# Using multiple Dirichlet distributions to improve parameter plausibility

Yue Gong, Joseph E. Beck and Neil T. Heffernan
{ygong, josephbeck, nth}@wpi.edu
Computer Science Department, Worcester Polytechnic Institute

Abstract. Predictive accuracy and parameter plausibility are two major desired aspects for a student modeling approach. Knowledge tracing, the most commonly used approach, suffers from local maxima and multiple global maxima. Prior work has shown that using Dirichlet priors improves model parameter plausibility. However, the assumption that all knowledge components are from a single Dirichlet distribution is questionable. To address this problem, this paper presents an approach to integrate multiple distributions and Dirichlet priors. We show that modeling groups of students separately based on their distributional similarities produces model parameters that provide a more plausible picture of student knowledge, even though the proposed solution did not improve the model's predictive accuracy. We also show Dirichlet priors might be hurt by outliers and models with trimming work better.

## 1 Introduction

In educational research, one fundamental goal is assessing students and estimating constructs, such as their knowledge levels, behaviors, goals and mental states, etc. Since most of those attributes are difficult to directly measure, the technique of student modeling has been widely used for estimating latent characteristics. A common evaluation of student modeling focuses on how well the model fits the training data and how well the model can generalize to unseen test data. However, there has been increasing research focusing on utilizing the model parameters to answer scientific questions [e.g., 1]. Since we are interpreting the model's parameters, we need some means of validating the model's *parameters*, not just its *predictions*. We call this property parameter plausibility. In this paper, we extended our prior work [2], investigating new approaches for improving the student model in terms of predictive accuracy and parameter plausibility. First, we provide some background into our student modeling framework, knowledge tracing, and its problems. We also illustrate the weaker points in our prior work and present a method that overcomes that limitation.

### *1.1 Knowledge tracing model*

Corbett and Anderson style knowledge tracing (KT) [3] has been successfully used in many tutoring systems to estimate a student's knowledge of a skill. It is based on a 2-state hidden Markov model where the student performance is observable, whereas his knowledge is latent. There are two parameters *slip* and *guess*, which mediate student knowledge and student performance. These two parameters are called the performance parameters in the model. The guess parameter represents the fact that the student may sometimes generate a correct response in spite of not knowing the correct skill. The slip parameter acknowledges that even students who understand a skill can make an occasional careless mistake.

In addition to the two performance parameters, there are two learning parameters. The first is *prior knowledge* ($K_0$), the likelihood the student knows the skill when he first uses the tutor. The second learning parameter is *learning*, the probability a student will acquire a skill as a result of an opportunity to practice it. Every skill to be tracked has these four parameters, *slip, guess, $K_0$,* and *learning*, associated with it.

### *1.2 The problem and proposed solution*

How to estimate the model parameters is an important issue. There are a variety of model fitting approaches. The Expectation Maximization (EM) algorithm is one of the most commonly used methods. It finds parameters that maximize the data likelihood (i.e. the probability of observing the student performance data). Compared to other model fitting approaches for KT, using EM to learn the parameters has been found to achieve the highest predictive accuracy [4]. However, it still suffers two major problems that are inherent in the KT model's search space: local maxima and multiple global maxima [2,5].

Local maxima are common in many error surfaces. The issue is that the algorithm has to start with some initial value of each parameter, and its final parameter estimates are sensitive to those initial values. The second difficulty, multiple global maxima, is known as identifiability and means that for the same model, given the same data, there are multiple (differing) sets of parameter values that fit the data equally well. Based on statistical methods, there is no way to differentiate which set of parameters is preferable to the others. Consequently, we have to be more careful to select the parameters' initial values when using EM to fit the model, as we want to neither be stuck with some local maxima, nor get unbelievable parameters which are meaningless for making scientific claims, even if those parameters make accurate predictions.

In order to solve the problems, in the previous work [2], we proposed that, rather than using a single fixed value to initialize the conditional probability table when training a knowledge tracing model, it is possible to use Dirichlet priors to start the algorithm. Briefly speaking, we assumed each parameter's values are drawn from Dirichlet distribution, which is specified by a pair of numbers ($\alpha$, $\beta$). The two numbers specify not only the most likely value for a parameter, but also the confidence in the estimate. The Dirichlet priors, which usually represent the researchers' prior beliefs, provide a reasonable starting point and bias the model-fitting process, thus decreasing the probability of ending with an implausible value.

Modeling all skills using the same set of Dirichlet priors assumes that all knowledge components are drawn from a single Dirichlet distribution. That is to say, knowledge components are assumed to have distributional similarities with each other in terms of all four attributes, prior knowledge, guess, slip and learning. Therefore, Dirichlet priors provide bias to all skills towards the mean of the distribution, especially to those abnormal outlier skills. In general, outliers could arise due to lack of sufficient observations. Specifically, with sparse data, the model is trained with few constraints from the evidence; thus although it achieves the highest predictive accuracy it could get, still generates implausible parameter estimates. In this situation, we argue that it is

preferable to have parameters which are more similar to the other, better-estimated, skills. As shown in Figure 1, Skill A and Skill B are at the tail of the distribution. By using Dirichlets, those outliers are biased towards the mean of the distribution. The hypothesis is that it is probably good that they are moved towards the center.

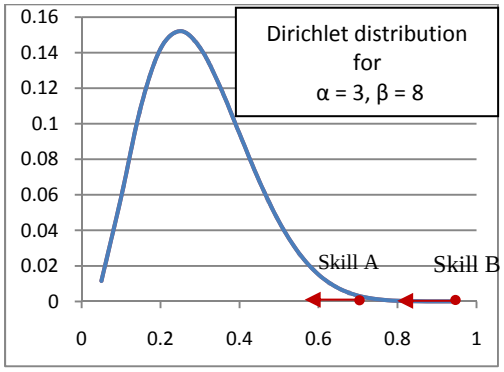


**Figure 1 Dirichlet distribution with two outliers**

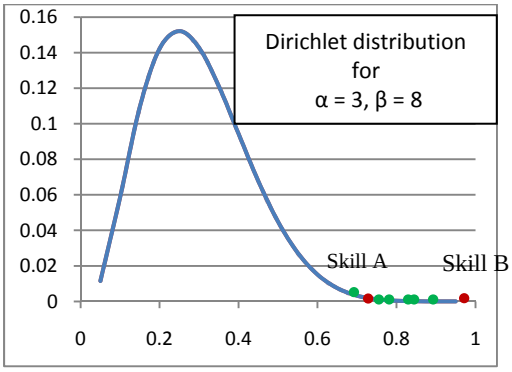


**Figure 2 Dirichlet with more "outliers"**

### *1.3 The problem with a single Dirichlet distribution*

Dirichlets has been shown to work well on positively biasing outliers [2,5]. However, a key question was overlooked: are the outliers really outliers?

Since the assumption of using Dirichlets is that skills in the domain are from a single distribution, those skills which are located further away from the mean are considered outliers. However, is it really true that all skills are from the same distribution? As shown in Figure 2, which has the same distribution as the one in Figure 1, if there are additional skills, with similar parameter estimates to Skills A and B, perhaps they are not really outliers. A plausible hypothesis is that they are from a cluster of skills which behave differently, i.e. they were not drawn from the same Dirichlet distribution as the other skills. If so, then moving them towards the mean may be inappropriate as they are better modeled as a separate distribution.

## 2 Methodology

### *2.1 Clustering*

We used clustering to discover which skills should be modeled separately with their own distribution. In the current context, a skill cluster is considered a region in the knowledge tracing parameter space where the skills share similar patterns with respect to the four knowledge tracing parameters. For example, possibly a group of skills might be described as "not previously known (low $K_0$), but easy to learn (high learning)", or "hard to learn, but students have partial incoming knowledge". The intuition is that the skills within a group are spatially located close to each other in the parameter space.

We used K-means clustering to identify the skill clusters. We did not use any self-adaptive clustering variants to automatically determine the number of clusters. The reason for this is that it is hard to evaluate the appropriate number of clusters, as our goal

is to find the clusters that will result in good predictive accuracy and parameter plausibility when modeled as Dirichlets. We had no *a priori* reason to believe that an automated clustering approach would optimize our metrics. Therefore, we used iteration until the number of clusters that works best on an unseen test set was found (i.e. we observed overfitting beginning to occur).

### *2.2 Trimming the data to improve Dirichlet parameter estimation*

There are several approaches to setting Dirichlet prior values. One approach is using knowledge of the domain [e.g. 5]. If someone knows how quickly students tend to master a skill or the likelihood of knowing a skill, that knowledge can be used to set the priors. One problem with this approach is that it is not necessarily replicable, as different domains, subjects, and experts may give different answers. Therefore, following the methods in [2], we automatically derived the Dirichlet priors from the data.

It is important to note, however, that automatically calculated Dirichlets are susceptible to undue influence from outliers. Similar to calculating the arithmetic mean, outliers can distort the parameter estimates. In order to address this problem, we trimmed outliers from the data using two different approaches.

The first approach was value-oriented, in which we processed the four knowledge tracing parameters separately and trimmed out the largest and smallest 5% of the values. For example, the 0.001 learning rate of the Pythagorean Theorem skill was in the lowest 5% and thus removed, while the more believable 0.45 prior knowledge estimate was not. The second approach was skill-oriented, in which we calculated the relative distance between a skill's parameters and its cluster's centroid. Those skills furthest away from the centroid were considered outliers, 10% of which we trimmed from the data used to calculate the Dirichlet priors.

### *2.3 Training with multiple Dirichlet distributions*

To compare the parameter plausibility and predictive accuracy of the fixed, single-Dirichlet prior and the multiple Dirichlet prior models, we trained a KT model on each of them using the following approach:

```
TrainWithMultiDrichlets (model, data)
1   [prior knowledge, guess, slip, learning] := EM (model, data, fixed prior[]);
2   for k :=1 to n
3         if (k !=1)  clusters[] := K-means ([prior knowledge, guess, slip, learning], k);
4         else cluster[1] := [prior knowledge, guess, slip, learning];
5         for i := 1 to k
6               for each dimension d from [prior knowledge, guess, slip, learning] k,i
7                     Dirichlet priors[] (α, β) := CalculateDirichlets(d);
8        [prior knowledge, guess, slip, learning]' k,i := EM (model, data in cluster k,i, Dirichlet priors[]);
```

We trained a knowledge tracing model for each skill using the same set of fixed priors for EM initialization. After finding each skill's parameter set (prior knowledge, guess, slip,

learning), we calculated the priors for a single Dirichlet distribution and reestimated the KT model. For multiple Dirichlet distributions, we classified the parameter sets into *k* clusters. For each cluster, we calculated its own Dirichlet priors, and then used those to initialize the EM algorithm and reestimated the models. We didn't specify an upper bound on the number of clusters (i.e. the value of *n*), as the number of clusters should depend on the improvement of predictive accuracy and parameter plausibility rather than the statistical properties of the clusters.

### *2.4 Data*

For this study, we used data from ASSISTment, a web-based math tutoring system. The data are from 345 twelve- through fourteen- year old 8$^{th}$ grade students in urban school districts of the Northeast United States. They were from four classes, each of which only lasted one month. These data consisted of 92,180 log records of ASSISTment during Dec. 2008 to Apr. 2009. Performance records of each student were logged across time slices for 105 skills (e.g. area of polygons, Venn diagram, division, etc). We took 20% of the students as the unseen test subjects. Their performance records are our test data.

## 3 Results

We used BNT-SM [7] to apply the EM algorithm to estimate the KT model's parameters. Following the above procedure in Section 2.3, we trained several models to fit the test dataset. We compared the models focusing on the model's predictive accuracy and parameter plausibility.

### *3.1 Predictive Accuracy*

We measured the models' predictive accuracy on the unseen test data set using two metrics: AUC (Area Under ROC Curve) and $R^2$.

In Table 1, for the three models with Dirichlets, the Dirichlet priors are calculated based on the trimmed data (since that gave slightly better results). We see the AUC values don't show any difference among the first four models. The values remain unchanged even we considered the possibility that skills come from multiple distributions (shown in the third and fourth rows). $R^2$ also didn't show any meaningful differences.

Since Ritter et al. have found that predictive accuracy when using the cluster centers is not much worse than when each individual skill's parameters are used, we decided to see if that result will replicate on our data set. Therefore, we evaluated the models using the cluster centers to predict the test data. The results in the last two rows of Table 1 showed that AUC values are similar to their counterparts, but $R^2$ values are lower (0.053 vs. 0.071 and 0.056 vs. 0.072), suggesting that compared to the predictions done by the models with parameters estimated for each skill, using generic cluster information to fit the data achieves less accurate, but possibly still acceptable, predictions.

These results show that predictive accuracy is not improved by using Dirichlets even with multiple distributions. Related to the prior work [4] where we evaluated the predictive

accuracy of the knowledge tracing model using a variety of model fitting approaches, and also the Performance Factor Analysis model [8], it seems that improving the model's predictive accuracy on the unseen students is a very difficult task.

**Table 1. Comparison of AUC and $R^2$**

| | AUC | $R^2$ |
|---|---|---|
| Fixed | 0.66 | 0.072 |
| Single Dirichlet | 0.66 | 0.071 |
| 2 Dirichlet distributions | 0.66 | 0.071 |
| 3 Dirichlet distributions | 0.66 | 0.072 |
| 2 distributions (cluster center) | 0.64 | 0.053 |
| 3 distributions (cluster center) | 0.65 | 0.056 |

### *3.2 Parameter plausibility*

In addition to using models for prediction, educational researchers also expect model parameters to be able to provide meaningful interpretations. Therefore, parameter plausibility is another important aspect for evaluating models. However, quantifying parameter plausibility or goodness is non-trivial due to the lack of gold standards. In our study, we used the two metrics we explored in [2].

For the first metric, we inspected the number of practice opportunities required to master each skill in the domain. We assume that skills in the curriculum are designed to neither be so easy to be mastered in three or fewer opportunities nor too hard as to take more than 50 opportunities. We define mastery as the same way as was done for the mastery learning criterion in the LISP tutor [9]: students have mastered a skill if their estimated knowledge is greater than 0.95. Based on students' prior knowledge and learning parameters, we calculated the number of practice opportunities required until the predicted knowledge exceeds 0.95. Then, we compared the number of skills with unreliable values in both cases (fewer than 3 and more than 50).

As seen in Table 2, the results might not be consistent in the two conditions. Fixed priors results in more skills with too fast mastery rate, whereas the other three models produce 5-6 more skills mastered too quickly. It is worth pointing out that the skills found to be slowly mastered by the fixed model is a subset of those found by the other three models. Furthermore, the skills with low mastery rates found by the three Dirichlet models have high overlap.

One possibility is the skills really are learnt that slowly. For example, if the students lacked preparation, they are unlikely to learn just through an ITS. All of the skills that required more than 50 opportunities to master were from the same distribution in the 2-distribution model. That distribution with "unlearnable" skills has the parameter estimates of 0.5, 0.36, 0.22 and 0.08 for prior knowledge, guess, slip and learning, respectively. Compared to the learning rate of the other distribution, 0.36, the skills are captured as ones that students have difficulties to learn, thus the mastery rates are very slow. Interestingly, in the 3-distribution model, the "unlearneable" skills are from two

distributions. One has higher prior knowledge, 0.62, but lower learning rates, 0.07. The other has lower prior knowledge, 0.39, but normal learning rate, 0.11. We know that both cases could result in a slow mastery progress. Therefore, although the numbers seems to suggest those skills are poorly-estimated, if there really are skills students don't learn, the models are better at finding them due to clustering.

**Table 2. Comparison of extreme number of practice until mastery**

| | # of skills with # of practices >=50 | # of skills with # of practices <=3 |
|---|---|---|
| Fixed | 22 | 2 |
| Single Dirichlet | 28 | 0 |
| 2 Dirichlet distributions | 27 | 0 |
| 3 Dirichlet distributions | 27 | 0 |

We also tried to evaluate parameter values directly by calculating the correlation between a skill's estimated prior knowledge and the grade at which that skill was taught. We assumed that the earlier the students learned the skills, the higher their incoming knowledge would be. However, we found our data suffer a severe problem that most items require multiple skills to answer, especially skills learned in earlier grades. Consequently, it confounds the relationship between the estimated prior knowledge and the grade at which the corresponding skill was taught, thus this approach was not viable. Therefore, we still followed the technique in [2]: using external measurement to evaluate parameter plausibility. The students in our study had taken a 33-item algebra pre-test before using ASSISTment. Taking the pre-test as external measure of incoming knowledge, we calculated the correlation between the *students'* prior knowledge estimated by the models and their pretest scores. In other to acquire the student's $K_0$ parameter, we used KT to model the students instead of skills (see [2] for details).

Table 3 shows four interesting results. The first and the most important one is that more Dirichlet distributions generally result in higher plausibility (shown in the second row). The correlation values of 0.88 and above are significantly higher than the baseline value 0.83 from the fixed prior model with p-values < 0.05. In the 7-distribtuion model, the value drops to 0.83. It suggests classifying students in a fine-grained level provides the models more confidence about the distributions where the data are from, thus taking the extra information specified by the Dirichlet priors, the models converge at more believable points. The second result is that we found the evidence of Dirichlet is hurt by outliers. As seen in the first column, the Dirichlet model produces lower correlation (0.80 vs. 0.83) compared to the fixed prior model. However, the Dirichlet model with trimming equals the fixed prior model, indicating the necessity of trimming for Dirichlets. However, the advantage from trimming decreases as the number of cluster increases, until eventually the untrimmed Dirichlet has better performance. Thus, the power from trimming is reduced as presumably the higher similarity of the students in a distribution reduced the problem of outliers.

The third result is the hypothesis that there is an interaction effect between using more distributions and using Dirichlets. To confirm that higher plausibility is not simply an

result of having additional distributions, we set each distribution's mean values as the fixed priors to train the models (first row of Table 3). We see that fixed prior models performance is independent of the number of distributions (except for possible overfitting with 6 clusters). Thus, the improvement from multiple Dirichlet distributions is not an artifact of multiple distributions necessarily resulting in better performance. The fourth result is shown at the last row of Table 3 where we used cluster centers to represent the individual student's prior knowledge. This approach achieves surprisingly high plausibility. With more distributions, it even outperforms the fixed prior models in spite of requiring less computation.

**Table 3 Comparison of correlation between prior knowledge and pretest, by number of clusters**

| | 1 cluster | 2 clusters | 3 clusters | 4 clusters | 5 clusters | 6 clusters |
|---|---|---|---|---|---|---|
| Fixed | 0.83 | 0.83 | 0.83 | 0.83 | 0.83 | 0.80 |
| Dirichlet | 0.80 | 0.83 | 0.85 | 0.86 | 0.88 | 0.91 |
| Dirichlet (trimmed) | 0.83 | 0.85 | 0.85 | 0.87 | 0.89 | 0.85 |
| Cluster center | N/A | 0.77 | 0.81 | 0.84 | 0.87 | 0.84 |

# 4 Contributions

This paper presents a new approach for strengthening the fundamental assumption of the usage of Dirichlet priors in order to improve the knowledge tracing model's predictive accuracy and parameter plausibility. Although Dirichlets are a solution to the problem of parameter plausibility, the assumption that all skills are from a single distribution is troubling. Rather than modeling skills as a single homogenous group, we acknowledge that similar skills should be modeled similarly. We used clustering techniques to identify groups of similar skills, and then modeled those groups with their own, independent Dirichlet priors.

In spite of using multiple Dirichlet distributions, we failed to find any improvement in predictive accuracy, which is consistent with the results in our previous work of investigating a single Dirichlet distribution. However, we confirmed that using distribution centers to fit the data isn't much worse than using the skill's individual parameter estimates [6].

For parameter plausibility of modeling skills, it appears using Dirichlets does not produce a more believable mastery rate, even when using multiple distributions. It is worth pointing out that if there really are skills that students don't learn, the Dirichlet approach is better at finding them. We also showed that using multiple Dirichlet distributions to model students results in high plausibility of the students' knowledge parameters. With multiple Dirichlet distributions (6 clusters), the correlation between the model's parameter estimates and the external standards reaches 0.9. We also showed that using the cluster centers, rather than individual student estimates, generates plausible results too, but with less computational work.

We found that Dirichlets are likely to be hurt by outliers, both with respect to predictive accuracy and parameter plausibility. For predictive accuracy, the models with trimming

perform comparable or even better than not using trimming. For the student knowledge parameter plausibility, trimming resulted in stronger results, except when six clusters were used. To understand this reversal requires additional experimentation.

Finally, our intuition that modeling a distribution as a single Dirichlet could be hurt since the "outliers" are the skills which are drawn from a different distribution has been partially supported by the results.

## 5 Future work and Conclusions

There are several unsolved problems related to this work. First of all, predictive accuracy is strongly desired in most student modeling applications. We have tried various approaches to improve accuracy in the knowledge tracing framework. However, we have found that there are no quick wins [2, 4, 5, 8]. We think perhaps only relying on the KT model with the basic structure might not be sophisticated enough to account for the substantial variability in student problem-solving efforts. One line of research is to consider integrating other useful information with KT, as it makes sense to be aware of other variables that might affect student performance such as question difficulty and student engagement. By accounting for other sources of variance, it enables us to better estimate the student's knowledge and (hopefully) consequently have a higher predictive accuracy and estimate more plausible parameters.

Second, considering the existence of multiple distributions seems reasonable and using multiple Dirichlet distributions is found to be beneficial in improving parameter plausibility. Dirichlet priors work fine in parameter plausibility on the student models, but don't have apparent benefit for skill models. It is an important task to understand how to overcome this issue, or even determine if it is a problem at all. At present, we lack the strong domain-driven parameter plausibility metric that was used in the initial work with Dirichlets for reading [5]. Determining better metrics for the domain of mathematics, or even better domain-independent metrics is a high priority. Human-generated Dirichlets might be a solution, as the single attempt [5] did result in more plausible parameters. Again, if we had more powerful parameter evaluation metrics we could better determine whether using human knowledge is a promising direction. It is interesting to see the outcomes from using other techniques to identify the distributions, such as latent Dirichlet allocation (LDA [10]), which is a generative model that allows sets of observations to be explained by unobserved groups. In this context, skills can be considered from several unobserved groups and each of them can be represented by a Dirichlet distribution. Thus, LDA is a promising technique rather than using clustering.

There is a limitation in this work. We took the benefit of looking at the test dataset for determining the number of clusters where the models result in the best performance. However, a better way that would be conducted in the future work is to use a tuning dataset besides the training and test datasets. This approach would enable us to tweak our models based on the models' performances on the tuning data, and then validate our models on the test data.

This paper has explored the idea of integrating multiple Dirichlet distributions with the knowledge tracing model. In terms of predictive accuracy, we failed to find any improvement contributed by the proposed technique. This work provides some additional support that using the using cluster centers is a reasonable approach. We found that, with multiple Dirichlet distributions, student knowledge parameters achieved high plausibility, even when using cluster centers to represent student knowledge. We have also found Dirichlet priors could be hurt by outliers, and found that first trimming the data before Dirichlet parameter estimations usually gives better performance.

## Acknowledgements

This research was made possible by the US Dept. of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant #R305A070440, NSF CAREER award to Neil Heffernan, the Spencer Foundation, and a Weidenmeyer Fellowship from WPI.

## References

[1] Beck, J. E., Mostow, J. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. 9th International Conference on Intelligent Tutoring Systems, Montreal, 353-362.
[2] Rai, D, Gong, Y, Beck, J. E. : Using Dirichlet priors to improve model parameter plausibility. Proceedings of the 2nd International Conference on Educational Data Mining, Cordoba, Spain, pp141-148.
[3] Corbett, A. and J. Anderson, Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 1995. 4: p. 253-278.
[4] Gong, Y, Beck, J. E., Heffernan, N. T. : Comparing Knowledge Tracing and Performance Factor Analysis by Using Multiple Model Fitting. The 10th International Conference on Intelligent Tutoring Systems, Pittsburgh, 2010
[5] Beck, J. E., Chang, K.-m.: Identifiability: A Fundamental Problem of Student Modeling. Proceedings of the 11th International Conference on User Modeling, Greece
[6] Ritter, S., Harris, T. K., Nixon, T., Dickison, D., Murray, R. C., Towle, B.: Reducing the Knowledge Tracing Space.Proceedings of the 2nd International Conference on Educational Data Mining, 2009, Cordoba, Spain.
[7] Kai-min Chang, Joseph Beck, Jack Mostow and Albert Corbett *:* A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems : Intelligent Tutoring Systems: Intelligent Tutoring SystemsVolume 4053/2006
[8] Pavlik, P. I., Cen, H., Koedinger, K.: Performance Factors Analysis - A New Alternative to Knowledge. Proceedings the 14th International Conference on Artificial Intelligence in Education, Brighton, UK, pp. 531-538
[9] Corbett, A.T. Cognitive computer tutors: Solving the two-sigma problem. International Conference on User Modeling. 2001. p. 137-147.
[10] Blei, D. M., Ng, A. Y., Jordan, M. I.. Latent Dirichlet allocation. Journal of Machine Learning Research. Vol. 3. Pp. 993-1022. 2003.