the ability to enhance existing intelligent tutoring systems to more accurately estimate when a student has reached mastery of a skill. Adaptation of instruction based on individualized knowledge and learning speed is discussed as well as the open research questions facing those that wish to exploit student and skill information in their user models.

Keywords: Knowledge Tracing, Bayesian Networks, Data Mining, Prediction, Intelligent Tutoring Systems

1 Intro

Our initial goal was simple; to show that with more data about students' prior knowledge, we should be able to achieve a better fitting model and more accurate prediction of student data. The problem to solve was that there existed no Bayesian network model to exploit per user prior knowledge information. Knowledge tracing is

[1] National Science Foundation funded GK-12 Fellow

the predominant method used to model student knowledge and learning over time. This model, however, assumes that all students share the same incoming prior knowledge and does not allow for per student prior information to be incorporated. The model we have engineered is a modification to knowledge tracing that increases its generality by allowing for multiple prior knowledge parameters to be specified and lets the Bayesian network determine which prior parameter value a student belongs to if that information is not known before hand. The improvements we see in predicting real world data sets are palpable, with the new model predicting student responses better than standard knowledge tracing in 36 out of the 42 problem sets with the use of external data to inform a prior per student that applied to all problem sets. Equally encouraging was that the individualized model predicted better than knowledge tracing in 30 out of 42 problem sets without the use of any external data. Correlation between actual and predicted responses also improved significantly with the individualized model.

1.1 Knowledge Tracing

Knowledge tracing has become the dominant method of modeling student knowledge. It assumes that each skill has 4 parameters; a guess rate, learn rate, slip rate and a background prior on the probability that the skill was known before hand. This method was first introduced by Atkinson in 1972 [1]. Corbett and Anderson introduced this method to the intelligent tutoring field in 1995 [2]. It is currently employed by the cognitive tutor, used by hundreds of thousands of students, and many others.

It might strike the uninitiated as a surprise that the dominant method of modeling student knowledge in intelligent tutoring systems, knowledge tracing, does not allow for students to have different learn rates even though it seems likely that students differ in their rate of learning. Similarly, knowledge tracing assumes that all students have the same probability of knowing the skill at their first opportunity.

Corbett and Anderson were interested in implementing the learning rate and knowledge individualization that was originally described as part of Atkinson's knowledge tracing model. They accomplished this but with limited success. They created a two step process for their model where the four parameters of the model were learned in the first step and the individual weights for each student were applied in the second step with a form of regression. Various factors were also identified for influencing the individual priors and learn rates [3]. The results [2] of their model showed that while the individualized model's predictions correlated better with the actual test results, their standard non individualized knowledge tracing model predicted test performance with greater overall accuracy. More recent work, however, has found utility in the contextual individualization of the guess and slip parameters [4]. In this paper, we hope to reinvigorate the field to further explore models that explicitly model the assumption that students differ in their background knowledge, learning rate and possibly their propensity to guess or slip as originally suggested by Atkinson.

1.2 The ASSISTment System

Our dataset consisted of student responses from The ASSISTment System, a web based math tutoring system for 7th-12th grade students that provides preparation for the state standardized test by using released math problems from previous tests as questions on the system. Tutorial help is given if a student answers the question wrong or asks for help. The tutorial help assists the student learn the required knowledge by breaking the problem into sub questions called scaffolding or giving the student hints on how to solve the question.

2 The Model

Our model uses Bayesian networks to learn the parameters of the model and predict performance. Reye [3] showed that the formulas used by Corbett and Anderson in their knowledge tracing work could be derived from a Hidden Markov Model or Dynamic Bayesian Network (DBN).. Corbett and colleagues later released a toolkit [4] using non individualized knowledge tracing to allow researchers to fit their own student models with DBNs.

2.1 The Prior Per Student model vs. standard Knowledge Tracing

The model we present in this paper focuses only on individualizing the prior knowledge parameter. We call it the Prior Per Student (PPS) model. The only difference between our model and Knowledge Tracing (KT) is the ability to represent a different prior knowledge parameter for each student.

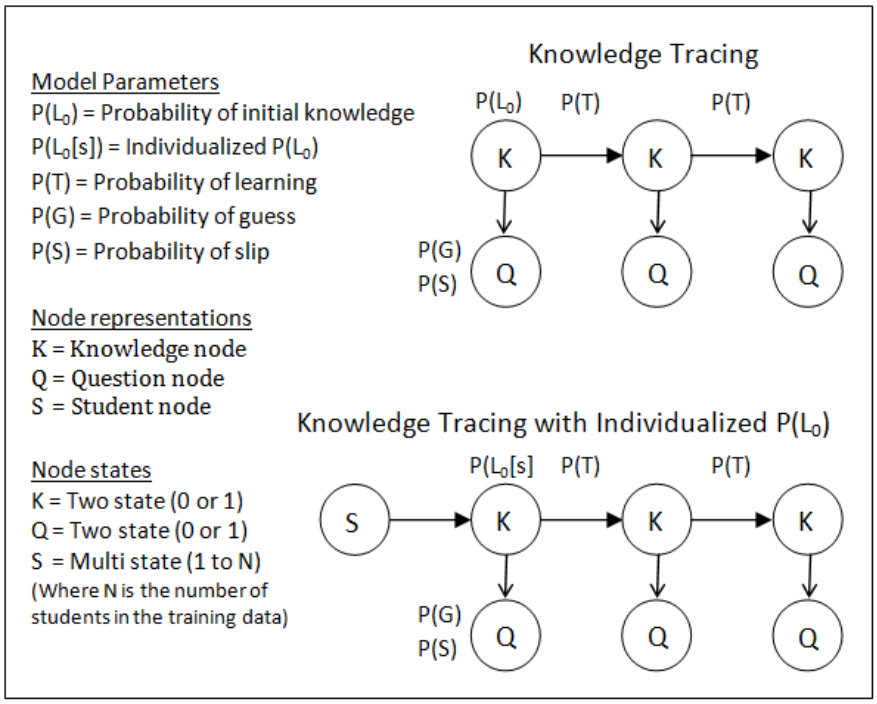

Figure 1. The topology and parameter description of Knowledge Tracing and PPS

The two model designs are shown in Figure 1. Initial knowledge and prior knowledge are synonymous. The individualization of the prior is achieved by adding a student node. The student node can take on values that range from one to the number of students being considered. The conditional probability table of the initial knowledge node is therefore conditioned upon the student node value. The student node itself also has a conditional probability table associated with it which determines the probability that a student will be of a particular value. The parameters for this node are fixed to be 1/N where N is the number of students. The parameter values for this node are not relevant since the student node is an observed node that corresponds to the student ID and need never be inferred.

This model can be easily changed to model individual learning rates by connecting the student node to the subsequent knowledge nodes thus training an individualized P(T) conditioned upon student as shown in Figure 2. Knowledge Tracing is a special case of this prior per student model and can be derived by setting all the priors to the same values or by specifying that there is only one student. This equivalence was confirmed empirically.

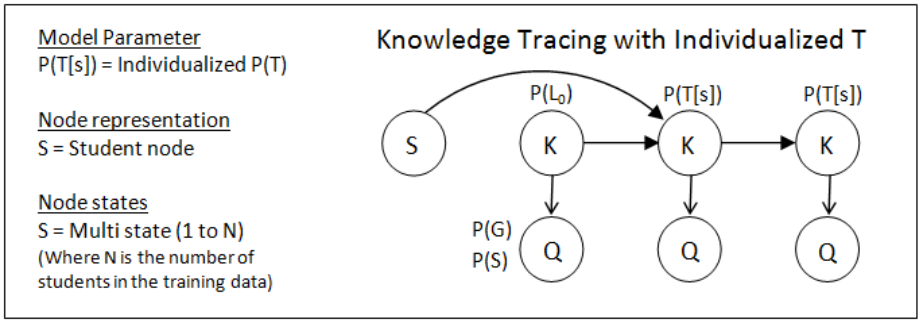

Figure 2. Graphical depiction of our individualization modeling technique applied to the probability of learning parameter. This model is not evaluated in this paper but is presented to demonstrate the simplicity in adapting our model to other parameters.

2.2 Parameter Learning and Inference

There are two distinct steps in knowledge tracing models. The first step is learning the parameters of the model from all student data. The second step is tracing an individual student's knowledge given their respective data. All knowledge tracing models allow for initial knowledge to be inferred per student in the second step. The PPS model allows for multiple priors to be learned along with the other parameters of the model in step one. We believe that if there is variance among student priors for a given skill, PPS will allow for more accurate guess and slip parameters to be learned. For example, if a student's knowledge is known to be zero and she answers the first question correctly, that performance can properly be attributed to the probability of guess. Similarly, if a student is known to have perfect knowledge and answers the first question incorrectly, that performance can properly be attributed to the probability of slip. However, if only a single prior is used to represent all students, the prior ends up being approximately the mean of all student priors, which in this two student example would be 0.50. With a prior of 0.50, the parameter learning procedure may struggle attributing either student's performance to a guess or slip. In our model each student has a student number identified by the student node. This number is presented during step one to associate a student with his or her prior. In step two, the individual student knowledge tracing, this number is again presented along with the student's respective data in order to associate that student with the individualized prior learned in the first step.

3 External Validity: Student Performance Prediction

In order to test the real world utility of the prior per student model, we used the last question of each of our problem sets as the test question. For each problem set we trained two separate models: the prior per student model and the basic knowledge tracing model. Both models then made predictions of each student's last question response and those responses tallied and compared to the students actual responses.

3.1 Dataset description

Our dataset consisted of student response to problem sets that satisfied the following constraints:

- Items in the problem set must have been given in a random order
- A student must have answered all items in the problem set in one day
- There are at least four items in the problem set of the exact same skill
- Data is from Fall 2008 to Spring 2010

Forty-two problem sets matched these constraints. Only the items within the problem set with the exact same skill tagging were used. The size of our resulting problem sets ranged from four items to thirteen. There were 4,354 unique students in total with each problem set having an average of 312 students (σ = 201) and each student completing an average of three problem sets (σ = 3.1).

Table 1. Sample of data format from a five item problem set

Student ID	1st response	2nd response	3rd response	4th response	5th response
750	0	1	1	1	1
710	0	1	1	1	0
714	1	1	0	1	0

In Table 1, each response represents either a correct or incorrect answer to the original question of the item. Scaffold responses are ignored.

3.2 Prediction procedure

Each problem set was evaluated individually by first constructing the appropriate sized Bayesian network for that problem set. In the case of the individualized model, the size of the constructed student node corresponded to the number of students with data for that problem set. All the data for that problem set, except for responses to the last question, was organized into an array to be used to train the parameters of the network using the Expectation Maximization (EM) algorithm. The initial values for the learn rate, guess and slip parameters were set to different values between 0.05 and 0.90 chosen at random. After EM had learned parameters for the network, student performance was predicted. The prediction was done one student at a time by entering as evidence to the network, the responses for the particular student except for the response to the last question. This enabled individual inferences of knowledge to be made about the student at each question including the last question. The probability of the student answering the last question correctly was computed and saved to later be compared to the actual response.

3.3 Approaches to setting the individualized prior knowledge values

In the prediction procedure, due to the number of parameters in the model, care had to be given to how the individualized priors would be set before the parameters of the

network were learned with EM. There were two decisions we focused on: 1) what initial values should the individualized priors be set to 2) whether or not those values should be fixed or adjustable during the EM parameter learning process. Since it was impossible to know the ground truth prior knowledge for each student for each problem set, we generated three heuristic strategies for setting these values, each of which will be evaluated in the results section.

3.3.1 Setting initial individualize knowledge to random values

One strategy was to treat the individualized priors exactly like the learn and guess and slip parameters by setting them to random values to then be adjusted by EM during the parameter learning process. This strategy effectively learns a prior per student per skill. This is perhaps the most naïve strategy that assumes there is no means of estimating a prior from other sources of information and no better heuristic for setting prior values. To further clarify, if there are 600 students there will be 600 random values between 0 and 1 set for each student. EM will then have 600 parameters to learn in addition to the learn, guess and slip parameters. For the non individualized model, the singular prior was set to a random value and was allowed to be adjusted by EM.

3.3.2 Setting initial individualized knowledge based on 1st response heuristic

This strategy was based on the idea that a student's prior is largely a reflection of their performance on the first question with guess and slip probabilities taken into account. If a student answered the first question correctly, their prior was set to one minus an ad-hoc guess value. If they answered the first question incorrectly, their prior was set to an ad-hoc slip value. Ad-hoc guess and slip values are used because ground truth guess and slip values cannot be known and because these values must be used before parameters are learned. The accuracy of these values could largely impact the effectiveness of this strategy. An ad-hoc guess value of 0.15 and slip value of 0.10 were used for this heuristic. Note that these guess and slip values are not learned by EM and are separate from the performance parameters. The non-individualized prior was set to the mean of the first responses and was allowed to be adjusted while the individualized priors were fixed. This strategy will be referred to as the "cold start heuristic" due to its bootstrapping approach.

3.3.3 Setting initial individualized knowledge based on global percent correct

This last strategy was based on the assumption that there is a correlation between student performance on one problem set to the next. This is also the closest strategy to a model that assumes there is a single prior per student that is the same across all skills. For each student, a percent correct was computed, averaged over each problem set they completed. This was calculated using data from all of the problem sets they completed except the problem set being predicted. If a student had only completed the problem set being predicted then her prior was set to the average of the other student priors. The single KT prior was also set to the average of the individualized priors for

this strategy. The individualized priors were fixed while the non-individualized prior was adjustable.

3.4 Performance prediction results

The prediction performance of the models was calculated in terms of mean absolute error (MAE). The mean absolute error for a problem set was calculated by taking the mean of the absolute difference between the predicted probability of correct on the last question and the actual response for each student. This was calculated for each model's prediction of correct on the last question. The model with the lowest mean absolute error for a problem set was deemed to be the more accurate predictor of that problem set.

Table 2. Prediction accuracy and correlation of each model and initial prior strategy

	Most accurate predictor (of 42)		**Avg. Correlation**	
L_0 Strategy	**PPS**	**KT**	**PPS**	**KT**
Percent correct heuristic	36	6	0.3457	0.1933
Cold start heuristic	30	12	0.3014	0.1726
Random parameter values	26	16	0.2518	0.1726

Table 2 shows the number of problem sets that PPS predicted more accurately than KT and vice versa in terms of MAE for each prior strategy. This metric was used instead of average MAE to avoid taking an average of averages. With the percent correct heuristic, the PPS model was able to better predict student data in 36 of the 42 problem sets. The binomial with $p = 0.50$ tells us that the probability of 36 success or more in 42 trials is $<< 0.05$, indicating a result that was not the product of random chance. The cold start heuristic, which used the 1st response from the problem set and two ad-hoc parameter values, also performed well - better predicting 30 of the 42 problem sets. According to the binomial the chance of 30 or more successes out of 42 < 0.05 making this result also statistically significantly reliable.

The correlation between the predicted probability of last response and actual last response using the percent correct strategy was also evaluated for each problem set. The PPS model had a higher correlation coefficient than the KT model in 32 out of 39 problem sets. A correlation coefficient was not able to be calculated for the KT model in three of the problem sets due to a lack of variation in prediction across students. This occurred in one problem set for the PPS model. The average correlation efficient across all problem sets was 0.1933 for KT and 0.3457 for PPS using the percent correct heuristic. Surprisingly, the correlation of the random parameter strategy using PPS was better than KT since the PPS random parameter strategy represents a prior per student per skill which could be considered an over parameterization of the model. Moreover, no effort is made to inform a reasonable prior per student, rather the values are set randomly. This is evidence to us that the PPS model may outperform KT in prediction under a wide variety of conditions.

3.4.1 Response sequence analysis of results

We wanted to further inspect our models to see under what circumstances they correctly and incorrectly predicted the data. To do this we looked at response sequences and counted how many times for PPS and KT, their respective prediction of the last question was right or wrong (rounding their probability of correct). For example: student response sequence [0 1 1 1] means that the student answered incorrectly on the first question but then answered correctly on the following three. The PPS and KT models were given the first three responses in addition to the parameters of the model to predict the fourth. If PPS predicted 0.68 and KT predicted 0.72 probability of correct for the last question, they would both be counted as predicting that instance correctly. We did this for the 11 problem sets of length four. There were 4,448 total student response sequence instances among the 11 problem sets. Tables 3 and 4 show the top sequences in terms of number of instances where both models predicted the last question correctly (Table 3) and incorrectly (Table 4). Tables 5-6 show the top instances of sequences where one model predicted the last question correctly but the other did not.

Table 3. Predicted correctly by both

Instances	Response sequence
1167	1 1 1 1
340	0 1 1 1
253	1 0 1 1
253	1 1 0 1
251	1 1 1 0

Table 4. Predicted incorrectly by both

Instances	Response sequence
251	1 1 1 0
154	0 1 1 0
135	1 1 0 0
106	1 0 1 0
72	0 0 0 1

Table 5. Predicted correctly by PPS only

Instances	Response sequence
175	0 0 0 0
84	0 1 0 0
72	0 0 1 0
61	1 0 0 0
24	1 1 0 1

Table 6. Predicted correctly by KT only

Instances	Response sequence
75	0 0 0 1
54	1 0 0 1
51	0 0 1 1
47	0 1 0 1
16	1 0 0 0

Table 3 shows the most frequently correctly predicted sequences, that happen to also be the top occurring sequences overall. The top sequence, where students answer all questions correctly, accounts for more than 1/3 of the sequences and is never predicted incorrectly by either model. The top 2-5 sequences predicted correctly by both have only one incorrect response, interestingly the incorrect response appears at an incremented position for each sequence. Table 4 shows that the sequence where students answer all questions correct except the last question is never predicted correctly by either model. In order to correctly predict this question individual learning rates may need to be modeled. PPS is able to predict the sequence where no problems are answered correctly, shown in Table 5. In no instances does KT predict the last question of sequences [0 1 1 0] or [1 1 1 0] correctly. This sequence analysis may not generalize to other datasets but it provides a means to identify areas the

model can improve in and where it is most weak. Figure 3 shows a graphical representation of the distribution of sequences predicted by KT and PPS versus the actual distribution of sequences. This distribution combines the predicted sequences from all of the four item problem sets.

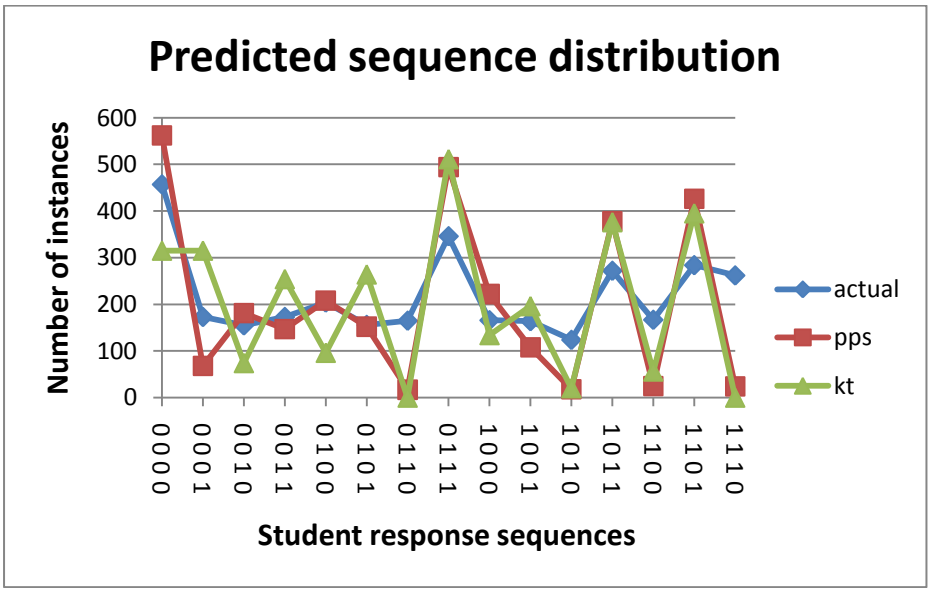

Figure 3. Actual and predicted sequence distributions of PPS and KT

The average residual of PPS is smaller than KT but as the chart shows, it is not by much. This suggests that while PPS has been shown to provide reliably better predictions, the increase in performance prediction accuracy may not be substantial.

4 Conclusion

In this work we have shown how individualization of the initial knowledge parameter in knowledge tracing can be accomplished with a simple technique and how that technique can easily be applied to model individualization of other parameters such as learn rate. The model we have presented allows for individualized and non individualized parameters of the model to be learned in a single step. We have also shown that the prior per student model, that individualizes initial knowledge, can be exploited to predict student data more accurately than standard knowledge tracing, which models a prior per skill.

5 Discussion

We hope this paper is the beginning of a resurgence in attempting to better individualize and thereby personalize a students' experience in intelligent tutoring systems.

We would like to know when using a prior per student is not a good idea. If all students had the same prior per skill then there would be no utility in modeling an individualized prior. On the other extreme, if student priors for a skill are highly varied then individualized priors are best since it allows the variation in that parameter to be captured.

Should we model the variations in prior among students as a discrete or a continuous probabilistic value? We have chosen in this model to represent the different priors as continuous probabilistic values but we point out that our cold-start strategy is an example of a bi-modal prior model since a student is assigned to a fixed prior based on their first response. While this heuristic worked, we suspect there are significantly superior representations. Ritter has recently shown that clustering of skills can drastically reduce the parameter space of Knowledge tracing while still maintaining high prediction accuracy [5]. Perhaps a similar approach can be employed to find clusters of student priors instead of learning the prior of each individual student.

Our work here has focused on just one of the four parameters in knowledge tracing. We are particularly excited to see if by explicitly modeling the fact that students have different rates of learning we can achieve higher levels of prediction accuracy. A student's learning could be indicative of motivational issues and if known, could prompt a type of tutor intervention personalized to the student. Guess and slip individualization is also possible, however, a potential pitfall with individualization is over parameterization and thus loosing the generality that allows for predictions to be made based on past and sometimes sparse observations.

We have shown that for a single skill there is a benefit in modeling students' variation in background knowledge and that utilizing prior information from other problem sets resulted in a marked improvement. We have shown that choosing a prior per student representation over the prior per skill representation of knowledge tracing is beneficial in fitting our dataset, however, an improved model is likely one that incorporates the student's prior with the skill's prior. How to design this model that properly treats the interaction of these two pieces of information is an open research question for the field.

In this work we focused on knowledge tracing, however, we know it is not the only model of learning. For instance, Draney, Wilson and Pirolli [7] have introduced a model they argue is more parsimonious than Knowledge tracing due to its fewer number of parameters. Additionally, EM is not the only method for fitting model parameters. Pavlik et al [8] have reported using different algorithms, as well as brute force, for fitting parameters with greater success than EM with their models. We also point out that more standard models used in educational measurement such as the Rasch model and item response theory have had large uses in and outside of the ITS field for estimating individual student and question parameters. We know there is value in these other approaches and strive as a field to learn how best to exploit information about students, questions and skills in our user models.

Acknowledgements

We would like to thank all of the people associated with creating the ASSISTment system listed at www.ASSISTment.org including investigators Kenneth Koedinger and Brian Junker at Carnegie Mellon. We would also like to acknowledge funding from the US Department of Education, the National Science Foundation, the Office of Naval Research and the Spencer Foundation. All of the opinions expressed in this paper are those solely of the authors and not those of our funding organizations.

References

1. Atkinson, R. C. and J. A. Paulson: 1972, An approach to the psychology of instruction. Psychological Bulletin, 78, 49-61.
2. Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4, 253–278.
3. Corbett A. and Bhatnagar A. (1997). Student Modeling in the ACT Programming Tutor: Adjusting a Procedural Learning Model with Declarative Knowledge. In Jameson A., Paris C., and Tasso C. (Eds.), User Modeling: Proceedings of the 6th International Conference, pp. 243-254.
4. Baker, R.S.J.d., Corbett, A.T., Aleven, V. 2008. More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. Proceedings of the 9th International Conference on Intelligent Tutoring Systems: 406-415.
5. Reye, J. (2004). Student modelling based on belief networks. International Journal of Artificial Intelligence in Education: Vol. 14, 63-96.
6. Chang, K.M., Beck, J.E., Mostow, J., & Corbett, A. (2006). A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. Proceedings of the 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan, 104-113
7. Ritter, S., Harris, T., Nixon, T., Dickison, D., Murray, C., Towle, B.(2009). Reducing the knowledge tracing space. In Barnes, Desmarais, Romero, & Ventura (Eds.). Proceedings of the 2nd International Conference on Educational Data Mining. pp. 151-160. Cordoba, Spain.
8. Draney, K. L., Pirolli, P., & Wilson, M. (1995). A measurement model for a complex cognitive skill. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), Cognitively diagnostic assessment (pp. 103–125). Hillsdale, NJ: Erlbaum.
9. Pavlik, P.I., Cen, H., Koedinger, K.R. (2009). Performance Factors Analysis - A New Alternative to Knowledge Tracing. In Proceedings of the 14th International Conference on Artificial Intelligence in Education. Brighton, UK, 531-538