Optimizing Assignment of Students to Courses based on Learning Activity Analytics

Atsushi Shimada
Department of Advanced Information Technology
Kyushu University, Japan
atsushi@ait.kyushu-u.ac.jp

Kousuke Mouri
Institute of Engineering
Tokyo University of Agriculture and Technology, Japan
mourikousuke@gmail.com

Yuta Taniguchi
Department of Advanced Information Technology
Kyushu University, Japan
taniguchi@ait.kyushu-u.ac.jp

Hiroaki Ogata
Academic Center for Computing and Media Studies
Kyoto University, Japan
hiroaki.ogata@gmail.com

Rin-ichiro Taniguchi
Department of Advanced Information Technology
Kyushu University, Japan
rin@ait.kyushu-u.ac.jp

Shin'ichi Konomi
Faculty of Arts and Science
Kyushu University, Japan
konomi@artsci.kyushu-u.ac.jp

ABSTRACT

In this paper, we focus on optimizing the assignment of students to courses. The target courses are conducted by different teachers using the same syllabus, course design, and lecture materials. More than 1,300 students are mechanically assigned to one of ten courses taught by different teachers. Therefore, mismatches often occur between students' learning behavior patterns and teachers' approach to teaching. As a result, students may be less satisfied, have a lower level of understanding of the material, and achieve less. To solve these problems, we propose a strategy to optimize the assignment of students to courses based on learning activity analytics. The contributions of this study are 1) clarifying the relationship between learning behavior pattern and teaching based on learning activity analytics using large-scale educational data, 2) optimizing the assignment of students to courses based on learning behavior pattern analytics, and 3) demonstrating the effectiveness of assignment optimization via simulation experiments.

Keywords

Student assignment to courses, optimization, learning activity analytics

1. INTRODUCTION

Due to the widespread use of digital learning environments in education, collecting large-scale educational data has become easier in recent years. For example, online course educational systems such as Massive Open Online Courses (MOOCs) generate clickstream data from users who access the course websites. E-Learning systems such as Blackboard [5] and Moodle [9] record clickstream data when users submit reports, access materials, complete quizzes, etc. Educational data can also be extracted from e-Book systems (digital textbook systems), which provide precise logs of actions such as page movement, bookmarks, highlights, text memos, and so on. These large-scale educational data play a crucial role in the research domains of learning analytics and educational data mining.

Learning analytics is defined as the measurement, collection, analysis, and reporting of data about learners and their contexts for understanding and optimizing learning and the environments in which it occurs [1]. Various studies thus far have focused on learning analytics, including learning activity analysis [25], identifying at-risk students [17, 21], understanding learning paths [7], pattern mining [15], performance prediction [6, 14], and learning support [20].

In this paper, we focus on optimizing the assignment of students to courses. The optimization of assignment is often discussed for the purpose of timetabling problem [2, 18], teacher assignment to courses[8, 16], student assignment to courses [12, 19], and so on. The objective is to reduce the time consuming cost of educational office persons and faculty members, or to maximize the satisfaction of students and teachers. For these reasons, assignment problem is often applied to multi-different courses with consideration of the classroom capacities and preference of students and teachers. In contrast to these existing studies, the target courses of our study are conducted by different teachers using the same syllabus, course design, and lecture materials. More than 1,300 students are mechanically assigned to one of ten courses taught by different teachers. Therefore, mismatches often occur between students' learning behavior patterns and teachers' approach to teaching. As a result, students may be less satisfied, have a lower level of understanding of the material, and achieve less. To solve these problems, we propose a strategy to optimize the assignment of students to courses based on learning activity analytics.

The research questions and contributions of this study are

Atsushi Shimada, Kousuke Mouri, Yuta Taniguchi, Hiroaki Ogata, Rin-Ichiro Taniguchi and Shinichi Konomi "Optimizing Assignment of Students to Courses based on Learning Activity Analytics" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 178 - 187

summarized as follows.

Research questions:

RQ1. Are learning activities common among courses or characterized by each individual course?

RQ2. Does better matching between the learning behavior pattern and teaching improve students' performance?

Contributions::

C1. Clarify the relationship between learning behavior pattern and teaching based on learning activity analytics using large-scale educational data.

C2. Optimize the assignment of students to courses based on learning behavior pattern analytics.

C3. Demonstrate the effectiveness of assignment optimization via simulation experiments.

In this paper, we review related research in the section 2 and then provide an overview of the proposed method including information about courses and the dataset in the section 3 The section 4 and section 5 discuss in detail the proposed method and strategy, and are followed by the discussion and conclusion in the section 6.

2. RELATED WORK

Optimization of assignment problem has been applied to several applications; such as timetabling problem [2, 18], classroom allocation problem [22], teacher assignment to courses [8, 16], student assignment to courses [24, 4, 12, 19], student grouping problem [11].

Elloumi et al. [2] defined the exam timetabling problem as the scheduling of exams to time slots, and the assignment of a set of exams to available classrooms. The objective was addressed to minimize the total capacity of the assigned classrooms. Phillips et al. [18] tackled the classroom assignment problem of university course timetabling. They solved an exact integer programming model for room assignment to get a Pareto optimal solution with respect to several solution quality measures on data from the university. Thongsanit [22] solved the classroom allocation problem. The number of students, the period of each course, the capacity of each classroom were used for optimization. Excel premium solver was applied to solve the problem.

Domenech et al. [8] solved the problem of teacher assignment to courses, taking teachers' preference into consideration. They developed a mixed integer linear programming model to balance teachers' teaching load and to maximize teachers' preference for courses. Ongy [16] also dealt with the teacher assignment problem to specific sections of particular courses. The assignment was solved to maximize the matching between teachers' competency to a specific subject. A mathematical model of the assignment process was formulated using mixed-integer programming.

Varone et al. [24] tackled the problem of course scheduling and assignment of students. They addressed students' preference for each course, a minimum number of students required to open a course, a maximum number of students for each course. The problem was defined as a generalization of the student project allocation problem, and was solved by an integer programming problem. Ivo et al. [12] dealt with the problem of assigning students to elective courses according to their preference. They presented an integer programming model that maximizes the total student satisfaction in line with a number of different constraints. Shannon et al. [19] proposed an evolutionary algorithm for assigning students to courses. They addressed a situation where each student specified a set of courses with preference, and capacity of each course was given. The object was to maximize the overall student satisfaction by assigning each student to a course as high on his/her preference as possible.

As introduced above, optimization problems are often defined as a family of integer programming problem. One of common criteria is the capacity information such as classroom size, the number of students required by each course. In addition, taking preference of students or teachers into consideration will improve the satisfaction of them. In contrast to these studies, our study focuses on compulsory courses which all students have to join. The courses are conducted by several teachers in parallel, because of the limited capacity of each classroom. In compulsory courses, considering preference of students does not make much sense. Therefore, our method introduces a matching between learning behavior pattern and teaching which are objectively observed through the analytics of learning logs, instead of using subjective preference of students. To the best of our knowledge, our study is the first case to introduce the learning activity analytics results to optimizing student assignment to courses.

3. OVERVIEW OF METHODS

3.1 Lecture Course and Dataset

The dataset used in this study was collected from e-Learning and e-Book systems. The target courses were a series of lectures that constitutes the "Primary Course of Cyber Security," which commenced in Kyushu University in April 2018. Overall, 1,354 students were assigned to one of the 10 courses in advance. The lectures were conducted by six teachers in face-to-face style over seven weeks. Teachers followed the same syllabus and used the same lecture materials in the courses. Table 1 provides detailed information on the courses: teacher, course id and number of students. Note that in each course, four teachers were assigned to give two lectures each.

Table 1: Course Information

teacher	course id	students
Te01	60ab104927	114
Te01	6b1900c56c	120
Te02	9a683161f5	171
Te02	86066cba6d	143
Te03	792efa2c1b	139
Te03	34451e8c77	129
Te04	24a65f29b6	137
Te04	dbed6c966a	140
Te05	39a67f80f4	133
Te06	65bb6224af	128

All students have their own laptops and bring them to ac-

cess the e-Learning and e-Book systems during the lecture. We collected the learning activity logs over seven weeks. 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. A total of 4,087,730 e-Book operation logs were collected.

3.2 Analytics Flow

The analytics flow of this study comprises two stages. The first stage involves extracting the analytics of learning activities and quiz scores from each course. Statistical summaries of e-book operations, the browsing time for each page, and the distribution of the quiz scores for each lecture are analyzed to gather the characteristics of the courses. We then perform further detailed analytics of learning activities over courses to investigate the relationship between learning behavior patterns and quiz scores. We will show the possibility of optimizing the assignment of students to courses based on the results of these analytics.

At the second stage, we tackle the optimization issue, aiming to match learning behavior patterns and teaching to improve students' understanding of course contents. To this end, we use students' quiz scores instead of their level of understanding of course contents. We solve the optimization problem as a generalized assignment problem. We define a new cost function to realize the best assignment. We investigate the effectiveness of our assignment of students through simulation experiments.

4. LEARNING BEHAVIOR PATTERN ANALYTICS

There are several existing approaches to analyze learning behavior patterns in Massive Open Online Courses (MOOCs)[13, 10, 3]. On the other hand, our study focuses on learning logs collected during in-class, i.e., face-to-face lecture time, and out-class activities. Therefore, we newly design a methodology to analyze learning activities of students.

4.1 Course Activity Summary

The learning logs consist of four types of datasets: e-book operation logs, lecture material information, lecture time information, and quiz scores. First, we divide the e-book operation logs into in-class activity logs and out-class activity logs by referring to the lecture time information. With ten courses in the dataset, we acquire ten sets of in-class and out-class activity logs after the division procedure. Second, the in-class and out-class activity logs are aggregated page by page. The aggregation procedure is performed for each week (for seven weeks). This allows us to analyze the page-wise activity of each week. In addition, we calculate students' browsing time on each page by subtracting the timestamps between successive page transition events. Consequently, we acquire the total length of students' browsing time for each page.

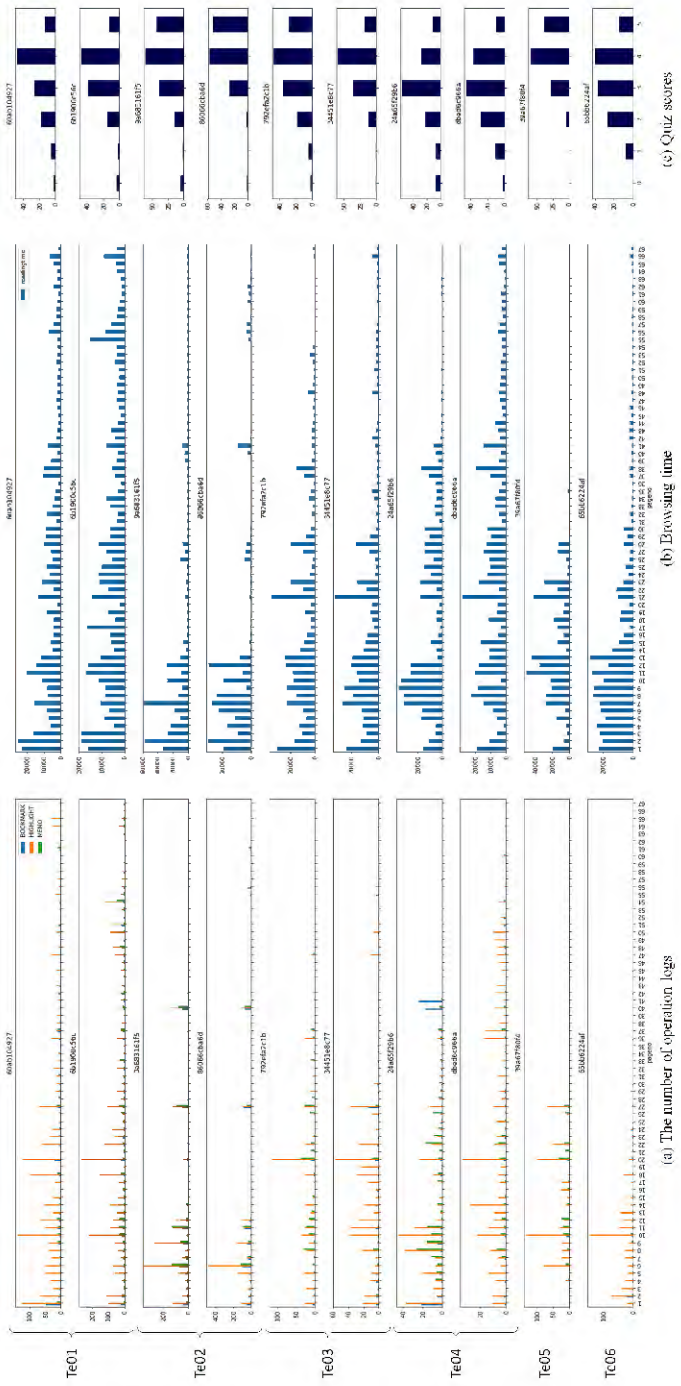

Figure 1: In-class Learning activity, browsing time and quiz scores of each course in the 1st week.

Figure 1 shows the visualization result of in-class activity, browsing time, and quiz scores for the first week (The figure is arranged with 90-degree rotation due to the page space limitation). The figure on the left displays page-wise e-book operations including "BOOKMARK," "HIGHLIGHT," and "MEMO." The horizontal axis represents the page number while the vertical axis shows the number of operations aggregated by the students. Each row corresponds to a single course. The central figure shows students' page-wise browsing time during lectures. The vertical axis of this figure is the timed duration (seconds). The figure on the right displays the quiz score distribution. After the lecture every week, students answered quizzes (averagely 5 questions). The quiz scores are normalized between 0 and 5 (full marks). The horizontal axis shows the scores and the vertical axis represents the number of students. The distributions of operations, browsing time, and quiz score are characterized for each course. For instance, the e-book operations are recorded in the former pages much more than latter pages. Regarding students' browsing time, a longer time was spent on the former pages rather than latter pages. The quiz scores are also characterized by courses. The courses in the seventh and eighth rows received lower scores compared with other courses.

Next, let us focus on sets of two specific courses conducted by the same teachers. Of six teachers, four (Te01, Te02, Te03, and Te04) have two courses, as summarized in Table 1. We can see that the visualized results are similar for courses conducted by Te01, Te02, and Te03 compared with those of other teachers. Especially in the case of Te01 and Te02, the frequency of the e-book operation logs and browsing time for e-books have common peaks. On the other hand, in the case of Te04, the distributions are not so similar between two courses compared with the cases of other teachers. Even so, the similarity of the two distributions are higher than the courses conducted by the other teachers. Figure 2 shows the summary of e-book operation usage and quiz scores of each course in the first week. In the case of bookmark, highlight and memo operations, the value represents the average usage of each operation per page. The quiz score is normalized between 0 and 1. The higher value implies that students used the operations frequently or received better quiz scores. This figure illustrates that the courses conducted by the same teachers have similar values. While we show the result of the first week only due to the page space limitation, a similar tendency was observed in the other weeks.

From the above results, we inferred the following points. First, teachers have their own teaching ways, which do not differ widely between courses. Second, students' learning activities are strongly affected by the teaching ways. To investigate these hypotheses, we further analyzed course characteristics.

4.2 Learning Activity Features

If learning activities are affected by teachers, the activities in each course should form a cluster, and the clusters related to the same teacher should have more similar features than the other clusters. For the investigation, we define a feature vector $F_{u,l}$ that represents the learning activities of student u for a lecture material l.

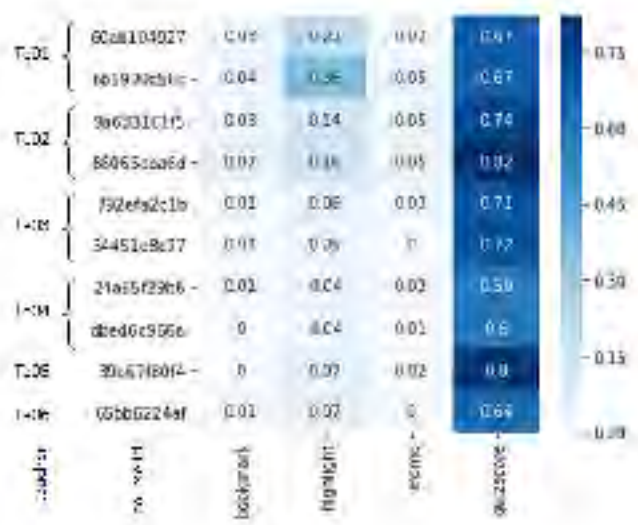

Figure 2: Learning activities in the 1st week.

To simplify the mathematical formulation, the notation u is omitted from the following explanation. Let f_p be a page-wise feature vector in page p of the lecture material. The f_p has eight elements;

$$f_p = (b_p^i, h_p^i, m_p^i, t_p^i, b_p^o, h_p^o, m_p^o, t_p^o), \quad (1)$$

where b_p^*, h_p^*, and m_p^* are the number of operation logs of "BOOKMARK," "HIGHLIGHT," and "MEMO" recorded during $(* = i)$/outside $(* = o)$ the lecture time. The t_p^i and t_p^o are the browsing time of page i during the lecture time and outside lecture time, respectively. A feature vector for a specific lecture material l containing l_N pages is defined by the concatenation of f_p as

$$F_l = (f_1, \ldots, f_p, \ldots, f_{l_N}). \quad (2)$$

For instance, when a lecture material l consists of 50 pages, the feature vector has 400 (8-dim $\times$50 pages) dimensions. Note that, in fact, the feature vector is calculated for each student u defined as $F_{u,l}$.

We apply t-SNE (t-Distributed Stochastic Neighbor Embedding) [23] to investigate the similarity and dissimilarity of feature vectors within the course and among the courses. t-SNE is a technique for dimensionality reduction. It is often used for the visualization of high-dimensional datasets. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of low-dimensional embedding and high-dimensional data. Figure 3 shows the visualization result in two-dimensional space. Courses are marked by color. We can see that the feature vectors distribute closely in the same course, while those of other courses make distinguishable clusters. From these results, we can say that learning activities are affected by teachers, as mentioned in the previous section.

4.3 Learning Activity vs. Quiz Score

Through analyzing e-book operation logs and learning activity features, we found that the learning activity itself is

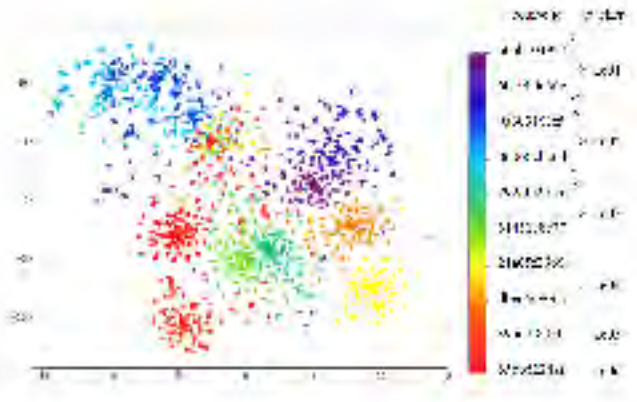

Figure 3: Visualization of feature vectors by t-SNE.

characterized by courses, i.e., teachers who conducted the lectures. On the other hand, the relationship between learning activities and quiz scores was not addressed in previous analytics. Although learning activity (i.e., feature vector $F_{u,l}$) is similar in the same course, quiz scores are distributed widely, as shown in the right part of Figure 1. Therefore, it is important to perform a relation analysis between learning activities and quiz scores.

To extract the characteristics of learning activities, we apply the k-means clustering method to the set of feature vectors $F_{u,l}$ acquired from all students. Then, we investigate the quiz score of each cluster and each course. Note that the clustering is performed for all feature vectors without considering the course id information. In other words, the learning activity features are purely analyzed to generate clusters. Afterwards, we put the course id again to each feature vector to investigate the clustering result. Figure 4 shows the result of the first week when the number of clusters was set to be 5. The horizontal axis is the course id and the vertical axis is the cluster id (from 0 to 4, totally 5 clusters). The value of each cell indicates the average quiz score. For example, in the left column and fifth row, the score is 3. This means that the students in course id "60ab104927," with learning activity in cluster id "4" received the score of 3 on an average. The cells with values of zero indicate that no student belongs to the cluster or the course. The detailed distribution of quiz scores in each cluster is shown in Figure 5. The horizontal axis is the cluster id, and the vertical axis is the number of students over courses. We can see that each cluster cannot be explained by quiz scores. Even in the same cluster, that is, even in the similar learning activity, some students received better scores while others received worse scores.

Figure 4 displays interesting and important characteristics of lectures. First, some clusters (e.g., cluster id 0) represent the characteristics of learning activities observed only in limited courses (e.g., course id "9a683161f5" and "86066cba6d"). In the case of cluster id 1, the corresponding learning activities are observed in all courses, but the average quiz scores are different. Students in course id "86066cba6d" received higher scores, while those in course id "24a65f29b6" received lower scores than in the other courses. On the other hand,

Figure 4: Average quiz scores of each cluster and each course when the number of clusters was 5 in the 1st week.

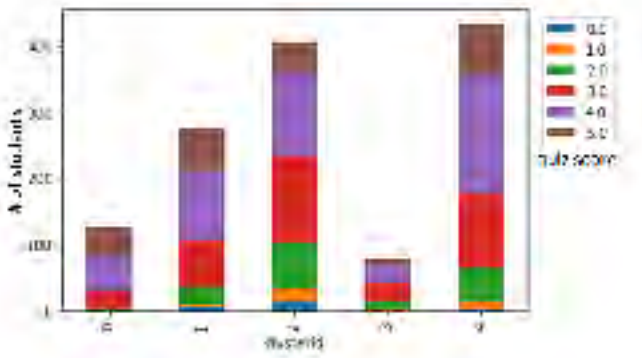

Figure 5: Quiz score description of each cluster in the 1st week.

each column also shows interesting characteristics of each course. For example, students belonging to cluster id 2 received high scores in course id "39a67f80f4," while those in cluster id 3 received lower scores. In the case of course ids "60ab104927" and "792efa2c1b," the scores of cluster id 3 are higher than those of cluster id 2. Therefore, our findings are summarized as follows. Even if learning activities are similar, quiz scores differ among courses. The characteristics of each course and its method of scoring quizzes are different for each cluster. We investigated these characteristics by changing the number of clusters from 3 to 29 (14 patterns) and found the same results. Due to length limitations, we only show additional results when the number of clusters was set to 15 in Figure 6. As the number of clusters increases, course-specific clusters appear, such as cluster id 1, 2, and 3.

5. OPTIMIZATION OF STUDENT ASSIGNMENT TO COURSES

Based on the results of the learning activity analysis and findings in the previous section, we optimize the process of assigning students to courses considering learning activities and course characteristics. In this section, we define the characteristic of courses c as the ability to give a student group g a quiz score $A_{c,g,l}$ on average for the lecture material l. Note that the lecture material l completely corresponds to each week, so that we can regard the l as the indicator of week. The c, g, and $A_{c,g,1}(l = 1)$ correspond

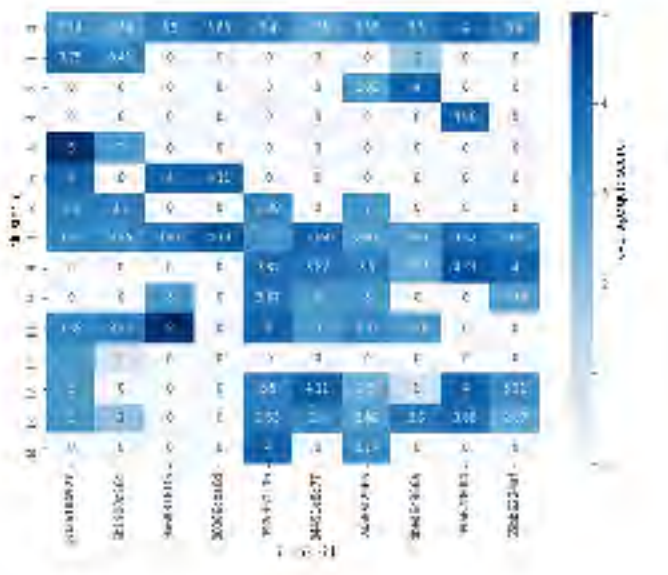

Figure 6: Average quiz scores of each cluster and each course when the number of clusters was 15 in the 1st week.

to the column, row, and element value in Figure 4, respectively. Our assumption is that students will get better quiz scores if they move to better-suited courses. For example, students in group $c =$"dbed6c966a" and $g = 3$ received an average quiz score of 2.0 (i.e., $A_{c,g,1} = 2.0$). If they could move to another course, "792efa2c1b," their quiz score would become 1.89 points higher (will receive 3.89 points on average). While this is an ideal situation, we suppose that a good match between course characteristics and learning behavior patterns will generate positive effects. Therefore, we propose an optimized strategy of assigning students to courses based on learning activity analytics.

5.1 Assignment Problem

The optimization of assignment can be considered the generalized assignment problem (GAP). The GAP is a problem in combinatorial optimization in which each agent in one set is matched to a single task in another set. Each task has a limited capacity for agents, and the goal is to minimize the sum of the costs or maximize the sum of profits. Formally, the problem can be stated as an integer programming problem.

In the case of our study, the agents and tasks can be replaced by the courses and students. The problem is:

$$\text{minimize} \sum_{c=1}^{C}\sum_{u=1}^{U} w_{c,u}x_{c,u} \tag{3}$$

$$\text{subject to} \sum_{u=1}^{U} x_{c,u} \geq S_c, \quad \text{for } c = 1,\ldots,C \tag{4}$$

$$\sum_{c=1}^{C} x_{c,u} = 1, \quad \text{for } u = 1,\ldots,U \tag{5}$$

where C is the number of courses, U is the number of students, and $w_{c,u}$ is the cost for the assignment of student u to course c. The detailed definition of $w_{c,u}$ will be explained later. The $x_{c,u}$ becomes 1 if student u is assigned to course c; otherwise, it is zero. The S_c is the minimum of students required in course c.

We define the cost $w_{c,u}$ as follows:

$$w_{c,u} = (H - A_{c,m(u)}) + b \tag{6}$$

where H is the maximum quiz score, $m(u)$ is a map function that presents the group g (i.e., cluster id) to which the student u belongs, and b is the bias term to penalize changing courses. The first term $(H - A_{c,m(u)})$ becomes smaller when student u is assigned to a course c in which student u will likely receive a higher quiz score. In other words, student u belonging to group $m(u)$ is likely to move to a course that gives higher quiz scores for group $m(u)$. Note that $m(u)$ indicates the student group (cluster id) that has a similar learning activity within the group, so that we can estimate a quiz score $A_{c,m(u)}$ for every assignment situation because $A_{c,m(u)}$ corresponds to an element in Figure 4. The bias term b gives an additional cost to constrain the course movement (change of the assignment from one course to another). If we give a large value to b, the $w_{c,u}$ also becomes large, which most likely results in students remaining in the current course. Note that the above equations are calculated in each week so that the notation l, which identifies the lecture material, hsould be put on each term such as $w_{c,u,l}, x_{c,u,l}, A_{c,m(u),l}$. In the avobe equations, we omitted the notation l to simplify the mathematical formulation.

5.2 Assignment Results

We conducted experiments to investigate the proposed assignment strategy. We varied the number of clusters from 3 to 29 (14 patterns) and the bias b from 0.0 to 1.0. We performed k-means clustering to acquire the relation matrix of the quiz scores between student groups g and courses c as shown in Figure 4. Next, we solved the generalized assignment problem by changing the bias value to 0.0, 0.1, 0.3, 0.5, and 1.0. Assignment optimization was conducted for each week individually so that we acquired a total of 490 assignment results (14 clustering patterns $\times$ 5 bias patterns $\times$ 7 weeks). In the following paragraph, we report how the assignment result changed according to the number of clusters and the strength of the bias.

First, we investigated how many students were assigned (moved) to the other courses. Figure 7 shows the assignment results when the number of clusters was 5 and the bias was 0.5. Course ids are arranged in horizontal and vertical lines. The horizontal line indicates the course id to which students belonged before the assignment, that is, the original course assigned by the university. The vertical line shows the course id to which students were assigned after the optimization of the assignment problem. The value of each cell is the number of students. For example, 137 of 139 students who originally belonged to "86066cba6d" remained in the same course, but 2 students moved to the course "9a683161f5." In the case of this result, the bias was set to be 0.5, which is a relatively strong bias to constrain course changes, resulting in students remaining in original courses (large value in diagonal element) rather than changing courses. Interestingly, a

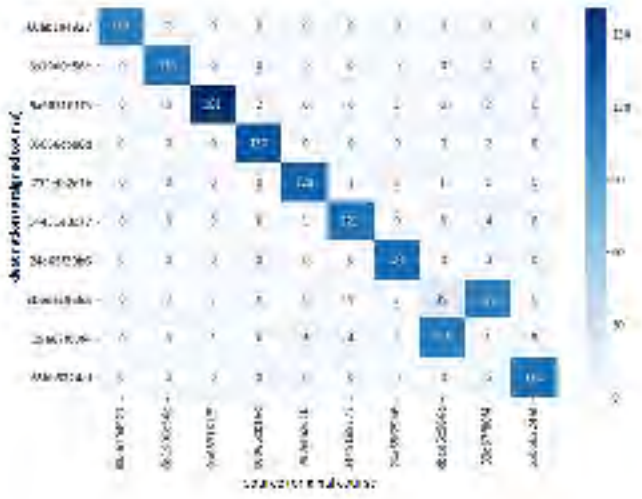

Figure 7: Movement matrix when the number of clusters was 5 in the first week.

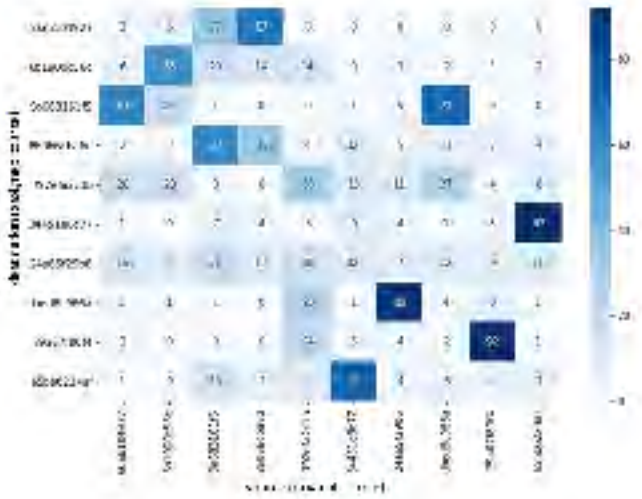

Figure 8: Movement matrix when the number of clusters was 15 in the first week.

large movement occurred between course "dbed6c966a" and "39a67f80f4." Figure 8 shows another result when the number of clusters was 15 and the bias was 0.0 (no bias). Compared with Figure 7, a larger number of students moved from their original courses to other courses.

Next, we investigated the total number of students who moved courses. Figure 9 shows the summarized result of course movement. The horizontal axis shows the number of clusters that we set when performing k-means clustering for learning activity logs. The vertical axis displays the percentage of students who were assigned to the other courses. The five lines represent the results with different values of bias. In general, as the number of clusters increased and as the value of bias decreased, many students were assigned to other courses. The larger number of clusters generated small clusters that precisely indicate the representative learning activities, which is why the flexibility of the matching between students and courses increased. In the case of no bias (bias $b = 0.0$), students moved among courses the most flexibly.

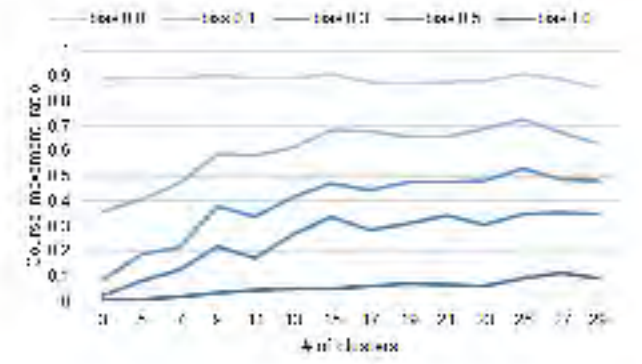

Figure 9: Course movement over 7 weeks.

The flexibility directly related to the encouragement of quiz scores. Figure 10 shows the improved quiz scores after the optimization of student assignments. The vertical axis is the value of the improved quiz score. In fact, the improvement in the quiz score $\hat{q}_{u,l}$ of each student u for the lecture material l (i.e., lth week) was calculated by:

$$\hat{q}_{u,l} = q^{org}_{u,l} - A_{c,\hat{m}(u),l} \tag{7}$$

where $q^{org}_{u,l}$ is the original quiz score of student u in lth week and $\hat{m}(u)$ is the map function that gives the group id (cluster id) to which student u was assigned. The line graphs indicate the average of score $\hat{q}_{u,l}$ of all students over seven weeks. We can see the similar tendency of the line graphs compared with Figure 9. We further investigated the improvement of quiz scores in each week. We identified three typical cases:

2nd week: w2 the original score was higher on average than in other weeks.

4th week: w4 the original score was lower on average than in other weeks.

6th week: w6 the original score was at an average level over 7 weeks.

Figure 11 shows the line graphs of three cases; the solid line and dashed line correspond to the different settings of bias value at 0.0 and 0.5, respectively. The lower the original score (fourth week), the more the score was improved. From these results, we can expect to improve the quiz scores through the optimization of student assignments. The level of improvement is affected by the number of clusters and the value of bias. In addition, the number of course movements is strongly affected by the bias.

Finally, we note again that the experiments in this section demonstrate the success of the proposed optimization strategy based on the analytics of learning activities. Although

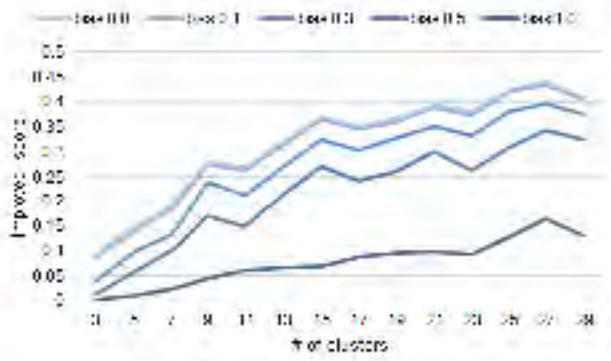

Figure 10: Score improvement after optimization of student assignment to courses.

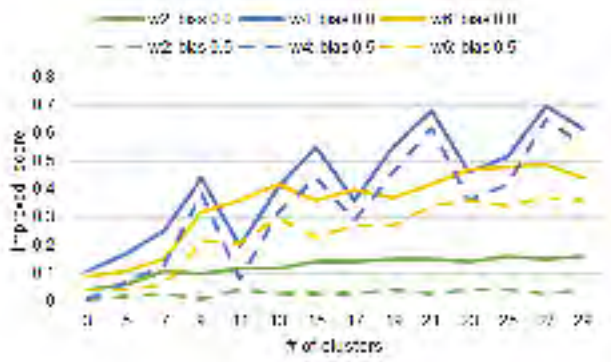

Figure 11: Three typical examples of score improvement.

the improvement of quiz score $\hat{q}_{u,l}$ is a simulated result, we can expect students to get better scores if matching between students and courses is optimized. This expectation comes from the fact that matching optimization provided better scores, as shown in Figure 4 and 6 of the previous section.

5.3 Simulation in a Realistic Situation

In this section, we conduct simulated experiments considering a more realistic situation. The proposed optimization method requires students' learning activity logs to analyze learning behavior patterns and quiz scores. Therefore, we assume that the optimization of student assignment is performed after the lecture of the first week. Using the data from learning activities and quiz scores collected in the first week, we optimize the assignment of students to courses. The assignment then remains the same after the second lecture. We simulated the quiz scores of students who were assigned to another course after the first lecture from the second to seventh week.

This simulation is difficult because although the actual quiz scores in the original courses are known, quiz scores after the optimization of assignments are unknown. Therefore, we must estimate students' quiz scores after optimization. We will now review the purpose of clustering learning activities. As concluded in the previous section, students received different quiz scores even when their learning activities were similar. After the optimization of student assignment to courses, students who received worse scores in the original course should be moved to another course in which they will receive better quiz scores. Therefore, we will focus on a student who has a similar learning activity in the course to which the target student is assigned. More specifically, let a target student be y and consider a situation where student y is assigned to course c. For all students who originally belonged to course c, we search for a student z who has the most similar learning activity to student y. Mathematically,

$$z = \arg\min_{u} |F_{y,l} - F_{u,l}| \tag{8}$$

where $F_{u,l}$ is a feature vector of learning activity of student u for the lecture material l. The lecture material l corresponds to a specific lecture, so that we can regard l as the lecture conducted in each week. Finally, we regard the original quiz score $q^{org}_{z,l}$ of student z as the estimated score of the target student y. Let $\hat{q}^{new}_{y,l}$ be the estimated quiz score of student y after the assignment. The score improvement ratio $r_{y,l}$ can be calculated by:

$$r_{y,l} = \frac{\hat{q}^{new}_{y,l} - q^{org}_{y,l}}{H - q^{org}_{y,l}} \tag{9}$$

where H is the maximum quiz score (the same with eq. 6) and q^{org}_{y} is the quiz score of student y in the original course. Note that $q^{org}_{y,l}$ is the actual quiz score and $\hat{q}^{new}_{y,l}$ is the estimated quiz score. Figure 12 illustrates the overview of the simulation strategy. In the Figure, student y is assigned to course A after the 1st week. The quiz scores from the 2nd week to 7th week have to be estimated because the student y has the quiz scores in the original course B. Our strategy explore the most similar (matched) feature vector $F_{u,l}$ from the students in course A. In the case of this figure, the learning activities of student 1 and student 2 are the best matched in the 2nd week and 3rd week, respectively. As the results, the quiz score of $q^{org}_{1,2}$ and $q^{org}_{2,3}$ are used for the estimated quiz scores of student y in the 2nd week and 3rd week, respectively.

Figure 13 shows the simulation result. The horizontal axis shows the number of clusters, and the vertical axis represents the score improvement ratio. The improvement ratio is averaged over six weeks (from the 2nd to 7th week) and over students who moved from the original course to the other course. Totally, the improvement ratio of every setting (any number of clusters, or any value of biases) was higher than zero, which means that the student assignments created positive effects for students. The scale of the effect was the largest when we set the bias b to 1.0. In contrast to the result in Figure 10, the largest value of bias provided the best result. This is because the improvement ratio is summarized by students who were assigned to the other course only. In the case of a large value of bias b, the movement from one course to another is constrained, so that a small number of students actually changed courses, as shown in Figure 9. As a result, the optimization of student assignment provided higher effects (i.e., made students receive better quiz scores) for a limited number of students. As the number of students increases (the bias decreases), the effect becomes smaller due to averaging calculation of improvement ratios. There was

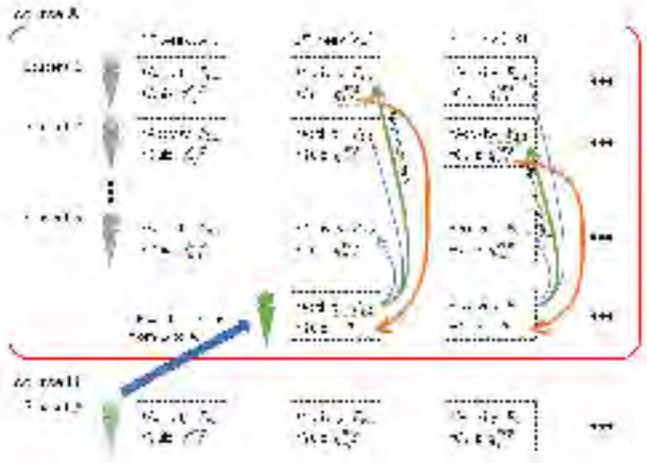

Figure 12: Overview of simulation experiment how to estimate the quiz scores in the newly assigned course

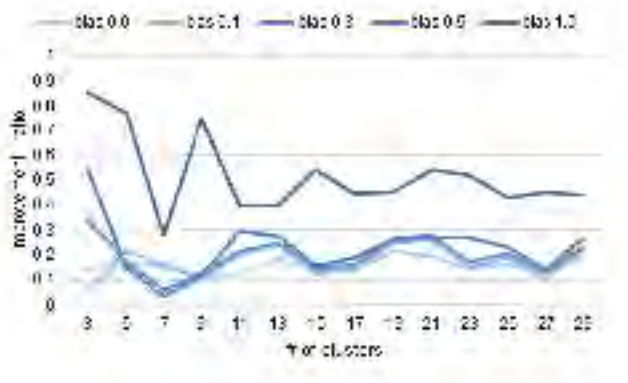

Figure 13: Average score improvement ratio over 6 weeks.

not a large difference of improvement ratios among the four settings where the bias value was 0.0, 0.1, 0.3, or 0.5 when the number of clusters was larger than five. In terms of the calculation cost and ease in explaining/interpreting learning activities, we prefer the smaller number of clusters, so that selecting the five clusters is one of the reasonable solutions.

Finally, Figure 14 shows the weekly improvement ratio when the number of clusters was fixed to be five. From the fourth (w4) to the seventh week (w7), a similar tendency was observed: The large value of bias b provided higher improvement. On the other hand, in the second (w2) and third weeks (w3), even the smaller value of bias b provided better results. We guess that the factor comes from the calculation of improvement ratio. When the original quiz scores are close to the maximum quiz score H, the denominator eq. 9 becomes smaller, resulting in a larger improvement ratio. In fact, the average of the original quiz scores in the second week was quite higher than in other weeks. In terms of maximization of the number of students who are supposed to get better quiz scores, the constraint bias should be relaxed as

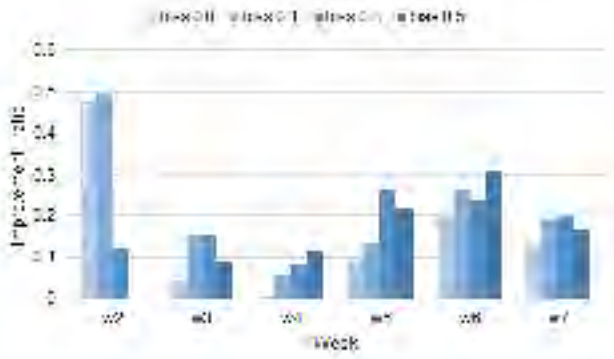

Figure 14: Average score improvement ratio of each week when the number of clusters was 5.

least as possible. Therefore, a reasonable guideline to set the number of clusters and bias is to set a smaller number of clusters (such as 5 or 7) and set a smaller value of bias (such as 0.1, 0.3, or 0.5).

6. DISCUSSION AND CONCLUSION

We proposed a strategy to optimize the assignment of students to courses based on learning activity analytics. This optimization is first intended to minimize the mismatch between students' learning behavior patterns and teachers, and second to maximize the improvement of students' quiz scores by assigning them to courses that are more suited to their learning behavior patterns. The success of these interventions are supported by the learning analytics results. Analyzing e-book operation logs and quiz scores collected from 1,354 students in 10 courses based on the same syllabus and lecture materials, we identified the following findings. From the macro perspective, learning activities are affected by teachers. Although teachers are not directly observed by teaching logs, the patterns implicitly appear as course-specific features, as shown in Figure 3. From the micro perspective, students' learning activities can be grouped into several clusters, each cluster representing a feature of such learning activities. Regardless of courses, students have similar learning activity features if they belong to the same cluster. On the other hand, quiz scores differ among students who belong to the same cluster. From these facts, we formulated the hypothesis that good matching between learning behavior pattern and teaching approach would provide better effects for students.

Our proposed approach requires learning activity logs to be acquired before students can be assigned to courses. Thus, in our experiments, we used learning logs collected in the first week and then optimized the assignment of students for subsequent weeks. Another promising solution to this matter could employ learning logs collected in other lecture courses or in past courses. If such logs are available, learning activities can be analyzed in advance and students can be optimally assigned to courses before the first week's lecture begins. This paper showed the effectiveness of an optimization strategy through the results of simulation experiments in which quiz scores improved. Meanwhile, the proposed method has another important aspect as a useful tool to

simulate the effects of assignments in advance. In future work, we will investigate the effectiveness of the optimization strategy in live settings based on effect simulations.

Acknowledgements

This work was supported by JST PRESTO Grant Number JPMJPR1505, JSPS KAKENHI Grand Number JP16H06304 and JP18H04125, Japan.

7. REFERENCES

[1] *https://solaresearch.org/* (SoLAR).

[2] Abdelkarim, E., Hichem, K., Bassem, J., and Abdelaziz, D. The classroom assignment problem: Complexity, size reduction and heuristics. *Appl. Soft Comput. 14* (Jan. 2014), 677–686.

[3] An, T.-S., Krauss, C., and Merceron, A. Can typical behaviors identified in moocs be discovered in other courses? In *The Tenth International Conference on Educational data Mining (EDM 2017)* (2017), pp. 220–225.

[4] Binyamin, K., Jon, L., and Daniel, N. Optimizing the assignment of students to classes in an elementary school. *INFORMS Transactions on Education* (2014), 38–44.

[5] Bradford, P., Porciello, M., Balkon, N., and Backus, D. The Blackboard Learning System: The Be All and End All in Educational Instruction? *Journal of Educational Technology Systems 35*, 3 (Apr. 2007), 301–314.

[6] Brinton, C. G., and Chiang, M. Mooc performance prediction via clickstream data and social learning networks. In *2015 IEEE Conference on Computer Communications (INFOCOM)* (April 2015), pp. 2299–2307.

[7] Davis, D., Chen, G., Hauff, C., and Houben, G. Gauging MOOC learners' adherence to the designed learning path. In *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016* (2016), pp. 54–61.

[8] Domenech, B., and Lusa, A. A milp model for the teacher assignment problem considering teachers' preferences. *European journal of operational research 249*, 3 (Jan 2016), 1153–1160.

[9] Dougiamas, M., and Taylor, P. Moodle: Using learning communities to create an open source course management system. In *Proceedings of EdMedia: World Conference on Educational Media and Technology 2003* (Honolulu, Hawaii, USA, 2003), D. Lassner and C. McNaught, Eds., Association for the Advancement of Computing in Education (AACE), pp. 171–178.

[10] Ferguson, R., and Clow, D. Consistent commitment: Patterns of engagement across time in massive open online courses (moocs). *Journal of Learning Analytics 2* (01 2015), 55–80.

[11] Hübscher, R. Assigning students to groups using general and context-specific criteria. *IEEE Transactions on Learning Technologies 3*, 3 (2010), 178–189.

[12] Ivo, B., and Josko, M. An integer programming model for assigning students to elective courses. *Croatian Operational Research Review 6* (2015), 511–524.

[13] Kizilcec, R. F., Piech, C., and Schneider, E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *LAK* (2013), ACM, pp. 170–179.

[14] Mouri, K., Okubo, F., Shimada, A., and Ogata, H. Bayesian network for predicting studentsÂĄf final grade using e-book logs in university education. In *IEEE International Conference on Advanced Learning Technologies(ICALT2016)* (2016), pp. 85–89.

[15] Oi, M., Okubo, F., Shimada, A., Yin, C., and Ogata, H. Analysis of preview and review patterns in undergraduatesÂĄf eÂĄ]book logs. In *The 23rd International Conference on Computers in Education (ICCE2015)* (2015), pp. 166–171.

[16] Ongy, E. E. Optimizing student learning: A faculty-course assignment problem using linear programming. *Journal of Educational and Human Resource Development 5* (2017), 1–14.

[17] Park, J., Denaro, K., Rodriguez, F., Smyth, P., and Warschauer, M. Detecting changes in student behavior from clickstream data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (2017), pp. 21–30.

[18] Phillips, A. E., Waterer, H., Ehrgott, M., and Ryan, D. M. Integer programming methods for large-scale practical classroom assignment problems. *Computers & OR 53* (2015), 42–53.

[19] Shannon, C. A., and McKinney, D. An evolutionary algorithm for assigning students to courses. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference* (2011), pp. 388–393.

[20] Shimada, A., Okubo, F., Yin, C., and Ogata, H. Automatic summarization of lecture slides for enhanced student preview-technical report and user study-. *IEEE Transactions on Learning Technologies 11*, 2 (2018), 165–178.

[21] Shimada, A., Taniguchi, Y., Okubo, F., Konomi, S., and Ogata, H. Online change detection for monitoring individual student behavior via clickstream data on e-book system. In *8th International Conference on Learning Analytics & Knowledge* (3 2018), pp. 446–450.

[22] Thongsanit, K. Solving the course - classroom assignment problem for a university. *Silpakorn University Science & Technologies Journal 8*, 1 (2014), 46–52.

[23] van der Maaten, L., and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605.

[24] Varone, S., and Schindl, D. Course opening, assignment and timetabling with student preferences. In *Proceedings of the 2nd International Conference on Operations Research and Enterprise Systems* (2013).

[25] Wang, G., Zhang, X., Tang, S., Zheng, H., and Zhao, B. Y. Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), CHI '16, pp. 225–236.