EAnalyst: Toward Understanding Large-scale Educational Data

Tao Huang[1], Zhi Li[2], Hao Zhang[1], Huali Yang[1,*], Hekun Xie[2]
National Engineering Research Center for E-learning, Central China Normal University, Wuhan, China
[1]{tmht, zhanghao, yanghuali}@mail.ccnu.edu.cn
[2]{zhili, xiehekun}@mails.ccnu.edu.cn

ABSTRACT
We present an educational data collecting, mining and analyzing system, EAnalyst, for learners in the K12 period, providing highly intellectual personalized analysis and recommendations for learners. EAnalyst consists of preprocess module, analysis module, dashboard module and recommendation module. To assess target learner's knowledge proficiency better, we extend the current deep knowledge tracing model to achieves the goal of performance predicting. The results on both open dataset and our platform dataset demonstrate the effectiveness of our model run on our platform.

Keywords
E-Learning; Personalized Analysis; Data Mining

1. INTRODUCTION
The rapid development of information technology has helped the "learner-centered" teaching mode attracting more and more attention. With the assistance of big data analysis and artificial intelligence, promoting large-scale data-driven personalized learning analysis has become realistic. EAnalyst is a system whose main goal is to provide intelligent, personalized, and novel assistance to learners.

To meet the increasing needs of personalized learning [1], some existing work focuses on single work or test of a target learner [2] without continuous tracking and analysis of the whole learning process. Chronological data contain hidden patterns that are difficult to detect [3]. There are some attempts on analyzing educational time series data [4], evaluating learners' emotional changes throughout learning process [5], but they didn't consider to make analysis on learners' cognitive level. Some work tried to do cognitive analysis of learning [6], but they didn't combine it with temporal data mining and consider using deep learning techniques.

An intelligent teaching environment helps educators to communicate with learners and be informed of recent states of learners. These technologies make traditional teaching and learning more accurate and intelligent. The quality of education relies more on data analysis than on the experience of educators. Learners are involved in drawing up their learning plans at the same time. Georgia state university tracks students from arrival to graduation in three years and has made a total of 100,000 active interventions based on the risk alert provided by the system, which has increased the graduation rate of students from 48% to 54% [7]. In Oregon's Beaverton, students' drop-off records, absenteeism records and various demographic information are used to help students adapt to school life better [8].

EAnalyst[1] solves the problem that learners have a hard time figuring out their own knowledge proficiency because of deficient assessment methods and inadequate guidance. Combing domain knowledge with educational data mining and analysis, EAnalyst enables learners to know their knowledge state from the dashboard and provides remedial learning strategy. EAnalyst is an end-to-end system that has been tested on both elementary schools and secondary schools. Thus, the system is designed mainly for learners in the K12 period. The system has been used by part of students of those schools since 2014 and gets notable results in controlled experiments.

2. DATASETS
The data of learners are collected cautiously and critically. Different datasets lead to different outputs. Data of too large or too small granularity can be harmful to the analysis process.

The main component of data collected by EAnalyst ranges from pre-class quiz, post-class quiz, homework, unit-test and term-test. We refer every quiz, homework or test as a collection of series exercises. The former three are mainly about inspecting learners' short-term mastery level on concepts they just learned and the latter two on a larger concept coverage area. Exercises can be both online and offline. Educators use tools provided by the platform to select questions from question bank to form test papers. While offline exercises are commonly used for learners at a young age using the traditional paper test, online exercises are mainly taken on digital devices which can help collecting more information from question answering process such as time spent per question.

3. SYSTEM ARCHITECTURE
We describe EAnalyst architecture illustrated in Figure 1. EAnalyst is composed of preprocess module, analysis module, dashboard module and recommendation module. Preprocess module takes test papers and answer sheets as inputs and outputs structured data; analysis module takes structured data as input, outputs analysis results; dashboard module and recommendation module take analysis results as input then output visualized analysis results and recommendation list.

[1] study.hub.nercel.com/#/

* Corresponding Author

Tao Huang, Zhi Li, Hao Zhang, Huali Yang and Hekun Xie "EAnalyst: Toward Understanding Large-scale Educational Data" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 620 - 623

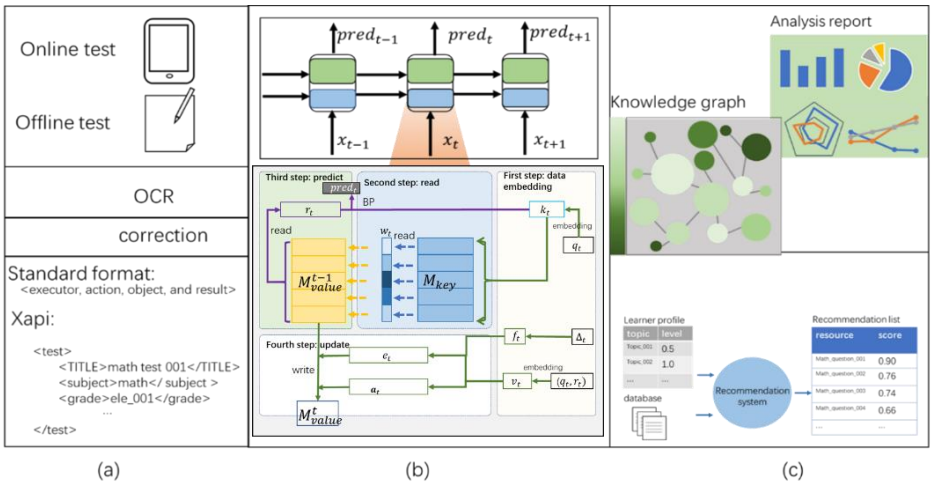

Figure 1. System architecture of EAnalyst: (a) Preprocess module. (b) Analysis module. (c) Dashboard module and recommendation module

3.1 Preprocess Module

Preprocess module uses optical character recognition (OCR) to transfer handwritten answers and correction marks to machine-encoded text. The module applies Transformer [9] which is one of natural language process (NLP) techniques to comprehensively learning question representation so that it can label questions with corresponding knowledge concepts. The response of learners to questions are recorded after being corrected by educators. The module then formalizes those heterogeneous educational data using the Experience API (Xapi), which makes the data readable for machine. Figure 1(a) illustrates EAnalyst's preprocess module.

3.2 Analysis Module

Learners interact with their coursework and generate sequences of learning process records. A sequence consists of multiple interaction record $x_0,\ldots,x_t$. The task of this module can be seen as predicting learner's future performance x_{t+1}. The record x_t at time step t can be represented as $x_t = (q_t, a_t)$ where q_t is a question learner attempts at time step t and $a_t \in \{0,1\}$ means learner's response (1 means correct and 0 means incorrect). Learning history is then analyzed by knowledge tracing model to reveal learners' learning status. From knowledge tracing prediction, educators can identify specific areas where learners need extra help. Educators can also analyze the data of the whole class to see their learning habit and adjust courses according to the feedback. Educators can even compare this information with that from other grades to determine which teaching methods are most effective.

The datasets that are used by knowledge tracing model are collected during the 2017-2019 school years. The datasets we conducted experiments on is on math subject, which has covered 652752 practice attempts of 3962 students on 4784 distinct questions. We filter learners who has fewer than three exercises to guarantee the reliability of knowledge tracing results since sequences that only contain one or two exercises barely contribute to tracing knowledge state of learners. We summarize some statistical features of two datasets in Table 1 and EAnalyst dataset distribution in Figure 2. For EAnalyst dataset, the average number of records per learner is 165. For EAnalyst dataset each learner interacts with more distinct questions than that in open dataset, which makes EAnalyst dataset more sparse.

Deep learning has made a huge success in tasks like image recognition, natural language processing (NLP), voice recognition and etc. Tasks which are good at handling sequential data use model like Long Short-Term Memory (LSTM) networks [10], a type of Recurrent Neural Networks (RNN), and get good results. Compared with models based on statistical graph like Bayesian Knowledge Tracing [11] and models based on matrix decomposition like Knowledge Proficiency Tracing [12], models based on deep learning, called Deep Knowledge Tracing (DKT) [13] are more flexible, which can be combined with effective mechanics so that they can make use of other information like content of questions and domain knowledge. DKT uses LSTM and its variation to cover previous learning records in a long time period to detect learners' knowledge state and memorize it in hidden vectors. This method has been combined with the attention mechanism to evaluate similarity among different question contents to improve prediction accuracy [14].

Table 1. Statistics of two datasets

Dataset Name	**EAnalyst Dataset**		**Assistment2009 Dataset**
Attribute of Dataset	*Original*	*Pruned*	*Original*
records	657573	652752	525534
learners	4285	3962	15931
questions	4788	4784	124

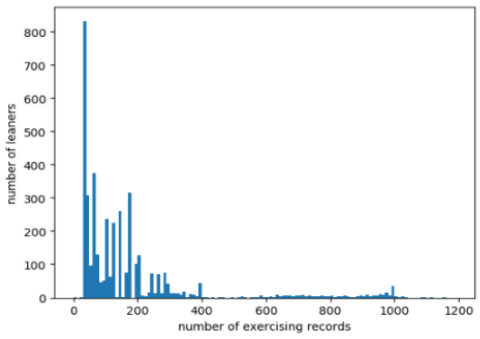

Figure 2. Distribution of EAnalyst Dataset on math subject.

Models like memory network [15] which has worked well in NLP field has also performed well at learning correlation between different questions. Model [16] using static key memory matrix to store question-concept relationships and dynamic value memory matrix to store and update concept-learner state relationships. This model performs well at knowledge tracing. We inherit the advantage of two memory matrices and apply convolution neural networks with some additional calculation to the reading process in the third step to reduce information loss in reading memory matrix process. We also consider the forgetting behavior of learners and add time interval of adjacent exercises to the updating step so that the model can simulate forgetting behavior. At first step, input data will be embedded. At second step, a question q_t is used to retrieve related concept position w_t in key matrix. At third step, position w_t is used in value matrix to query corresponding concept state. Finally, the concept state is used to predict learner's future performance on q_t. At fourth step, only related concept state will be updated in value matrix. The overall structure is illustrated in Figure 1(b).

We compare the prediction accuracy on both our dataset and public benchmark dataset—Assistment2009 [17]. Assistment is an online platform which teaches and assesses learners in elementary school mathematics. It is also the largest available public knowledge tracing dataset. We use Area Under a ROC Curve (AUC) to measure performance of the traditional model and deep learning model. AUC value ranges from 0.5 to 1 where the former value indicates the prediction result by random guessing and the latter represent precise prediction.

We set all sequences to be length of 150 and use -1 to pad short sequences to the expected length. The parameters are initialized randomly using Gaussian distribution. We set batch size for Assistment2009 dataset to 32 and that for Eanalyst dataset to 16 due to limitation of gpu memory. For momentum, it is set to be 0.9 and for norm clipping threshold to be 50.

The performance of different models is listed in in Table 2. The comparison results lead to findings that EAnalyst model can produce relative good result on Assistment2009 and better prediction results on EAnalyst dataset considering EAnalyst dataset are much sparser than Assistemt2009. And Our model does not come into the problem of overfitting due to its complexity compared to DKT's LSTM network.

Table 2. Performance of different models on two datasets – Eanalyst dataset and Assistment2009 dataset (AUC)

Model	EAnalyst Dataset	Assistment2009 Dataset
Bayesian Knowledge Tracing	0.69	0.73
Variant of Bayesian Knowledge Tracing	0.75	0.82
Deep Knowledge Tracing on EAnalyst platform	**0.85**	**0.86**

3.3 Dashboard Module

Dashboard module is a visualization tool for learners displaying results of analysis on knowledge graph, which is illustrated in Figure 1(c) upper part. Educators and experts in education field construct the knowledge graph manually according to textbooks and their experience. Knowledge graph constructs a network of knowledge concepts, which are connected by lines with relevant knowledge concepts. The size of each concept is related to its importance. The importance level is valued by corresponding syllabus. The more important a concept is, the bigger is a node. Color depth of a node indicate how a leaner mastery a concept node. Each subject includes multiple knowledge graphs divided by school year while some concepts can appear in one or more graphs. Knowledge graph is a precondition of accurate analysis of learners' overall cognitive levels, knowledge state and appropriate learning path recommendation. A learner and his or her educator can locate weak spots easily. And having a big picture of one's knowledge state helps the learner to carry out the following remedial activities.

Analysis report giving a more detailed description of a learner's learning report. History of exercises will be evaluated in a statistical point of view. Different types of charts such as histogram, pie chart, radar chart and line chart. These charts can well represent changes in learning indicator of learners over time, break out learners of a class by percentage of accuracy they have got, show distribution of a leaner's overall quality and give a rough comparison between the learner and the average level of his or her class and grade. Figure 3 gives a partial screenshot of a learner's dashboard in elementary school mathematics.

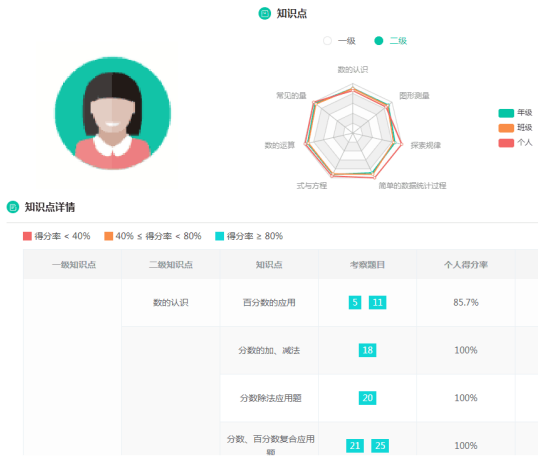

Figure 3. concepts mastery level in a radar chart and statistical report

Analysis report giving a more detailed description of a learner's learning report. History of exercises will be evaluated in a statistical point of view. Histogram represents change in learning indicator like accuracy over time. Pie chart breaks out learners of a class by percentage of accuracy they have got. Radar chart shows distribution of a leaner's overall quality. Line chart gives a rough comparison between the learner and the average level of his or her class and grade.

Dashboard contains statistical reports generated from analysis module and knowledge graph presenting learner's knowledge proficiency. The report displays learner's test results, test analysis. The circle in the graph represents separate entities. The importance of the entity is distinguished by size, and the depth of color indicates the learners' mastery level of each entity. The line between two circles displays relation existing between two corresponding entities. Dashboard works as an effective tool to promote learners to define and achieve goals.

3.4 Recommendation Module

Recommendation module mines learner features and course features, uses learners' rating of learning materials as supervised labels to filter recommendation materials like reading material, exercises, notes and outstanding answers from learning partners. We form a learner-course feature vector matrix by combining learners' behavior data with attributes data from learners and courses. This module first uses extraction capabilities of deep belief networks (DBN) to collect features from learner-course matrix to represent learners' preference. This feature extraction part is composed of bottom-up unsupervised pretraining using layers of restricted Boltzmann machine (RBM) and top-down supervised parameter fine-tuning using Backpropagation (BP) in the last level of the DBN. The trained DBNs from unsupervised part and corresponding rating score labels are used as inputs to the BP supervised part [18]. Then the recommendation model can be used to rating learning materials with scores. Materials with scores are ranked and those with higher scores are recommended to learners. The process is illustrated in Figure 1(c) lower part. This recommendation list will be updated dynamically according to newly generated learning tracks to match learners' changing needs.

4. CONCLUSION

We present EAnalyst, a learner's assistant developed by applying deep learning techniques for large-scale educational data mining and analysis. The system takes temporal data analysis aligned with knowledge graph, presents learners with multidimensional analytical reports, and recommending learning paths by offering relative learning materials. In the future, we intend to solve the "cold start" problem of learners' performance evaluation process and improve the analysis model by adding question content so that the deep relation between questions and learners' state can be exploited.

5. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under grants No. 61977033.

6. REFERENCES

[1] Wang Y, Liao H C. 2010. Data mining for adaptive learning in a TESL-based e-learning system. *Expert Systems with Applications*, 2011,38(6): 6480-6485. DOI=https://doi.org/10.1016/j.eswa.2010.11.098

[2] Brown N C C, Kölling M, McCall D, et al. 2014. Blackbox: a large scale repository of novice programmers' activity. *Proceedings of the 45th ACM technical symposium on Computer science education*. ACM, 2014: 223-228. DOI=https://doi.org/10.1145/2538862.2538924

[3] Allevato A, Thornton M, Edwards S, et al. 2008. Mining data from an automated grading and testing system by adding rich reporting capabilities. *Educational Data Mining 2008*. 2008.

[4] Baker R S, Inventado P S. 2014. Educational data mining and learning analytics. *Learning analytics*. Springer, New York, NY, 2014: 61-75. DOI=https://doi.org/10.1007/978-1-4614-3305-7_4

[5] Le N T, Boyer K E, Chaudry B, et al. 2013. The First Workshop on AI-supported Education for Computer Science (AIEDCS). *International Conference on Artificial Intelligence in Education*. Springer, Berlin, Heidelberg, 2013: 947-948. DOI=https://doi.org/10.1007/978-3-642-39112-5_159

[6] Schunk D H, Greene J A. 2017. Handbook of self-regulation of learning and performance. Routledge, 2017.

[7] Executive Office of the President, Munoz C, Director D P C, et al. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights. *Executive Office of the President*, 2016.

[8] West D M. 2014. Big data for education: Data mining, data analytics, and web dashboards. *Governance studies at Brookings*, 2012, 4(1).

[9] Vaswani A, Shazeer N, Parmar N, et al. 2017. Attention is all you need. *Advances in neural information processing systems*. 2017: 5998-6008.

[10] Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural computation*, 1997, 9(8): 1735-1780. DOI=https://doi.org/10.1162/neco.1997.9.8.1735

[11] Yudelson M V, Koedinger K R, Gordon G J. 2013. Individualized bayesian knowledge tracing models. *International conference on artificial intelligence in education*. Springer, Berlin, Heidelberg, 2013: 171-180. DOI=https://doi.org/10.1007/978-3-642-39112-5_18

[12] Chen Y, Liu Q, Huang Z, et al. 2017. Tracking knowledge proficiency of students with educational priors. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017: 989-998. DOI=https://doi.org/10.1145/3132847.3132929

[13] Piech C, Bassen J, Huang J, et al. 2015. Deep knowledge tracing. *Advances in neural information processing systems*. 2015: 505-513.

[14] Su Y, Liu Q, Liu Q, et al. 2018. Exercise-enhanced sequential modeling for student performance prediction. *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[15] J. Weston, S. Chopra, and A. Bordes. 2015. Memory networks. *International Conference on Learning Representations*, 2015.

[16] Zhang J, Shi X, King I, et al. 2017. Dynamic key-value memory networks for knowledge tracing. *Proceedings of the 26th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017: 765-774. DOI=https://doi.org/10.1145/3038912.3052580

[17] Feng M, Heffernan N, Koedinger K. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 2009, 19(3): 243-266. DOI=https://doi.org/10.1007/s11257-009-9063-7

[18] Zhang H, Huang T, Lv Z, et al. 2019. MOOCRC: A Highly Accurate Resource Recommendation Model for Use in MOOC Environments. *Mobile Networks and Applications*, 2019, 24(1): 34-46. DOI=https://doi.org/10.1007/s11036-018-1131-y