

# DIRT: Deep Learning Enhanced Item Response Theory for Cognitive Diagnosis

Song Cheng<sup>1</sup>, Qi Liu<sup>1,\*</sup>, Enhong Chen<sup>1</sup>, Zai Huang<sup>1</sup>,  
Zhenya Huang<sup>1</sup>, Yuying Chen<sup>1,2</sup>, Haiping Ma<sup>3</sup>, Guoping Hu<sup>3</sup>

<sup>1</sup>Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China,  
{chsong, huangzai, huangzhy, cyy3322}@mail.ustc.edu.cn; {qiliuql, cheneh}@ustc.edu.cn;

<sup>2</sup>Ant Financial Services Group, yuying.cyy@antfin.com

<sup>3</sup>IFLYTEK Co.,Ltd., {hpma, gphu}@iflytek.com

## ABSTRACT

Cognitive diagnosis is the cornerstone of modern educational techniques. One of the most classic cognitive diagnosis methods is *Item Response Theory* (IRT), which provides interpretable parameters for analyzing student performance. However, traditional IRT only exploits student response results and has difficulties in fully utilizing the semantics of question texts, which significantly restricts its application. To this end, in this paper, we propose a simple yet surprisingly effective framework to enhance the semantic exploiting process, which we termed *Deep Item Response Theory* (DIRT). In DIRT, we first use a proficiency vector to represent student proficiency on knowledge concepts and represent question texts and knowledge concepts by dense embedding. Then, we use deep learning to enhance the process of diagnosing parameters of student and question by exploiting question texts and the relationship between question texts and knowledge concepts. Finally, with the diagnosed parameters, we adopt the item response function to predict student performance. Extensive experimental results on real-world data clearly demonstrate the effectiveness and the interpretability of DIRT framework.

## CCS CONCEPTS

• **Information systems** → *Data mining*; • **Social and professional topics** → *K-12 education*;

## KEYWORDS

Cognitive diagnosis; Item response theory; Deep learning

### ACM Reference Format:

Song Cheng, Qi Liu, Enhong Chen, Zai Huang, Zhenya Huang, Yuying Chen, Haiping Ma, Guoping Hu. 2019. DIRT: Deep Learning Enhanced Item Response Theory for Cognitive Diagnosis. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM'19)*, November 3–7, 2019, Beijing, China. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3357384.3358070>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CIKM '19, November 3–7, 2019, Beijing, China

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6976-3/19/11...\$15.00

<https://doi.org/10.1145/3357384.3358070>




	1. In $\triangle ABC$ , $AB = AC$ , $\angle BAC = 108^\circ$ . $AD$ , $AE$ and $BC$ intersect at point $D$ and $E$ . And $\angle BAC$ is divided into three equal parts, what is wrong? <b>Concepts:</b> 1. Similar triangle properties 2. Similar triangle judgement 3. Proportional line segment	✓
	2. Calculate $4\sin 60^\circ + \tan 45^\circ - 2\sqrt{3}$ <b>Concepts:</b> 1. Quadratic root operation 2. Special trigonometric function	✗
	3. What is the minimal positive period of function $y = 1 - \cos(2x)$ ? <b>Concepts:</b> 1. Period 2. Trigonometric	✗

Figure 1: A toy example of student question records

## 1 INTRODUCTION

A large number of educational systems (e.g., massive open online courses) provide a series of computer-aided applications for better tutoring, such as computer adaptive test [7] and knowledge tracing [9]. Among these applications, the cognitive diagnosis that discovering the latent traits of students is becoming increasingly important. To execute cognitive diagnosis more effectively, the classic framework of Item Response Theory (IRT) [10] has been proposed, which introduces interpretable parameters with item response function to analyse students' performance.

Though IRT has achieved great successes in cognitive diagnosis area, there is still an important issue limits its usefulness. Specifically, it only considers student responses, right (e.g., 1) or wrong (e.g., 0)—that is, it ignores the rich semantics in the other question materials. As shown in Figure 1, the question texts and the knowledge concepts on the underline with the same color are closely related, which is helpful for modelling questions [5]. It motivates us to integrate semantics to improve and enhance traditional IRT.

To this end, we propose a novel and general *deep item response theory* (DIRT) framework to enhance item response theory. Specifically, we first create a proficiency vector to represent the student proficiency on each knowledge concept and embed questions. Then, to diagnose the latent trait  $\theta$  of students, the discrimination  $a$  and the difficulty  $b$  of questions [10], we introduce the deep learning methods (e.g., DNN, LSTM) for parsing semantics from question texts and the relationship between question texts and knowledge concepts. Finally, with the parameters diagnosed by deep learning methods, it can predict whether the student can answer the question correctly by item response function. Extensive experimental results present that DIRT surpasses traditional IRT by a large margin.

## 2 PRELIMINARIES

### 2.1 Cognitive diagnosis Task

Suppose there are  $L$  students,  $M$  questions and total  $P$  knowledge concepts. The history records that  $L$  students do  $M$  questions are represented by  $R = \{R_{ij} | 1 \leq i \leq L, 1 \leq j \leq M\}$ , where  $R_{ij} = \langle S_i, Q_j, r_{ij} \rangle$  denotes the student  $S_i$  obtains score  $r_{ij}$  on question  $Q_j$ .  $Q_j = \langle QT_j, QK_j \rangle$  is composed of question texts  $QT_j$  and knowledge concepts  $QK_j$ . Given students' responses  $r_{ij}$ , question texts  $QT_j$  and knowledge concepts  $QK_j$ , our goal is to build a model  $\mathcal{M}$  to diagnose students' proficiency on each knowledge concept. Since there is no ground truth for diagnosis results, following previous works [14], we adopt performance prediction task to validate the effectiveness of cognitive diagnosis results.

### 2.2 Related Models

**2.2.1 Item Response Theory.** IRT is one of the most important psychological and educational theories which roots in psychological measurement [10]. With the student latent trait  $\theta$ , question discrimination  $a$  and difficulty  $b$  as parameters, IRT can predict the probability that the student answers a specific question correctly with item response function. The item response function is defined as follow:

$$P(\theta) = \frac{1}{1 + e^{-Da(\theta-b)}}, \quad (1)$$

where  $P(\theta)$  is the correct probability,  $D$  is a constant which often set as 1.7.

**2.2.2 Multidimensional Item Response Theory.** MIRT is extended from IRT to meet the demands of multidimensional data [13]. With student latent traits  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^T$ , knowledge concept discriminations  $\mathbf{a} = (a_1, \dots, a_m)^T$  and intercept term  $d$  of the question as parameters, MIRT can also predict the probability of the student answers a specific question correctly with multidimensional item response function. The multidimensional item response function is defined as follow:

$$P(\boldsymbol{\theta}) = \frac{e^{\mathbf{a}^T \boldsymbol{\theta} + d}}{1 + e^{\mathbf{a}^T \boldsymbol{\theta} + d}}, \quad (2)$$

where  $P(\boldsymbol{\theta})$  is the probability same as IRT.

## 3 DIRT FRAMEWORK

To enhance item response theory for cognitive diagnosis, DIRT contains three modules, i.e., input, deep diagnosis and prediction module. Input module initializes a proficiency vector in each knowledge concept for the student, and embeds question texts and knowledge concepts to vectors. Deep diagnosis module diagnoses latent trait, discrimination and difficulty with deep learning to enhance the model. Prediction module predicts the probability that the student answers the question correctly with item response function. In the section bellow, we give a specific implementation of DIRT which is shown in Figure 2.

### 3.1 A Specific Implementation of DIRT

**3.1.1 Input Module.** Given a student  $S$ , we initialize a proficiency vector  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_p)$  with randomly, it is not belong to the training process, where  $\alpha_l \in [0, 1]$  represents the degree a student masters the knowledge concept  $l$ .

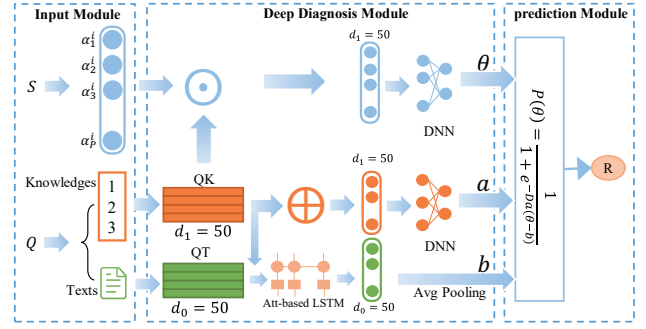


Figure 2: The Specific Implementation of DIRT.

For a question  $Q$ , question texts are composed of a sequence of words  $QT = \{w_1, \dots, w_u\}$ , where  $u$  is the length of  $QT$ ,  $w_i \in \mathbb{R}^{d_0}$  is a  $d_0$ -dimensional *Word2Vec* [8] vector, as for mathematical formulas, we regard each symbols as a word. Knowledge concepts are represented by one-hot vectors  $QK = \{K_1, \dots, K_v\}$ ,  $K_i \in \{0, 1\}^P$ , where  $v$  is the number of knowledge concepts. Then, we utilize a  $d_1$ -dimension dense layer to acquire the dense embedding for each knowledge concept  $K_i$  for better training, the dense embedding of  $K_i$  as  $k_i$ , and  $k_i \in \mathbb{R}^{d_1}$ :

$$k_i = K_i W_k, \quad (3)$$

where  $W_k \in \mathbb{R}^{P \times d_1}$  are the parameters of the dense layer.

**3.1.2 Deep Diagnosis Module.** Deep diagnosis module is mainly achieved by deep learning techniques (e.g., DNN, LSTM) to diagnose latent trait, discrimination and difficulty. The details are as follows.

**Latent Trait.** Latent trait  $\theta$  has strong interpretability for students' performance on questions, it is closely related to the proficiency of knowledge concepts [13]. In order to learn high-order features for latent trait diagnosing, we may use some nonlinear models (e.g., DNN), here we adopt deep neural network [15]. Specifically, given the proficiency vector  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)$  of the student  $s$  and a question  $q$ , we multiply the corresponding proficiency in  $\boldsymbol{\alpha}$  with the concepts dense embedding of the questions and get a  $d_1$ -dimension vector  $\Theta \in \mathbb{R}^{d_1}$ . Then we input  $\Theta$  into DNN to learn the latent trait, which is defined as follow:

$$\theta = \text{DNN}_\theta(\Theta), \quad \Theta = \boldsymbol{\alpha} \odot \mathbf{k} = \sum_{k_i \in \mathcal{K}_q} \alpha_i k_i, \quad (4)$$

where  $\mathcal{K}_q$  is the set of the knowledge concepts of question  $q$ .

**Discrimination.** Discrimination  $a$  can be applied to analyse student performance distribution on the question. Inspired by the relationship between Multidimensional Item Discrimination (MDISC) and knowledge concepts [13], we learn question discrimination  $a$  from knowledge concepts corresponded to the question. Also, since deep neural network can learn high-order nonlinear features automatically [15], we use another DNN to diagnose question discrimination  $a$ . Specifically, we sum the dense embedding of knowledge concepts in  $\mathcal{K}_q$  to get a  $d_1$ -dimensional vector  $A \in \mathbb{R}^{d_1}$ . Then, we input  $A$  into the DNN to diagnosis question discrimination. We normalize the discrimination to meet the requirements that the range of  $a$  should be  $[-4, 4]$  [1] and the definition of  $a$  is as follow:

$$a = 8 \times \text{sigmoid}(\text{DNN}_a(A) - 0.5), \quad A = \mathbf{k} \oplus \mathbf{k} = \sum_{k_i \in \mathcal{K}_q} k_i, \quad (5)$$

where the structure of  $DNN_a$  is same as  $DNN_\theta$ , but the parameters are not shared between them.

**Difficulty.** Difficulty  $b$  determines how difficult the question is. The first perspective is to diagnose difficulty by exploiting semantics of question texts [5]. Following previous works [11], LSTM can handle and represent long time sequence texts from semantic perspective which have strong robustness, we adopt LSTM to model difficulty  $b$  from question text perspective. As for the second perspective, the depth and width of knowledge concepts examined by the question also have a great impact on difficulty. The deeper and wider the knowledge concepts are examined, the more difficult the question is. Obviously, the depth and width of the examined concepts can be reflected by the relevance between question texts and knowledge concepts. We adopt an attention mechanism to capture the relationship between question texts and knowledge concepts. Totally, we design an attention-based LSTM to integrate question texts and knowledge concepts for diagnosing question difficulty  $b$ . Specifically, the sequence input to this LSTM is  $\mathbf{x} = (x_1, x_2, \dots, x_N)$ , where  $N$  is the max step of the attention-based LSTM. The  $t$ -th input step of attention-based LSTM is defined as follow:

$$x_t = \sum_{k_i \in \mathcal{K}_q} \text{softmax}\left(\frac{\xi_j}{\sqrt{d_0}}\right) k_i + w_t, \quad \xi_j = w_t^T k_i, \quad (6)$$

where  $\sqrt{d_0}$  is the scaling factor.  $\xi_j$  is the relevance between word  $w_t$  and the knowledge concepts in  $\mathcal{K}_q$ . After that, an average-pooling operation is utilized to obtain parameter  $b$ . Also, we normalize the difficulty to meet the requirements that the range of  $b$  should be  $[-4, 4]$  [1] and the definition of  $b$  is as follow:

$$b = 8 \times (\text{sigmoid}(\text{averagePooling}(h_N)) - 0.5), \quad (7)$$

where *averagePooling* is an operation that calculates the mean of all elements in the last step vector  $h_N$  of LSTM.

**3.1.3 Prediction Module.** The prediction module is used to preserve the ability of performance prediction and the interpretation power of student latent trait, question discrimination and difficulty in traditional item response theory. We input parameters diagnosed by deep diagnosis module into the item response function Eq.(1) [10] to predict the student performance on the specific question.

**3.1.4 DIRT Learning.** The whole parameters to be updated in DIRT mainly exist in two parts: input module and deep diagnosis module. In input module, the parameters need to be updated contain proficiency vector  $\alpha$ , question embedding weights and knowledge concept dense embedding weights  $\{\mathbf{W}_Q, \mathbf{W}_K\}$ . In the deep diagnosis module, the parameters need to be updated contain the weights of three neural networks  $\{\mathbf{W}_{DNN_a}, \mathbf{W}_{DNN_\theta}, \mathbf{W}_{LSTM}\}$  which are used to learn the latent trait, discrimination and difficulty respectively. The objective function of DIRT is the negative log likelihood function. Formally, for student  $i$  and question  $j$ , let  $r_{ij}$  be the actual score,  $\widetilde{r}_{ij}$  be the score predicted by DIRT. Thus the loss for student  $i$  on question  $j$  is defined as:

$$\mathcal{L} = r_{ij} \log \widetilde{r}_{ij} + (1 - r_{ij}) \log(1 - \widetilde{r}_{ij}), \quad (8)$$

in this way, we can learn DIRT by directly minimizing the objective function using Adam optimization [6].

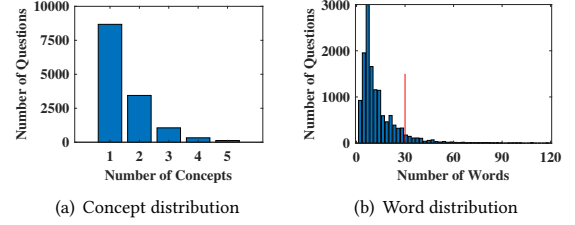


Figure 3: Distribution of words and knowledge concepts.

## 4 EXPERIMENTS

### 4.1 Dataset Description

Since DIRT needs to exploit question texts, only one private dataset can be used, which is composed of the mathematical data supplied by iFLYTEK Co., Ltd collected from Zhixue<sup>1</sup>. We filter out the students with less than 15 records and the questions that have not been answered by students. After pruning, the distribution of knowledge concepts number and question texts length are shown in Figure 3. Also, some statistics of the dataset are shown in Table 1. We can observe that each student has done about 62.09 questions, and each question requires about 1.49 knowledge concepts.

Table 1: The statistics of the dataset.

Statistics	Original	Pruned
# of history records	65,368,739	5,068,039
# of students	1,016,235	81,624
# of questions	1,735,635	13,635
# of knowledge concepts	1,412	621
Avg. questions per student	/	62.09
Avg. concepts per question	/	1.49

### 4.2 Baselines and Evaluation Metrics.

We compare the performance of DIRT with several methods: IRT [10] and DINA [3] are continuous and discrete cognitive diagnosis methods respectively, MIRT [13] is a multidimensional cognitive diagnosis method extend from IRT, (PMF) [4] and (NMF) [12] are matrix factorization methods, DIRTNA is a variant of DIRT without attention mechanism.

We evaluate the performance of DIRT from two perspectives, regression perspective [2]: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and classification perspective [14]: Area Under Curve (AUC) and Prediction Accuracy (ACC).

### 4.3 Experimental Results

**4.3.1 Performance Prediction Task.** Here, we conduct extensive experiments on performance prediction task at different data sparsity by splitting dataset into training and testing dataset with different ratio: 60%, 70%, 80%, 90%. The results on all metrics are shown in Figure 4. We can observe that compares with all the baselines, especially IRT, MIRT. DIRT performs the best, it illustrates that DIRT can make full use of question texts, benefiting the prediction. Comparing with DIRTNA, DIRT performs better, it proves that attention mechanism is effective for exploiting the relationship between question texts and knowledge concepts and helpful for prediction. We can also observe that DIRT and IRT perform better than MIRT, which is mainly because MIRT is sensitive to the concept on which student has high proficiency. Therefore, DIRT

<sup>1</sup><http://www.zhixue.com>

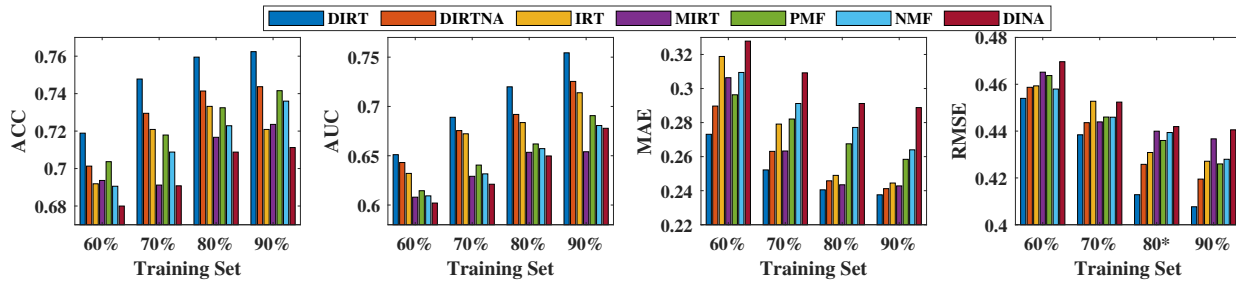


Figure 4: Overall results of student performance prediction on four metrics.

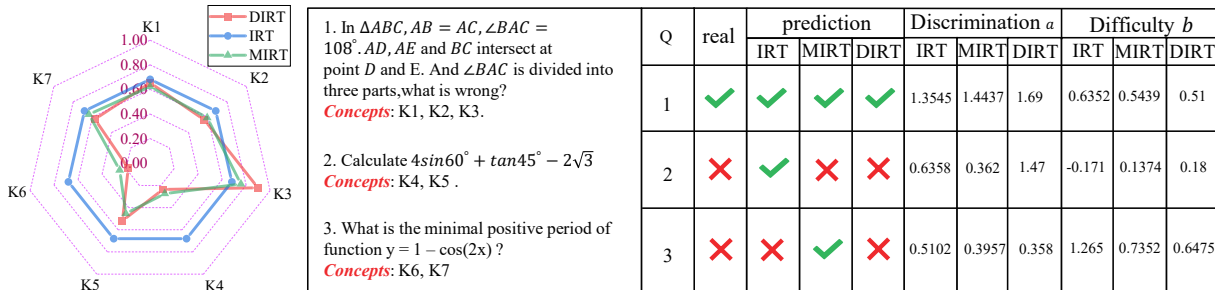


Figure 5: Visualization of a student's proficiency on knowledge concepts and the parameters of three questions.

framework is more reliable than MIRT to the concept on which students have a high proficiency.

4.3.2 Case Study. Here, we give an example of cognitive diagnosis of student knowledge proficiency. As shown in Figure 5, the radar chart shows a student's concepts proficiency diagnosed by IRT, MIRT and DIRT. Since IRT only diagnoses student latent trait which has the same value on all questions, so the diagnosis result of IRT is a regular polygon in Figure 5. Thus, DIRT can provide more accurate diagnosis results on knowledge concepts than IRT. We can also observe that DIRT predicts all three questions correctly, but IRT gets a wrong result on the second question, that because IRT obtains a wrong value -0.171 of difficulty  $b$  compares with DIRT and MIRT. Also, MIRT gets a wrong result on the third question, which is because MIRT is sensitive to concepts on which student has high proficiency [13] such as  $K7$ . Totally, DIRT can enhance traditional IRT with deep learning for cognitive diagnosis by exploiting question texts.

## 5 CONCLUSIONS

In this paper, we proposed a general DIRT framework to enhance traditional IRT to exploit the rich semantics in the question texts, as well as the relationship between question texts and knowledge concepts for cognitive diagnosis. Extensive experiments on a large scale real-world dataset clearly validated the effectiveness and the interpretation power of DIRT.

## ACKNOWLEDGMENT

This research was partially supported by grants from the National Key Research and Development Program of China (No. 2018YFC0832101), the National Natural Science Foundation of China (Grants No. 61672483, U1605251), and the Science Foundation of Ministry of Education of China & China Mobile (No. MCM20170507). Qi Liu gratefully acknowledges the support of the Young Elite Scientist

Sponsorship Program of CAST and the Youth Innovation Promotion Association of CAS (No. 2014299).

## REFERENCES

- [1] Frank B Baker. 2001. *The basics of item response theory*. ERIC.
- [2] Tianyou Chai and Roland R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE).
- [3] Huilin Chen and Jinsong Chen. 2016. Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly* 13, 3 (2016), 218–230.
- [4] Nicolás Fusi, Rishit Sheth, and Melih Elilbol. 2018. Probabilistic Matrix Factorization for Automated Machine Learning. In *NeurIPS*.
- [5] Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question Difficulty Prediction for READING Problems in Standard Tests. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- [6] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [7] Andrew J Martin and Goran Lazendic. 2018. Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology* 110, 1 (2018), 27.
- [8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
- [9] q. liu, Z. Huang, Y. Yin, E. Chen, H. Xiong, Y. Su, and G. Hu. 2019. EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction. *IEEE Transactions on Knowledge and Data Engineering* (2019), 1–1.
- [10] Georg Rasch. 1960. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. (1960).
- [11] Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 175–184.
- [12] Chenyang Yang, Mao Ye, Zijian Liu, Tao Li, and Jiao Bao. 2014. Algorithm for Non-Negative Matrix Factorization.
- [13] Lihua Yao and Richard D Schwarz. 2006. A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied psychological measurement* 30, 6 (2006), 469–492.
- [14] Run ze Wu, Guandong Xu, Enhong Chen, Qi Feng Liu, and Wan Ng. 2017. Knowledge or Gaming?: Cognitive Modelling Based on Multiple-Attempt Response. In *WWW*.
- [15] Liang Zhang, Keli Xiao, Hengshu Zhu, Chuanren Liu, Jingyuan Yang, and Bo Jin. 2018. CADEN: A Context-Aware Deep Embedding Network for Financial Opinions Mining. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 757–766.