Detecting Incomplete Learners in a Blended Learning Environment among Japanese University Students

doi:10.3991/ijet.v4i1.659

M. Nakayama, H. Kanazawa and H. Yamamoto
CRADLE, Tokyo Institute of Technology, Tokyo, Japan

***Abstract*—To examine the feasibility of identifying incomplete participants who had not eventually completed a course in a blended learning environment using current learning behavioral data, access log data of complete and incomplete participants were analyzed. There is a significant difference between the two sets, and the number of accesses correlates with the final test score. Discrimination analysis was conducted using several variables across the learning process, and the ratio of those taking part in online tests was significant. Discrimination performance improved in relation to the number of accesses. The estimation performance was determined for two disparate courses in order to detect incomplete participants.**

***Index Terms*—Blended learning, Incomplete participants, Access log, Discrimination.**

I. INTRODUCTION

It is often suggested that many students do not complete contemporary online learning courses, and that a learning support system is required to avoid this occurrence and to promote effective learning using online materials. One possible way to reduce incomplete participants is to find out who those students are in advance during the learning process [1]. To maximize the effectiveness of this support, incomplete participants' symptoms should be accurately detected in advance. A previous study suggested that passing a number of progress tests affects a student's possibility of completing a course [2]. A progress test may depend on course content; this should be examined carefully.

If there was a procedure to predict which students will not complete a course or to estimate participants' learning situation, it would assist in the efficient management of online learning. Towards this purpose, Ueno [3, 4] has developed a procedure using a computational model, and implemented it with an LMS (Learning Management System) for assisting online learning. This computational model, based on a decision tree theory, can accurately estimate the number of incomplete participants using a database of many student's study records. This procedure requires the preparation of a database in advance. Therefore, it can not be applied to newly developed courses or to traditional courses which do not have a database. Most online courses are developed by their teaching staff through trial and error, and the contents are frequently modified. In those types of online learning courses, it is not easy to create a database of participants' learning patterns as they progress through the course contents. For those students, the ability to predict incomplete participants is preferred over estimating the accuracy of the detecting them.

The aim of this paper is to examine the feasibility of identifying incomplete participants using current learning behavioral data such as access log data, but without a systematic database or any other information in advance. For this purpose, the first step is to compare the characteristics of access log data of complete and incomplete participants across the learning process. The second step is to investigate the possibility of predicting instances of incomplete participants, using the characteristics of the access log.

II. METHOD

A. Blended learning

To extract characteristics of incomplete participants, a set of access log data for an ordinary university course, conducted using blended-learning, was analyzed. Blended learning is defined as learning based on face-to-face classes and online learning session outside of the class.

Course: Trends in Information Industry and Society.
Participants: 81 Freshman undergraduates.
Term: Autumn, 2005.

One session of the online module was designed to correspond with each face-to-face class session, and these were held every week. Students can access the online module at any time, and are informed that the frequency of accessing the online module will be considered in their final assessment. The online module was originally developed for courses that are completely online, so that it is possible to complete all of the course using this online module.

The following types of logged data were collected and analyzed.

1) The number of accesses of the online course per week
2) The cumulative standardized number of accesses of the online course per week
3) The moving average of the number of accesses over a three week period
4) The ratio of those taking part in online tests of the weekly modules
5) Standard deviation of the number of accesses
6) The difference in the number of accesses between two weeks
7) Days elapsed from course start to first instance of access to the online module

The logged data consists of 17 weeks in total; 15 weeks of classes and a 2 week break. This permits the weekly statistics to be used as possible indices of learning activity. The course lasted for 17 weeks, and there were a number of unique events during the course, such as a guest lecture day, and a quiz day. These events affected the numbers accessing the online modules, and therefore the data varied dramatically from week to week. To compensate for this varying distribution, the number of accesses was standardized using standard deviation, and the moving average over three weeks was also extracted to obtain a wider indication of the weekly changes. The cumulative number of accesses, the moving average of the accesses and other statistics were also defined and calculated.

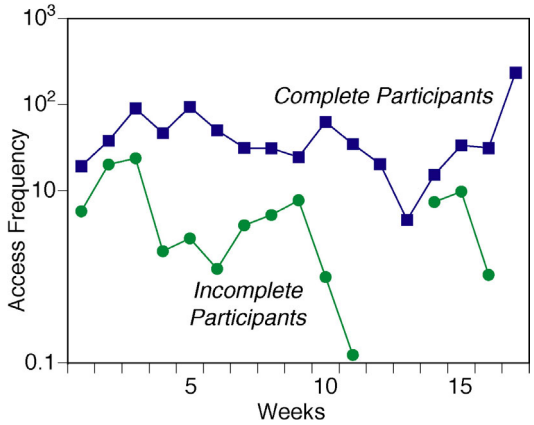

Figure 1. The number of accesses of online modules

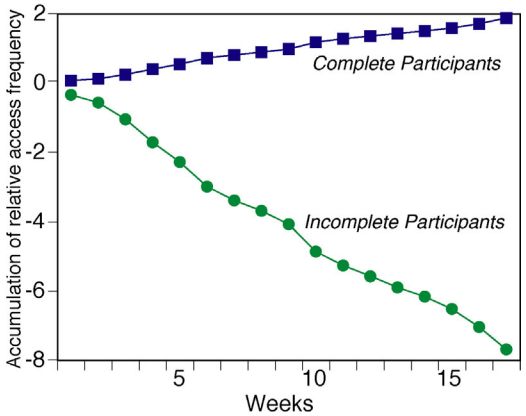

Figure 2. The cumulative standardized number of accesses.

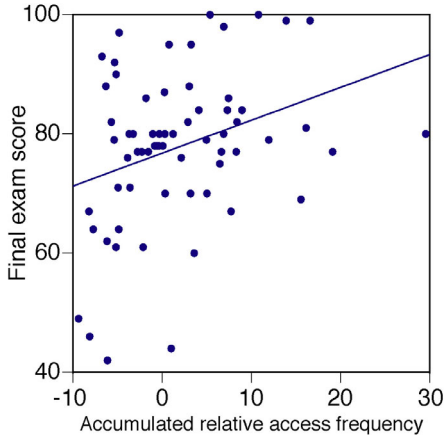

Figure 3. The relationship between final examination scores and the cumulative standardized number of accesses at the 4th week.

B. Results of statistics for accessing online modules

As a result, 65 students completed this course and 16 students did not. Most incomplete participants did not attend the final exam, as they may have considered abandoning their learning program before the exam. Therefore, 81 participants can be divided into two groups, complete and incomplete.

The number of accesses to online modules are compared between the two groups, and the results are illustrated in Figure 1. The horizontal axis shows weeks 1 to 17; the vertical axis shows the number of accesses, in logarithm scale due to the large number. For students who completed the course, the number was relatively large during its early stage, and the number also increased rapidly towards the end of the course. Participants reviewed all content before the final exam. On the other hand, the number of incomplete participants decreased gradually as the course progressed. In examining the differences between the two groups, there are significant differences ($p<0.05$) for most weeks. This suggests that there are behavioral characteristics for those accessing online materials, and that there is a possibility of detecting incomplete participants during the process using this index. The difference in the number of times, the online module was accessed increased week by week because incomplete participants had gradually stopped learning over time.

The cumulative standardized number of accesses is illustrated in Figure 2. In this figure, the cumulative number of complete participants increased monotonically along with their progress. Symmetrically, the cumulative number for incomplete participants decreased. The numbers of these students was relatively small, so that standardized numbers were negative values. Therefore the cumulative number decreased dramatically, and the difference increased week by week. In examining the difference using a t-test, there were significant differences in the cumulative number of accesses between the two groups after the fourth week (for example, $t(66)=4.86$, $p<0.01$ at 4th week).

According to the results in Figure 2, the number of accesses for incomplete participants decreased after the 4th week. This indicates that the symptom will appear around the 4th week of the course. This also suggests that some differences in learning performance may appear amongst participants who complete the course in accordance with the number of accesses to the online modules.

The relationship between the final examination score and the cumulative standardized number of accesses at the 4th week is illustrated in Figure 3. Here, incomplete participants did not take the exam, so that the scatter-gram is shown without them. According to a regression line, there is a positive correlation between final exam scores and the cumulative number of accesses, and the correlation coefficient is significant ($r=0.27$, $p<0.05$). Furthermore, correlation coefficients between final examination scores and the cumulative standardized number of accesses for every week are significant from the 2nd week to the 17th week ($r=0.26$-0.32,$p<0.05$). This provides evidence that the number of accesses can be an index of the learning situation, in particular in the data after the 4th week. Here, the number of accesses is one of the indices indicating learning behavior; those indices may be useful in estimating complete and incomplete participants.

III. ESTIMATION FOR INCOMPLETE PARTICIPANTS

According to the analysis in the previous section, there is a significant difference in the number of accesses between complete and incomplete participants. The possibility of predicting incomplete participants using information from their behavioral data is interesting. In this section, the possibility of using various information from the fourth week, by trial and error, is discussed.

A task to predict incomplete participants can be defined as a two-class discrimination of complete and incomplete participants using their behavioral data. Discrimination analysis was conducted in a stepwise fashion on all 7 of the variables, using regression with variable selection. In the subsequent steps, one or more variables were removed or reintroduced. As a result, only the ratio of those taking part in online tests was significant under discrimination analysis. This result supports the previous report [2].

According to the results of discrimination analysis, a function equation with the ratio of those taking part in online tests (x_1) can be created for discriminate participants as follows: complete participants ($y > 0$) and incomplete participates ($y < 0$).

$$y = 7.32x_1 - 3.52 \quad (1)$$

This suggests that the threshold of discrimination is 50% for the ratio of those taking part in online tests. A discrimination calculation using this function is shown in Table 1, with vertical cells showing final results and horizontal cells showing estimated results. Table 1 shows that a function can predict 13 out of 16 incomplete participants using the data from the fourth week. The relationship between estimated and final results using a chi-square test was significant (chi-square =27.4, df=1, $p<0.01$). The results determine that a linear function of the ratio of those taking part in online tests can detect incomplete participants significantly, but there are some inconsistencies between the course's final results and the estimated results, which are errors of discrimination. In Table 1, there are inconsistencies for 13 out of 81 students, and the error rate is 0.16.

In estimating the performance of incomplete participants, the precision rate (incomplete participants / total of estimated incomplete participants) is 0.57, the recall rate (estimated incomplete participants / total of incomplete participants) is 0.81. Again, the precision rate shows the exact rate of incomplete participants per the estimated rate, and the recall rate shows the correct rate of estimation for distinct incomplete participants. Therefore, the value of the precision rate is not sufficient, and the value of the recall rate is high.

TABLE I.
ESTIMATION RESULT OF COMPLETERS AND INCOMPLETERS WITH THE RATIO OF THOSE TAKING PART IN ONLINE TESTS

	Estimation		
Final result	**Complete**	**Incomplete**	**Total**
Complete	55	10	65
Incomplete	3	13	16
Total	58	23	81

If a detailed analysis of inconsistency is made, the influence of learning support may, in a sense, be small during the early stage if incomplete participants can complete the course. When the precision rate for detecting incomplete participants was low, learning support should be provided to participants who are likely to complete the course. This may create large increase in the workloads of support providers. A serious problem is that the exact number of participants who are estimated to complete the course but do not complete it can not be reduced to zero for certain. This sort of error rate requires careful consideration. The error rate was 0.04 for Table 1.

Various procedures for the improvement of estimation performance may be possible. One of them is to consider the levels of the ratio of participants taking part in online tests. Because the discrimination model consists of one variable. Here, the discriminations are considered by three levels, as follows:

- the rate is 0
- the rate is less than 50%
- the rate is more than 50%

All students who did not take part in online tests were categorized as incomplete participants. The error rate was 0.24. For participants in the second group, a discrimination function was created using the ratio of participants taking part in online tests (x_1) and the accumulated standardized number of accesses (x_2). The function is as follows:

$$y = -4.14x_2 - 2.26 \quad (2)$$

For participants of the third group, a separate discrimination function of the ratio of participants taking part in online tests (x_1), the accumulate standardized number of access (x_2) and the number of days taken to access the online module after the commencement of face to face classes (x_3) was created using the same procedure. The function is as follows:

$$y = 0.21x_2 - 0.25x_3 + 1.2 \quad (3)$$

The revised results using multiple discrimination functions for three groups are shown in Table 2. According to the table, miss-classification of incomplete participants as complete participants can be 0, however many participants completing the course were classified as not completing it. There is a tradeoff relationship between the two groups. Additionally, these two variables may also show some characteristics of students' accesses although they are not significant in the initial analysis.

TABLE II.
ESTIMATION RESULT OF COMPLETERS AND INCOMPLETERS WITH THREE INDICES AT THE FOURTH WEEK: RATIO OF THOSE TAKING PART IN ONLINE TESTS, THE ACCUMULATE STANDARDIZED NUMBER OF ACCESS AND THE NUMBER OF DAYS COSTING TO ACCESS THE ONLINE MODULE AFTER THE FACE TO FACE CLASS.

	Estimation		
Final result	**Complete**	**Incomplete**	**Total**
Complete	40	25	65
Incomplete	0	16	16
Total	40	41	81

IV. APPLYING THE PROCEDURE TO OTHER ONLINE COURSES.

A discrimination procedure for incomplete participants was developed using student's behavioral data for one lecture. As this was based on a case study of a single course, the possibility of making generalizations is not certain. To determine the validity of this procedure, it was applied to participants of two other courses which use online modules, as follows:

Course A: a complete online course for 15 Bachelor students

Course B: a blended learning course for 68 Graduate students

These courses were essentially different, but the topics were similar in content to the class which was analyzed for developing the procedure.

A. Estimation results for class A:

All variables were created using students' behavioral data, and the prediction results are summarized in Table 3, using the same format. According to these results, the procedure can estimate that 4 out of 6 Bachelor students will not complete the course. The miss-classification of incomplete participants as participant who completed the course was 0.13 (2/15), however. The relationship between final results and estimated results was significant using a chi-square test (chi-square=8.2, df=1, $p<0.01$). This supports the idea that the estimation procedure can predict incomplete participants and it is significant, and that the prediction of incomplete participants has to be carefully considered. Also, this suggests the possibility of applying this procedure to fully online courses.

TABLE III.
ESTIMATION RESULT FOR CLASS A USING THE DEVELOPED PROCEDURE AT THE FOURTH WEEK

	Estimation		
Final result	**Complete**	**Incomplete**	**Total**
Complete	9	0	9
Incomplete	2	4	6
Total	11	4	15

B. Estimation results for class B:

The results for class B, where subjects were Graduate school students, were examined using the same procedure. Their performance was summarized in Table 4, using the same format. The relationship between final results and estimated results was not significant (chi-square=2.89, df=1, n.s.). The miss-classification of incomplete participants as participants who complete the course was 0.03 (2/68). This was relatively small, however, and suggests that one must consider the type of class when applying the estimation procedure that has been developed.

TABLE IV.
ESTIMATION RESULT FOR CLASS B USING THE DEVELOPED PROCEDURE AT THE FOURTH WEEK

	Estimation		
Final result	**Complete**	**Incomplete**	**Total**
Complete	33	26	59
Incomplete	2	7	9
Total	35	33	68

Comparing the performance of incomplete participant estimation between classes A and B, the performance was significant for Bachelor students who are in the same category of students the procedure was developed for, but the performance was not significant for Masters students. There was a difference in the learning style between class A, which was a fully online course, and the class used for the development of the procedure, a blended learning course. Although performance was not significant, the learning style for class B and for the class used for the development of the procedure were both blended learning. This suggests that the estimation performance for incomplete participants may depend on the type of group, such as Bachelors students or Masters students. Nakayama et al. [5] have reported that there are some differences in the types of characteristics of Bachelors and Masters students for blended learning. In particular, the learning strategy is significantly different. This result may be related to the differences in the characteristics of students who are in Bachelor degree courses or Master's degree courses.

V. CONCLUSION

This paper examines the feasibility of identifying incomplete participants who do not complete courses in the early stage of a blended learning environment course, using current learning behavioral data. The estimation procedure was developed using a simple discrimination model and several behavioral data, such as access log data for online modules, the ratios of those taking part in online tests, etc. There is a significant difference in the cumulative standardized number of accesses between participants who complete and do not complete the course. According to the results of discrimination analysis, an estimation procedure for predicting which students will fail to complete a course was developed, and the performance was evaluated. The possibility of applying the procedure to other classes was also examined. The performance was significant for Bachelor students who are in the same category as the students the procedure was developed for, but for Masters students the performance was not significant. This suggests that the estimation performance for incomplete participants may depend on its group, such as Bachelor students or Masters students. This result supports the idea that estimation procedures can predict incomplete participants, and prediction of this requires careful consideration.

The development of a more robust estimation procedure and the encouragement of the development of supporting methodologies for monitoring and assisting incomplete participants will be subjects for our further study.

ACKNOWLEDGMENT

This research is partially supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (B-19300274), 2007-2009.

REFERENCES

[1] T. Matsuda, N. Honna and H. Kato, "Development of e-Mentoring Guideline and its Evaluation", *Japan Journal of Educational Technology*, vol. 29, pp.239-250, 2004.

[2] C. Kougo and E. Nojima, "Student dropout in e-learning and its symptom", *Proc. of JSET annual conference*, vol. 20, pp.997-998, 2004.

[3] M. Ueno, "Online Outlier Detection for e-Learning Time Data", *The IEICE Trans. on Information and System*, vol. J90-D, pp.40-51, 2007.

[4] M. Ueno, "Data Mining in e-Learning", *Japan Journal of Educational Technology*, vol. 31, pp.271-284, 2007..

[5] M. Nakayama, H. Yamamoto and R. Santiago, "The Impact of Learner Characteristics and Learning Performance in Hybrid Courses Among Japanese Students", *The Electronic Journal of e-Learning*, vol.5, issue 3, pp.195-206, 2007.

AUTHORS

M. Nakayama is with CRADLE (the Center for Research and Development of Educational Technology), Tokyo Institute of Technology, Tokyo, 152-8552 Japan (e-mail: nakayama@cradle.titech.ac.jp). Correspondent author.

H. Kanazawa was with Tokyo Institute of Technology, Tokyo, 152-8552 Japan. He is now with Uchida Yoko Co., Ltd.

H. Yamamoto is Professor Emeritus, Shinshu University, Japan. He is now with CRADLE, Tokyo Institute of Technology, 152-8552 Japan.

Manuscript received 18 September 2008. Published as submitted by the authors.