A Learning Analytics System for Cognition Analysis in Online Learning Community

Yinan Wu and Wenjun Wu

State Key Lab of Software Development Environment, Department of Computer Science and Engineering, Beihang University, Beijing, China
wuyinan@nlsde.buaa.edu.cn

Abstract. While cognitive behaviors and social network structure in Online Learning Community (OLC) have been studied in the past, few research has proposed a model linking the two important factors to analyze students’ cognitive learning gains, even though it has been widely acknowledged that interaction is a significant way for students to exchange knowledge and obtain learning gains. In this paper, for a better indication of cognitive gains, we introduce an analytic model to quantify the students’ learning gains by using a redesigned taxonomy of cognitive behaviors while considering the flow of knowledge among students in discussion forums. And further, we implement a learning analytics system to streamline the data analysis pipeline of social network analysis, cognition classification and learning gain calculation and visualize the analytic results from multiple-level views including student, discussion thread and forum. We demonstrate the results on a MOOC course and confirm the effectiveness of our model. Our model and analytic system enable instructors and TAs to take active mediation among online discussions of students to improve their cognitive gains through OLCs.

Keywords: MOOC discussion forums · Social learning network Cognitive learning gains · Taxonomy of cognition

1 Introduction

Massive Open Online Courses (MOOC) has attracted millions of registered users to learn over the Internet, which scaled distance education to a magical size that everyone can participate in courses developed by numerous universities and educational institutions [1]. However, many problems regarding MOOCs such as high dropout rate and low terminal efficiency remain unsolved. To address these problems, Online Learning Community (OLC) had been provided to increase the teacher-to-student ratios and face-to-face interaction [2]. In addition, the instructor and TAs are able to monitor the learning progress based on the posts [3]. Previous research efforts have qualitatively and quantitatively proved that students’ participation in online discussion is positively correlated to their learning gains [4, 5]. For exploring the factors in forums that would influence students’ learning gains, many methods are proposed from the two main aspects: content analysis and social structural analysis.
Currently, most researchers adopt content-related analysis, which focuses on students’ posts observed in discussion forums, such as topic analysis, semantic analysis and emotional analysis. OLCs offer a new environment that provides computer supported collaborative learning activities, where social existence might reflect cognitive existence [6]. Besides, interpersonal relationships among students can provide cognitive and emotional support that ultimately benefit the learning process [7]. Online learning community as a social learning network (SLN), provides a place where teachers and students exchange information and share their ideas and insights into course topics through discussion threads. Thus, structural connectivity between each pair of students, is also an important factor of affecting students’ learning gains [8]. However, the social structure of OLCs is not well applied on analyzing students’ cognitive learning gains in previous investigations.

In this work, we aim to combine SLN structural analysis of OLCs and cognitive analysis to build a new analytic model to quantitatively assess the students’ learning gains. Further on, considering the continued rising in popularity of OLCs [9], we parallelized the algorithm in our cognitive gains model and developed a Spark-based analytics system to streamline the large-scale data analysis from forum data collecting, data analysis to result visual presentation. In particular, we contribute to the existing literature by (1) redesigning an easy labeled coding scheme of cognitions based on Wang’s framework [10] for precisely categorizing students’ cognitive behaviors in OLCs; (2) developing a classification workflow including posts preprocessing, labeling, features extracting and modeling; (3) visualizing the learning gains from the perspectives of student, thread and course forum respectively.

The remainder of the paper is organized as follows: Related work section describes the related work and previous theoretical basis that we take advantage of in our research. Dataset section describes the datasets for the research of the paper, and in Methods section, we elaborate our redesigned coding scheme, post processing methods, cognitive gains model and the selection of latent parameters. Results section describes the implementation of our analytics system and experimental results of our analytic model. Conclusion section gives the conclusion and points out the future work.

2 Related Work

2.1 Social Network Analysis on OLCs

Online learning community provides students an online environment to interact with their learning partners. One of the issues about social learning pointed out by Putnik [11] is to analyze the students’ interactions using techniques from the social network analysis field. The interactions between students in OLCs mainly includes asking questions, offering answers and giving opinions, and all of them promote the flow of knowledge exchange between students and bring them benefits. Brinton et al. [12] proposed a framework for modeling SLN efficiency of information exchange between students and evaluated students’ gains in topic level. The study defines student’s benefit from his/her learning benefits by asking questions and his/her teaching benefits by
providing answers. The study has set up a good foundation for analyzing students’ gains in social structure of OLCs.

However, it is not enough to express the gains from questions or answers in different levels of cognition. For example, novel questions would benefit students more than simple questions and reasoned answers would benefit students more than repeated answers. Such an observation motivates more fine-grained analysis method to be defined for classifying the gain sources in multiple cognitive levels.

2.2 Cognitive Analysis on OLCs

Investigating the online behaviors in cognition level in OLCs has been one major theme for research. Elouazizi [13] used linguistic cues to measure cognitive engagement in forum data and found evidence of low cognitive engagement. Wong [14] used content analysis to investigate the tendency of different levels of cognitive learning in OLCs. Wang [10] classified different behaviors into three cognition categories including active, constructive and interactive behaviors based on Chi’s ICAP framework [15] and explored the relationships between each cognitive behavior and students’ scores. Their finding imply that students’ interactive discussion behaviors will benefit more than constructive behaviors, and constructive behaviors will benefit more than active behaviors.

These research results give our work strong theoretical support, especially Wong’s study that provides a quantification basis for each cognition.

However, only the quantity of student’s behaviors in each cognition is considered to affect the scores in Wong’s study, but a student wouldn’t obtain gains without information exchanges (i.e. no one replies his/her post) in the forum. Hence in our work, we take others’ contribution into account to calculate students’ cognitive gains.

3 Dataset

Data for this research comes from two different data sources. For testing the classification result of cognitive behaviors, we used the Stanford MOOC Posts dataset [16], which contains 29,604 posts from OLCs within the Education, Humanities and Medicine domain areas. It is convenient for testing because the opinion, question and answer classifications are labeled in the dataset.

But the Stanford dataset doesn’t cover students’ information required in our cognitive gains analysis. We have to collect the second dataset of an OLC from an online educational institution (referred as SETC), made available upon successfully fulfilling application requirements. The SETC dataset contains 15,536 posts and 128,202 answers in 1540 questions, covering 2,858 students in an OLC within the 420 high school courses.

To analyze students’ cognitive gains over the social network, we select the course “Mathematic Thinking Method” from the second dataset, which includes the 1592 posts made by 403 students. “Mathematic Thinking Method” was designed as a 11-week course. For each week of class, students focused on a major topic, watched the
video lecture, completed quizzes, and discussed in the forum. A screenshot of the course discussion forum is shown in Fig. 1.

![Screenshot of discussion forum in the course Mathematic Thinking Method](image)

**Fig. 1.** Screenshot of discussion forum in the course Mathematic Thinking Method

## 4 Methods

### 4.1 Taxonomy of Cognition

Among the taxonomy of cognition, Chi’s ICAP framework [15] has been widely used to discriminate cognitive behaviors and interpret the learning result [10, 17]. The framework classified the learning behaviors into three categories, which is, active, constructive and interactive behaviors. The active behaviors infer to the degree that students engaged in the learning process. The constructive behaviors indicate how much students produce new information beyond the presented materials. And the interactive behaviors involve interaction and cooperation with partners. We choose this framework as the basis of our discourse classification algorithm and extend it into a three-level classification method.

For cleaning posts of futility in cognitive classification, we firstly distinguish on-task and off-task discourse in the dataset. Off-task posts refer to those discourses that talking about administrative issues about the course [10], or only have contents irrelevant to academics and nonsense words, such as greeting, self-introduction and emoticons. Within the on-task labeled posts, we drew on constructs of coding scheme on cognitive domain as articulated by Wang [10], based on Chi’s ICAP (Active-Constructive-Interactive) framework [15]. For suiting our analysis and for convenience of labeling, we adjusted the coding scheme and determined abbreviated definitions of each category in Table 1.
Table 1. Cognition coding examples

<table>
<thead>
<tr>
<th>Cognition level 1</th>
<th>Cognition level 2</th>
<th>Cognition level 3</th>
<th>Abbr.</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Off-task</td>
<td></td>
<td>Off.</td>
<td></td>
<td>Talk about administrative issues about the course, contents irrelevant to academics and nonsense words</td>
</tr>
<tr>
<td>On-task</td>
<td>Active</td>
<td>Provide simple/repeat answer/opinion</td>
<td>Act. A/O</td>
<td>The student states an opinion or provides an answer without reasons, or just repeats the information that’s already covered in the course material</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ask simple question</td>
<td>Act. Q</td>
<td>The student proposes a question that just repeats the information given in the material without his/her own understanding</td>
</tr>
<tr>
<td>Constructive</td>
<td>Provide reasoned answer/opinion</td>
<td>Con. A/O</td>
<td></td>
<td>The student supports his/her answer or opinion with evidence, e.g., giving examples, comparing or connecting with external resources</td>
</tr>
<tr>
<td></td>
<td>Ask novel question</td>
<td>Con. Q</td>
<td></td>
<td>The student proposes a novel question based on his/her own understanding</td>
</tr>
<tr>
<td>Interactive</td>
<td>Acknowledgement or expand on</td>
<td>Int.</td>
<td></td>
<td>The student acknowledges others’ statements, or expand on them</td>
</tr>
<tr>
<td></td>
<td>Defend and challenge</td>
<td></td>
<td></td>
<td>The student challenges others’ ideas, or defends his/her own ideas, when there is a disagreement</td>
</tr>
</tbody>
</table>

4.2 Post Processing

4.2.1 Data Wrangling

As the first step in processing forum post data, we aim to extract the useful discourse information from every post by:

- Removing all punctuations except for question marks (probably a question), quotation marks (probably a repeat) and exclamation mark (probably an Interactive post)
- Converting all URLs and images to text notation (probably providing reasons)
- Removing all emails and emoticons
- Removing all stopwords from an aggressive stopword list
- Stemming all words
- Segmentation to get bag of words

After the data wrangling, we will get enough formatted information for feature extracting.
4.2.2 Labeling
In supervised learning, ground truth is significant for training, which requires expertise to distinguish each category. Hence we introduce distinctive features for each category high distinction that can be recognized via some simple features, as list in Table 2.

<table>
<thead>
<tr>
<th>Cognitive behavior</th>
<th>Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Off-task</td>
<td>Could be indicated by administrative words, e.g., “homework”, “submit”, “download”, and nonsense words, e.g. “hah”, “lol”</td>
</tr>
<tr>
<td>Provide simple/repeat answer/opinion</td>
<td>Could be indicated by quotation marks, a short length of text, and a simple statement without details</td>
</tr>
<tr>
<td>Ask simple question</td>
<td>Could be indicated by question and quotation marks, topics about “what” and “where”, and a short length of text</td>
</tr>
<tr>
<td>Provide reasoned answer/opinion</td>
<td>Could be indicated by longer length and cognitive action verbs, e.g., “propose”, “imagine”, as well as images and URLs</td>
</tr>
<tr>
<td>Ask novel question</td>
<td>Could be indicated by question marks, topics about “why” and “how”, and a longer length of text</td>
</tr>
<tr>
<td>Interactive</td>
<td>Could be indicated by longer length and interactive action verbs, e.g., “agree”, “disagree”, as well as exclamation mark</td>
</tr>
</tbody>
</table>

Certainly, a post may contain multiple cognitive behaviors, but it makes no difference for analysis with two or more labels on one post because we calculate all behaviors distribution and put them in model, and for the sake of accuracy, we advise to do so.

4.2.3 Feature Extracting
On Cognition Level 1, we take the following features to make a distinction between on-task and off-task posts:

- The number of the top 30 linguistic words for content-related and non-content-related posts proposed by Cui [18]
- The number of words in the post

On Cognition Level 3, for identifying cognition related words, we adopt the action verbs for each level of revised Anderson and Krathwohl cognitive taxonomy [19] and extract the features below:

- The number of words in the post. Kovanović [20] pointed out the longer the message is, the higher the chances are for the message to display higher levels of cognitive presence
- The number of cognitive action verbs on each level
- The number of each punctuation mentioned in preprocessing section
- The number of text notation of URLs and images
- The number of the first person pronouns
• The number of the second person pronouns (probably an Interactive post)
• The number of upvotes (suggesting a gainful post)

Experimental results of Sect. 5 confirm that these feature choices can achieve more accurate classification on Cognition Level 3.

4.2.4 Classification Modeling
For binary classification of on-task and off-task, we designed 3 methods including SVM, Bayesian model and logistic regression model on the following three courses: Education How to Learn Math (DS1), Medicine Sci Write Fall2013 (DS2) from Stanford dataset and Mathematic Thinking Method (DS3) from SETC dataset. Table 3 shows that on the forum data of these courses, the logistic regression model can achieve the best classification effect of AUC over 0.8, which illustrates that the results are within a reasonable range for our further analysis.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SVM</th>
<th>Bayesian model</th>
<th>Logistic regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>0.82</td>
<td>0.79</td>
<td>0.85</td>
</tr>
<tr>
<td>DS2</td>
<td>0.79</td>
<td>0.77</td>
<td>0.83</td>
</tr>
<tr>
<td>DS3</td>
<td>0.84</td>
<td>0.78</td>
<td>0.86</td>
</tr>
</tbody>
</table>

For predicting the on-task into categories in Cognition Level 3, considering a post may contain one or more cognitive behaviors, e.g., one asks a question immediately after he formulates his ideas or interacts with others in one post, we adopted 5 binary classifier using logistic regression model. On the forum data of the three courses, the average accuracy from 10-fold cross validation for each category has been made bold in Table 4.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Active</th>
<th>Constructive</th>
<th>Interactive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q</td>
<td>A/O</td>
<td>Q</td>
</tr>
<tr>
<td>DS1</td>
<td>0.78</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>DS2</td>
<td>0.76</td>
<td>0.75</td>
<td>0.79</td>
</tr>
<tr>
<td>DS3</td>
<td>0.82</td>
<td>0.80</td>
<td>0.83</td>
</tr>
</tbody>
</table>

4.3 Cognitive Gains Model
4.3.1 Model Design
When a student posts a comment or replies to someone in a discussion thread, his cognitive behavior influences both how much others gain from his post and what cognition level others can behave following his post in this thread. To analyze
cognitive gains in OLC discourses, we introduce a new cognitive gain model to analyze student’s cognitive gains in fineness of threads.

We define $f_{u,t,i}$ in (1) to be the probability of user $u$ posting on cognitive behavior $c$ in thread $t$, where $p \in P_{u,t}$ captures all the posts made by user $u$ in thread $t$, $c_i \in C$ denotes cognition $c_i$ in the set $C$ of all the categories in Cognition Level 3 and $\varphi_{p,i} \in [0, 1]^{[P] \times |C|}$ is post-cognition distribution resulting from 6 binary classifier, giving the probability that the cognitive behavior in post $p$ contains $c_i$.

$$f_{u,t,i} = \log \left( 1 + \sum_{p \in P_{u,t}} \varphi_{p,i} \right)$$ (1)

To analyze cognitive gains of each user, we should know how much is the possibility that the spread of cognition from user $u$ to user $v$ in the course, that is, the probability of user $u$ replying to user $v$ if user $v$ makes a post. Hence we set $r_{u,v}$ in (2) to quantify it.

$$r_{u,v} = \frac{\sum_t n_{u,v,t}}{N_u}$$ (2)

Since in some cases the number of times that user $u$ replies to user $v$ may be so small that the probability calculated in (2) would be unrepresentative, we introduce a heuristic definition instead: in the same thread $t$, if user $u$ makes a post later than user $v$, we define that user $u$ replies to user $v$, denoted by $n_{u,u,t}$. Let $N_u = \sum_t P_{u,t}$ be the total times that user $u$ posts in the course.

By now, the total cognitive gains of user $u$ in thread $t$ can be modeled as:

$$G_{u,t} = \sum_{c_i \in C} g_{u,i} \log \left( 1 + \sum_v r_{v,u} f_{v,t,i} \right)$$ (3)

Here, $r_{v,u} f_{v,t,i}$ captures the theoretic amount of cognitive response provided from user $v$ to user $u$ in thread $t$. And further on, we can get the gains of each user as $G_u = \sum_t G_{u,t}$ and sort out the threads by $G_t = \sum_u G_{u,t}$ where users obtain the most gains. We adopt a gain rate $g_{u,i}$ from the spread of others’ cognitive behavior, considering that student $u$ can obtain different degrees of cognitive gains from different cognitive behaviors $c_i$. The selection of gain rate is discussed in next section.

### 4.3.2 The Selection of Gain Rate

To evaluate the gain rate of each cognition, we take the test scores of students into two parts according to answering time: the scores before and after students participating the OLC’s discussion. And then we use Huang’s approach [21] to simplify the evaluation computing of gain rate.
Firstly, we model $s^{(T)}_{u,k,i}$ as a Bernoulli random variable, which is observation binary test score of the student $u$ answering the question $q_k$ at time instance $T$ in terms of cognition $c_i$, where $c_i \in C$:

$$s^{(T)}_{u,k,i} \sim \text{Ber}\left(p^{(T)}_{u,k,i}\right)$$

$$(4)$$

$$p^{(T)}_{u,k,i} = \Phi\left(a_{k,i}c^{(T)}_{u,i} - b_{k,i}\right)$$

$$(5)$$

Here, $\Phi$ is the inverse logit link function $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathcal{N}(t) dt$, where $\mathcal{N}(t) = 1/\sqrt{2\pi}\exp(-t^2/2)$ is the standard normal distribution. Let #C be the number of cognition in level 3, and the $c^{(T)}_{u,i}$ represents student $u$’s latent cognition state of cognition $c_i$ at time instance $T$. The $a_{k,i}$ and $b_{k,i}$ are properties of question $q_k$ in terms of cognition $c_i$ that can be estimated using Item Response Theory (IRT). For notational simplicity, we will omit the student index $u$ and cognition index $i$ in the following formulas, e.g., the quantities $s^{(T)}_{u,k,i}$ and $c^{(T)}_{u,i}$ are replaced by $s^{(T)}_{k}$ and $c^{(T)}$. Hence, the likelihood of an observation $s^{(T)}_{k}$ can be written as:

$$P\left(s^{(T)}_{k} | c^{(T)}\right) = p^{(T)}_{k} \left(1 - p^{(T)}_{k}\right)^{1-s^{(T)}_{k}}$$

$$(6)$$

Then, we model the latent state transition between time instance $T - \tau$ and $T$ as:

$$P\left(c^{(T)} | c^{(T-\tau)}, g\right) = \mathcal{N}\left(c^{(T)} | g^\top c^{(T-\tau)}, \tau\sigma^2\right)$$

$$(7)$$

Where $\mathcal{N}(x|\mu, \sigma^2)$ represents a Gaussian distribution with mean $\mu$ and covariance $\sigma^2$. The $g$, shorthand for $g_{u,i}$, is student $u$’s gain rate of the cognition $c_i$. The covariance $\sigma^2$, shorthand for $\sigma^2_{u,i}$ characterizes the uncertainty induced in the student $u$’s cognition state transition by acting the $c_i$ behavior.

Therefore, we can estimate the student’s cognition state after participating in the discussion and gain rate parameter through the student’s history answers of tests. And for simplifying the computing, we use the approximate result by the following methods:

$$P\left(c^{(T)}, g | s^{(1:T)}_{1:m_T}\right) \propto P\left(s^{(1:T)}_{1:m_T} | c^{(T)}, g\right) \cdot P\left(c^{(T)}, g\right)$$

$$(8)$$

Where $m_T$ is the number of tests that student have done at time instance $T$.

We assume that the student’s answers are independent of each other, the student previous answering will not impact on the current one. So the first item on the right side of (8) can be expressed as:
\[ P(s_{1:m}^{(1:T)}|c^{(T)}, g) = \prod_{T=1}^{T} \prod_{k=1}^{mr} P(s_k^{(T)}|c^{(T)}, g) \]
\[ = \prod_{T=1}^{T} \prod_{k=1}^{mr} \int P(s_k^{(T)}|c^{(T)}, g) \cdot P(c^{(T)}|c^{(T)}, g) \, dc^{(T)} \]
\[ = \prod_{T=1}^{T} \prod_{k=1}^{mr} \int P_k^{(T)} s_k^{(T)} \left(1 - p_k^{(T)}\right)^{1-s_k^{(T)}} \cdot \mathcal{N}\left(c^{(T)}|g^{T-T'c^{(T)}}, (T-T')\sigma^2\right) \, dc^{(T)} \]

(9)

By using the Eq. (10) and definition of \( \bar{p}_k^{(T)} \) in (11), we can simplify (9) into (12):

\[ \int \Phi(ax - b)\mathcal{N}(x|\mu, \Sigma) = \Phi\left(\frac{b - a\mu}{\sqrt{1 + a^2\sigma^2}}\right) \]  
(10)
\[ \bar{p}_k^{(T)} = \Phi\left(\frac{b_k - a_k g^{T-T'c^{(T)}}}{\sqrt{1 + a_k^2(T-T')\sigma^2}}\right) \]  
(11)
\[ P(s_{1:m}^{(1:T)}|c^{(T)}, g) \approx \prod_{T=1}^{T} \prod_{k=1}^{mr} P_k^{(T)} s_k^{(T)} \left(1 - \bar{p}_k^{(T)}\right)^{1-s_k^{(T)}} \]  
(12)

Then, by putting log on both sides, we can get:

\[ \log P(c^{(T)}, g|s_{1:m}^{(1:T)}) = \log P(c^{(T)}, g) \]
\[ + \sum_{T=1}^{T} \sum_{k=1}^{mr} s_k^{(T)} \log \bar{p}_k^{(T)} + \left(1 - s_k^{(T)}\right) \log \left(1 - \bar{p}_k^{(T)}\right) \]

(13)

With:

\[ \log P(c^{(T)}, g) = \log \mathcal{N}\left(c^{(T)}|1, \sigma^2\right) + \log g \]

(14)

Finally, our goal latent state variable \( g \) can be estimated by using BFGS-B algorithm [22] to maximize the objective function:

\[ \max_g \sum_{\mathcal{U}} \log P(c^{(T)}, g|s_{1:m}^{(1:T)}) - \gamma\|g_1\| \]

Here, we define \( \mathcal{U} \) as all the students that answer the tests between time instances 1 and \( T' \). To prevent overfitting, we impose an \( \ell_1 \)-norm penalty on \( g \). The stop condition is the difference between the two iterations is less than 0.0001.
4.4 System Design

We developed our analytics system as inspired by the Lambda architecture [23], and identify six functional phases from data aggregation to visualization (Fig. 2): (1) Data Aggregation, (2) Data Ingest, (3) Data Storage, (4) Streaming Analysis, (5) Batch Analysis and (6) Visualization. The OLCs data come from various of data sources provided by education institutions. We aggregate these data by using xAPI [24], an e-learning software specification software specification that allows learning content and learning systems to speak to each other in a manner that records and tracks all types of learning experiences. The aggregated data are temporarily stored in a Learning Record Store (LRS), hold by PostgreSQL in our system. Periodically, Sqoop (http://sqoop.apache.org/) imports raw data from PostgreSQL into Hadoop HDFS for permanent storage. For streaming analysis, we ingest high velocity forum data from relational database by using Flume (http://flume.apache.org/), and then forward to Spark Streaming for preprocessing. The preprocessed data store back into HDFS via Flume agent for further analysis, such as batch analysis. Apache Spark is chosen as the core component for batch analysis because it is efficient in iterative computing, provides various data sources supports, and can run on Hadoop YARN with multiple programming languages. In the batch analysis step, our analytic model is executed to generate the social network of the discussion forums and calculate the students’ cognitive gains. At last, the analysis results are written back to HDFS for visualization and query.

5 Results

5.1 Student Cognitive Gains Analysis

We firstly select some typical students with their quantifiable data in the course to provide an intuitionistic view of students’ behaviors. Table 5 lists the top 3 students who obtain the highest gains and the top 3 students who obtain the lowest gains in the course, along with the number of each cognition (the most probable cognition) in their
posts and their total posts count (#Posts). An interesting observation is that some students with high posting quantity, e.g., No.32143, No. 7036 and No. 8149 student, have low cognitive gains, contrary to some previous research results which indicated that the more students posted, the higher learning gains they achieved [10]. For exploring the causes, we investigate their forum behaviors in the course. We found that the cognitive behaviors of those who have low gains are relatively concentrated in off-task and active cognition. Besides, most of the discussion threads for these students cease after their posts because no one makes comments on their thread posts, which means they receive almost no gains in these threads. These observations suggest that course instructors and TAs need to actively get involved in the discussion with these students in time.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q/A/O</td>
<td>Q/A/O</td>
<td>Q/A/O</td>
</tr>
<tr>
<td>36856</td>
<td>40.38</td>
<td>49</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>20</td>
<td>9</td>
</tr>
</tbody>
</table>
| 27405      | 33.10  | 54     | 1        | 10      | 5    | 16   | 13   | 9
| 6547       | 27.04  | 42     | 0        | 9       | 0    | 18   | 5    |
| 32143      | 3.92   | 69     | 38       | 24      | 4    | 2    | 1    |
| 7036       | 4.01   | 57     | 22       | 20      | 12   | 3    | 0    |
| 8149       | 4.02   | 59     | 23       | 29      | 4    | 1    | 2    |

For intuitively presenting the students’ cognitive gains in each course, we plot the structure of students’ social network with their cognitive gains and other intuitive information, based on our system. As the Fig. 3 shows, the links refer to the interaction between the two students, and node’s size depicts the relative gains of student in the course and node’s color depth stands for the number of posts made by the student. By observing the graph, we should pay more attention to the node with deep color but small size, which indicate the student makes many posts but gains little.

To better understand the cognition trends of each student, we extend the graph in Fig. 3 and by clicking a node, a sunburst graph (Fig. 4) will be shown to represent the student’s cognitive behaviors distribution and quizzes performance every week in the course. In the graph, the black arc around shows the number of correct quizzes out of all quizzes the student answered, and each sector indicates the proportion of each cognitive behavior in cognition level 3. Instructors and TAs could find problems according to the proportion of each student’s cognitive behaviors.

5.2 Thread Cognition Analysis

Students obtain gains by posting in interested threads where they can get information from others. However, by inspecting the threads with low gains, we found that most of them have few participants or talk about off-task topics. It is for the reason of the first
post with insufficient details or asking questions that no one is interested. Hence it is necessary for students to notice how to create or participate in threads that benefit them more and for teachers to track the trends of each thread and make timely adjustments.

To students, the Table 6 below with the top 5 threads where students obtain the most cognitive gains could give inspiration. From the table, we can see the threads with high gains usually start with a novel question with enough information (e.g., images), mostly followed by constructive answers or opinions, and interactive behaviors are also contributive.

**Fig. 3.** Visualization of Students’ cognitive gains in social network graph of course Mathematic Thinking Method

**Fig. 4.** A Student’s weekly cognitive behaviors distribution and quizzes performance
To instructors and TAs, how to get involved in the threads that need help in time?

We design a thread cognitive tree graph (Fig. 5) to handle this problem, where teachers can intuitively view each thread’s cognitive changes via the circle colors (e.g., yellow refers to off-task, green refers to active, blue refers to constructive and red refers to interactive cognition). Instructors and TAs could intervene in a discussion thread when it is mostly covered by off-task and active cognitive behaviors, or a thread seems to be ignored. Also, teachers could pick out essential threads by their total gains and rank them on the top of the course discussion forum to benefit more students.

<table>
<thead>
<tr>
<th>First post in thread</th>
<th>Total gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why does the angle of the problem about billiards not give? And does the last time the ball enters the box count a collision?</td>
<td>10.45</td>
</tr>
<tr>
<td>Excuse me, I don’t understand the meaning that Xiao Ai walks up 1000 m along the vertical direction with the slope. How does she go? Is that it? ##IMAGE##</td>
<td>7.50</td>
</tr>
<tr>
<td>I would like to ask, how does the stroke theorem prove out, is it just that Euler has tried it through constant experimentation?</td>
<td>6.72</td>
</tr>
<tr>
<td>The stroke theorem should not just contain two singular point, right? A singularity should not work.</td>
<td>6.04</td>
</tr>
<tr>
<td>The problem about power supply can’t always be symmetrical. I would like to ask you whether there is better solution?</td>
<td>5.58</td>
</tr>
</tbody>
</table>

Table 6. Top 5 threads with highest total gains in the course Mathematic Thinking Method

Fig. 5. Threads’ cognition tree of course Mathematic Thinking Method

5.3 Course Forum Cognition Analysis

For overviewsing the cognitive gains obtained by all students in the whole course forum, we draw a line chart to show the trends of total students’ gains over time based on our system. As shown in Fig. 6, at the beginning and end of the course, the gains grow slower than the middle. It is due to the off-task and active cognition taking up a majority of posts at the beginning and less posts in vacations of the International Labor Day (May 1st) at the end. By intervening the preliminary posting patterns, such as increasing the number of higher level cognitive posts, stopping off-task topic threads and ranking gainful threads on top, instructors and TAs could increase the slope of
students’ gains. Also, the slope could indirectly reflect how deep students understand the lecture in every week, by which teachers could adjust the content or difficulty of the lecture.

6 Conclusions

In this paper, we propose a cognitive gains model on social learning network to represent students’ cognitive benefits through interaction. With an easy-labeled coding scheme and common posting features defined in our model, teachers are able to conveniently and accurately classify students’ postings into six cognitions (off-task, active question, active answer/opinion, constructive question, constructive answer/opinion and interactive cognition). The classification results on our OLC dataset show that the model is viable to estimate student’s gains from conversation in cognitive level. Moreover, we have integrated the model into our learning analytics system for enabling instructors and TAs to assess the cognitive performance of students through three perspectives (student, thread and course).

Acknowledgement. This paper is supported by the NSFC (61532004), State Key Laboratory of Software Development Environment (Funding No. SKLSDE-2017ZX-03).

References


17. Marzouk, Z., Rakovic, M., Winne, P.H.: Generating Learning Analytics to Improve Learners’ Metacognitive Skills Using nStudy Trace Data and the ICAP Framework, in LAL@ LAK, pp. 11–16 (2016)


24. xAPI: https://www.adlnet.gov/xAPI