Chen, W. et al. (Eds.) (2017). Proceedings of the 25th International Conference on Computers in Education.
New Zealand: Asia-Pacific Society for Computers in Education

Cross Analytics of Student and Course Activities from e-Book Operation Logs

Atsushi SHIMADA[a*] & Shin'ichi KONOMI[b]
[a]*Faculty of Information Science and Electrical Engineering, Kyushu University, Japan*
[b]*Faculty of Arts and Science, Kyushu University, Japan*
*atsushi@ait.kyushu-u.ac.jp

Abstract: In this paper, we propose a cross analytics methodology of student activities and course activities using e-Book operation logs collected in 15 courses with face-to-face lecture style over 4 weeks. These courses commonly use the same lecture materials, but are conducted by different teachers. The new aspect of our research is that we perform cross analysis over courses. Most past researches focus on students' activities in a specific course, and give discussions about how the students behaved, how the behaviors differ from each other. In contrast, our research focuses on the course activities and conducts a comparison among courses. First, we begin with data alignment for row data to rectify a student activity every 10 seconds. Through our analytics, it becomes clear that whether students' activities varies with teachers or their teaching styles. In the experiments, we applied the proposed analytics to 1.1-million operation logs, and found out interesting characteristics through the comparison across courses.

Keywords: Learning analytics, educational data mining, e-Book logs, learning activity

1. Introduction

Much attention has been paid to learning analytics, which is defined as the measurement, collection, analysis, and reporting of data about learners and their context for the purpose of understanding and optimizing learning and environments in which it occurs (https://solaresearch.org/). Thanks to digital learning systems such as Blackboard and Moodle, large scale educational data can be easily collected. This has changed the trend of research from data measurement and collection to analytics and feedback for improvement of teaching and learning.

One of the major studies of educational data analytics is to understand students' activities during a course period. Many studies have focused on clickstream data collected from Massive Open Online Courses (MOOCs), and analyzed the data for prediction of course completion (Crossley 2016), understanding students' learning path (Davis 2016), discovering relationship between activities and culture (Liu 2016), and change detection of students' behavior (Park 2017). Another possible educational data is available on an e-Book system (a digital textbook system) and e-Learning system. For example, the data was analyzed for pattern mining of preview and review activities (Oi 2015), understanding learning behavior of students (Yin 2015), browsing pattern mining (Shimada 2016), and performance prediction (Okubo 2016). As introduced above, learning analytics has been addressed from various aspects, but mainly has been focused on a specific course.

In this paper, we focus on e-Book operation log data collected from approximately 2,700 students across 15 courses over 4 weeks. There are approximately 1.1 million operation logs that record students' time-series activities during lecture periods. It must be noted that all the courses follow the same course design, but are conducted by different teachers. Therefore, it is possible to analyze students' activities not only within a specific course, but also across courses. The aim of this study is to investigate whether students' activities varies with teachers or their teaching styles. First, we apply a data alignment technique that can cleanse data for the ease of analytics. Then, we analyze the browsing speed of each individual student. Following this, we apply a simple anomaly detection method to the time-series operation logs to discover changes over time for each student. We summarize the individual and course activities during a lecture in the form of an 8-dimensional activity vector containing e-Book operation features, anormal states, and browsing speeds.

2. Datasets of e-Book Operation Logs

The dataset that we used in our study was collected from the Kyushu University e-Learning system and e-Book system. The target course is a series of lectures that constitute the "Primary Course of Cyber Security," which commenced in April 2017. All first-year students (more than 2,600 students, including both arts and science students) in Kyushu university are required to take this course. All students have their own laptops and bring them to access the e-Learning system and e-Book system during the lecture every week. There are a total of 15 courses (approximately 180 students/course on an average), in which teachers use the same lecture materials.

Table 1 shows the detailed course information: course id, teacher, and number of students. Note that five teachers are assigned two lectures each (e.g., Te01 is assigned C01 and C04). We analyzed e-Book operation logs collected over a 4-week period (approximately 1 million operation logs). For courses C08 and C13, which were conducted by Te06, the number of operation logs in the 3rd and 4th week eventually decreased. According to the teacher, the reason is that the e-Book system did not work well in the 3rd week. Thus, the teacher gave students a PDF version of the lecture material. In the 4th week, the teacher presented a different topic from that prescribed for the other courses. We ignored the logs of these two courses in the 3rd week and 4th week in our analysis.

Table 1: The details of course information and the number of e-Book logs in each course

			Operation Logs				
Course	Teacher	Students	1st week	2nd week	3rd week	4th week	Course Subtotal
C01	Te01	208	14,186	12,927	13,131	10,495	50,739
C02	Te02	185	11,839	42,130	33,777	28,859	116,605
C03	Te03	208	23,267	30,820	50,910	48,107	153,104
C04	Te01	157	9,092	28,301	22,082	15,985	75,460
C05	Te04	192	18,862	9,185	18,460	13,571	60,078
C06	Te05	175	7,423	4,418	14,777	12,506	39,124
C07	Te03	175	16,029	18,702	42,674	39,029	116,434
C08	Te06	200	18,983	22,737	1,541	235	43,496
C09	Te02	137	2,473	20,343	20,689	15,167	58,672
C10	Te07	176	21,195	19,361	24,215	23,721	88,492
C11	Te08	144	33,890	22,698	27,486	25,488	109,562
C12	Te05	180	9,246	3,825	10,848	8,498	32,417
C13	Te06	178	22,564	19,103	304	18	41,989
C14	Te09	213	20,002	6,568	3,808	17,319	47,697
C15	Te10	171	24,482	30,891	26,968	20,773	103,114
	Week Subtotal		253,533	292,009	311,670	279,771	1,136,983

3. Proposed Method

3.1 Data Alignment

When an e-Book is operated, its timestamp, user id, material id, page number and operation name are automatically recorded as an operation event. There are many types of operations; for example, OPEN indicates that a student has opened the e-Book file and NEXT indicates that the student has clicked the next button to move to the subsequent page. Students can bookmark a specific page, highlight selected characters, and make notes on a page. These operations correspond to the events ADD BOOKMARK, ADD MARKER, and ADD MEMO, respectively. There are two issues to consider when data analysis is performed for these event logs. First, the operation timing of students are not completely synchronized. We have to consider a slight time delay of operation. Second, the e-Book system only records the event logs and does not know a student's activities between successive event logs. We have to complement the missing activities.

Let N be the number of students in a course, and i be an index that refers to an individual student such that $i = 1, \ldots, N$. We assume discrete time, where each time-point is associated with a short period called time window. In other words, a 90-min time period is divided into short time windows, where $t = 1, \ldots, T$ is an index that runs from the first to the last time period. In our study, we define the length of time window to be 10 seconds; therefore, T equals 540 (because 90 minutes = 5,400 seconds, 5,400/10 seconds = 540 time windows).

Let $S_{i,t}$ be the state of student i at t. $S_{i,t}$ consists of five status: $S_{i,t} = (e_{i,t},\ p_{i,t}, b_{i,t}, m_{i,t}, n_{i,t})$, where each element stores the following information.

$e_{i,t}$: the number of events operated by student i at t
$p_{i,t}$: the page number browsed by student i at t
$b_{i,t}$: the number of events of type "ADD BOOKMARK" operated by student i at t
$m_{i,t}$: the number of events of type "ADD_MARKER" operated by student i at t
$n_{i,t}$: the number of events of type "ADD_MEMO" operated by student i at t

After the alignment, the status of every student can be represented by the uniformly sized vector $S_{i,t}$.

3.2 Pace Estimation

It is meaningful to know how many students open and browse the same page, and how many students are delayed. We assume that most students are following teacher's explanation and reading the same page at t, so $M_t = median(p_{i,t})$ is calculated as the page taken to be the one being read by the majority of all students. Let $v_{i,t}$ be the pace of student i at t, whose value is calculated by the following formula.

$$v_{i,t} = p_{i,t} - M_t$$

If $v_{i,t}$ is a negative value, it means that student i is behind; otherwise, the student is ahead compared with the majority of students taking the course. Figure 1 depicts a visualization of students' paces in a 90-min lecture. The blue color and the red color correspond to $v_{i,t} < 0$ and $v_{i,t} > 0$, respectively. When $v_{i,t} = 0$, the painted color is black. In the ideal case, all students open the same page at the time, and the visualization result is a black colored rectangle. In Figure 1, most of students browsed the same page at the beginning part of the lecture. From the middle part, some students began to be delayed. In our experiments below, we compared the speed status among courses across 4 weeks.

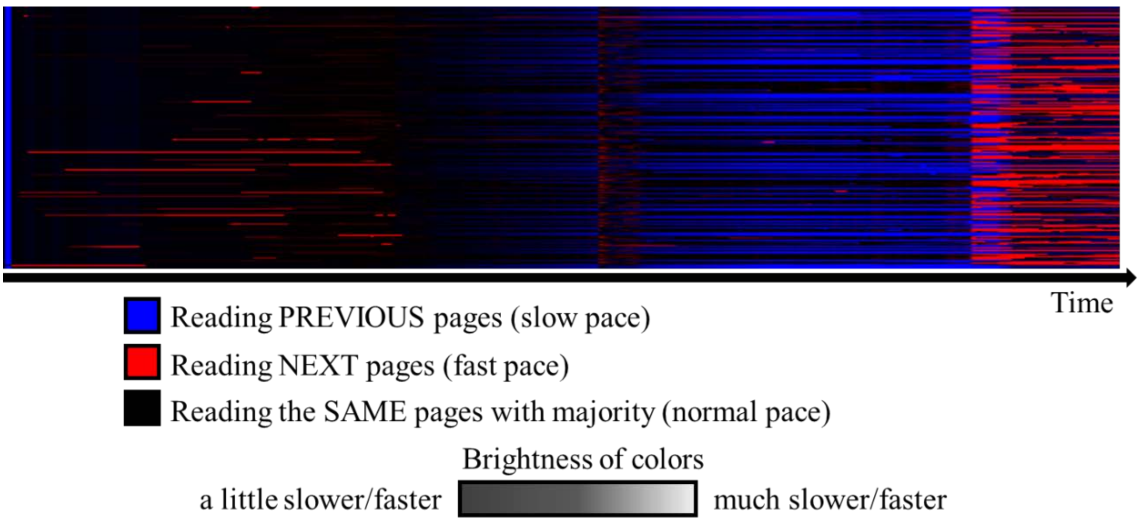

Figure 1. Visualization of browsing/reading speed in 90min lecture

3.3 Anomaly Detection

Change detection or anomaly detection is a useful technique used to discover a student who is doing something else when compared with other students. In our study, we detect two kinds of anormal states of a student. One is the anormal page view, which indicates that a student is browsing a different page from the majority. To determine whether the page view is anormal or not, we introduce a threshold th. Let $apv_{i,t}$ be the anormal page view state of student i at t, whose value is defined as follows.

$$apv_{i,t} = \begin{cases} 1, & |v_{i,t}| > th \\ 0, & |v_{i,t}| \leq th \end{cases}$$

The other is an anormal event number that indicates the state of a student with less operation logs than those of other students. Let $aen_{i,t}$be the anormal event state of student i at t.

$$aen_{i,t} = \begin{cases} 1, & e_{i,t} < me_t - se_t \\ 0, & otherwise, \end{cases}$$

where me_t and ms_t represent the mean value and standard deviation of $e_{i,t}$ of all the students in a course. The anomaly detection results are aggregated to summarize a student's activity within a 90-min lecture. The details are explained in the following section.

3.4 Student Activity and Course Activity

The individual student activity a_i is defined as

$$a_i = \left(\frac{\sum_t b_{i,t}}{\sum_t e_{i,t}}, \frac{\sum_t m_{i,t}}{\sum_t e_{i,t}}, \frac{\sum_t n_{i,t}}{\sum_t e_{i,t}}, \frac{\sum_t aen_{i,t}}{T}, \frac{\sum_t apv_{i,t}}{T}, \frac{\sum_t f(slower)}{T}, \frac{\sum_t f(normal)}{T}, \frac{\sum_t f(faster)}{T} \right),$$

where $f(X)$ is a function that computes the number of time windows that satisfy a given condition X. The three conditions are defined as follows.

$$slower = v_{i,t} < -th$$
$$normal = |v_{i,t}| \leq th$$
$$faster = v_{i,t} > th$$

The activity a_i consists of eight elements. The first three elements are ratios of three operations related to bookmark, marker, and memo. The fourth and fifth elements are the ratio of the number of time windows detected as anormal states over time. The last three elements are the ratios of time windows over time when above three speed conditions are satisfied.

The course activity C is acquired by averaging a_i in the course. A_i can be written as

$$C = \frac{\sum_i a_i}{N},$$

where the numerator calculates elementwise summation. The course activity can be used as a barometer of a specific course, and can be easily compared with those of other courses. In our study, we calculated 15 course activities over 4 weeks.

4. Experimental Results

We applied our analytics method to the rectified data (alignment data in 540 time windows) of all courses. Figure 2 shows the visualization results of speed analytics for the 4 week datasets. The red/blue color denotes that the reading/browsing speed of the student was faster/slower than the majority of the students. The brighter the color, the larger the degree of pace.

Each week, the color of speed visualizer is different among the courses. For instance, the main color of C04 is red, while C07 has a predominantly blue color. We categorized the speed characteristics into the following four types. A description is provided for each.

Type1: Most students were browsing a page at the same pace with other students. The visualized result is shown in black. For example, C08, C10, and C13 in the 1st week.

Type2: Many students were delayed in browsing. The visualized result is shown in blue color for a 90-min period. For example, C06, C12, and C14 in the 3rd week.

Type3: Many students browsed later pages as the lecture went on. The visualized color changes from blue (or dark blue) to red. For example, C02, C03, and C11 in the 1st week.

Type4: Many students were getting delayed gradually as the lecture went on. The visualized color changes from red (or dark red) to blue. For example, C04 and C13 in the 2nd week.

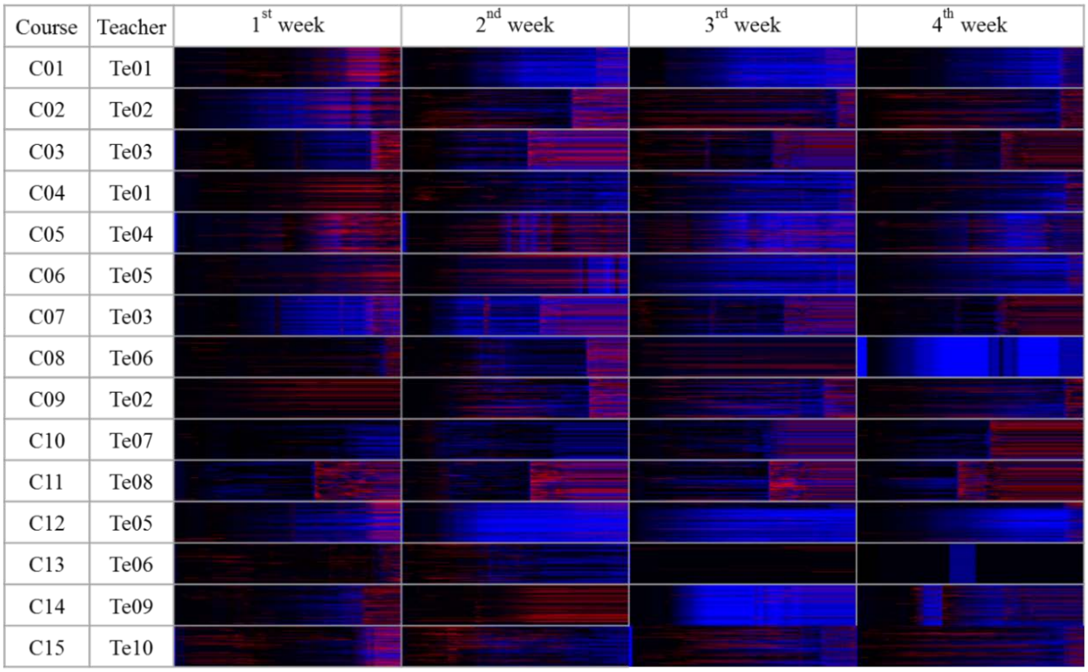

Figure 2. Comparison of speed status across courses and weeks

As summarized above, the characteristics of courses are obviously different from each other despite the fact that all of them use the same lecture organization and materials. We suppose that the teacher's personality influences the progression of a lecture. Five teachers were assigned two courses each; C01 and C04 by Te01, C02 and C09 by Te02, C03 and C07 by Te03, C06 and C12 by Te05, and C08 and C13 by Te06. Among these, courses conducted by the same teacher were found to have similar characteristics. For example, the color maps of C03 and C07 are very similar when compared with those of other courses. Across the weeks, the similar color maps were drawn in each row, which implied that the method of conducting a lecture strongly depended on the teacher.

In Table 2, a summary of eight kinds of activities (i.e., course activity C explained in Section 3.4 for each course) in the 1st week and the average of 4 weeks is shown. In the figure, a larger score in each column is highlighted using red color. The bookmark function was only used by the students in C1, C02, C03, C04, and C07 in the 1st week. The memo function was also used in limited courses. On the other hand, the marker function was used by students in all courses. This might be because the marker function is easy to use when compared with the other two functions. A student only has to trace a cursor over characters after selecting the marker function. Although the bookmark function does not require a complicated operation, a marker function is convenient in terms of marking specific sections (not a page but a content in the page). In contrast, when a student uses a memo function, it will take a

Table 2: Summary of activities in each course. Left: 1st week, Right: average of 4 weeks

Course	Teacher	bookmark	marker	memo	anormal event	anormal page view	slower	normal	faster
C01	Te01	0.00	0.03	0.00	0.25	0.62	0.18	0.62	0.20
C02	Te02	0.01	0.03	0.00	0.13	0.62	0.35	0.44	0.21
C03	Te03	0.03	0.03	0.00	0.19	0.49	0.22	0.70	0.08
C04	Te01	0.01	0.08	0.01	0.14	0.38	0.00	0.76	0.24
C05	Te04	0.00	0.01	0.00	0.38	0.75	0.19	0.60	0.21
C06	Te05	0.00	0.03	0.00	0.12	0.52	0.22	0.60	0.19
C07	Te03	0.01	0.02	0.00	0.21	0.57	0.27	0.62	0.11
C08	Te06	0.00	0.03	0.01	0.17	0.30	0.08	0.85	0.07
C09	Te02	0.00	0.03	0.00	0.03	0.28	0.00	0.75	0.25
C10	Te07	0.00	0.01	0.01	0.08	0.23	0.14	0.85	0.01
C11	Te08	0.00	0.01	0.00	0.13	0.49	0.24	0.60	0.15
C12	Te05	0.00	0.04	0.00	0.13	0.48	0.27	0.64	0.09
C13	Te06	0.00	0.02	0.02	0.16	0.28	0.07	0.86	0.07
C14	Te09	0.00	0.03	0.00	0.21	0.48	0.18	0.72	0.10
C15	Te10	0.00	0.02	0.00	0.22	0.49	0.15	0.70	0.15

Course	Teacher	bookmark	marker	memo	anormal event	anormal page view	slower	normal	faster
C01	Te01	0.01	0.07	0.00	0.17	0.49	0.23	0.66	0.11
C02	Te02	0.00	0.02	0.00	0.08	0.44	0.22	0.61	0.17
C03	Te03	0.02	0.03	0.00	0.08	0.47	0.23	0.60	0.17
C04	Te01	0.01	0.07	0.00	0.10	0.32	0.13	0.76	0.11
C05	Te04	0.00	0.01	0.00	0.29	0.74	0.27	0.48	0.25
C06	Te05	0.00	0.01	0.00	0.15	0.54	0.27	0.56	0.17
C07	Te03	0.01	0.03	0.00	0.11	0.50	0.26	0.58	0.15
C08	Te06	0.00	0.01	0.00	0.25	0.48	0.24	0.63	0.13
C09	Te02	0.00	0.01	0.00	0.11	0.43	0.19	0.64	0.17
C10	Te07	0.00	0.01	0.00	0.08	0.36	0.20	0.70	0.10
C11	Te08	0.01	0.01	0.00	0.10	0.52	0.25	0.55	0.20
C12	Te05	0.00	0.02	0.00	0.12	0.49	0.31	0.59	0.11
C13	Te06	0.00	0.01	0.01	0.21	0.35	0.09	0.84	0.07
C14	Te09	0.00	0.01	0.00	0.22	0.56	0.20	0.62	0.18
C15	Te10	0.00	0.01	0.00	0.14	0.48	0.21	0.64	0.14

longer time to type characters on the form. Therefore, a student has to complete typing as quickly as possible. Otherwise, he/she will be delayed in comparison to the pace of lecture.

The course activities throughout the 4 weeks exhibited tendencies similar to that of the activities in the 1st week. In fact, the correlation of each activity between the 1st week and average of 4 weeks was very high (0.92, 0.72, 0.85, 0.66, 0.73, 0.67, 0.65, 0.41 for each of the activities). The findings suggest that a specific course activity does not change much over the weeks, which implies that each teacher has his/her own teaching style.

5. Conclusion

In this paper, we analyzed e-Book operation logs collected from approximately 2,700 students enrolled in 15 courses over 4 weeks. We developed new analytics methodologies and applied them to the 1.1 million log data entries. The new findings from this study are summarized below. The approach to teaching differed even though teachers followed the same course design and used the same lecture materials. Each teacher tended to keep his/her own teaching style over weeks.

In future work, we plan to analyze the analytics of e-Book operation logs across other courses over years to establish the ease of use of the proposed method for long-term and wide-spread educational data. In addition, we will develop a feedback system for teachers and students. A feedback regarding real-time anomaly detection will be useful for teachers to change the pace or provide detailed explanation during lectures. It will be helpful for students to know the activities of other students and compare self-activity with other students. Through feedback using quantitatively represented learning activities, we would like to improve education in the big data era.

References

Crossley, S., Paquette, L., Dascalu, M., D. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. *LAK'16*, 6-14.

Davis, D., Chen, G., Hau, C. & Houben, G.-J. (2016). Gauging MOOC learners' adherence to the designed learning path. *EDM Conference*, 54-61.

Lee, D., Seung, H. (1999). Learning of the parts of objects by non-negative matrix factorization. *Nature*, 788–791.

Liu, Z., Brown, R., Lynch, C., Barnes, T., Baker, R. S., Bergner, Y. and McNamara, D. S. (2016). MOOC learner behaviors by country and culture; an exploratory analysis. *EDM Conference*, 127-134.

Monti, S., Tamayo, P., Mesirov, J, & Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *MACHINE LEARNING*, 91-118.

Oi, M., Okubo, F., Shimada, A., Yin, C. & Ogata, H. (2015). Analysis of preview and review patterns in undergraduates' e book logs. *the 23rd International Conference on Computers in Education*, 166–171.

Okubo, F., Yamashita, T., Shimada, A. & Ogata, H. (2017). A Neural Network Approach for Students' Performance Prediction. *LAK'17*

Park, J., Denaro K., Rodriguez, F., Smyth, P., & Warschauer, M. (2017). Detecting Changes in Student Behavior from Clickstream Data. *LAK'17*.

Roschelle, J., Dimitriadis, Y., Hoppe, U (2013). Classroom orchestration: Synthesis. *Journal of Computers & Education*, 523-526.

Shimada, A., Okubo, F., Ogata, H. (2016). Browsing-Pattern Mining from e-Book Logs with Non-negative Matrix Factorization. *the 9th International Conference on Educational Data Mining*, 636–637.

Yin, C., Okubo, F., Shimada, A., Hirokawa, S., Ogata, H. &Oi, M. (2015). Identifying and analyzing the learning behaviors of students using e-books. *the 23rd International Conference on Computers in Education*, 118–120.