Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Joint Activity on Learner Performance Prediction using the BookRoll Dataset

Brendan FLANAGAN[a*], Weiqin CHEN[b] & Hiroaki OGATA[a]
[a] *Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b] *Oslo Metropolitian University, Norway*
*flanagan.brendanjohn.4n@kyoto-u.ac.jp

Abstract: In this paper, we introduce the BookRoll dataset that was provided for analysis in the joint activity on learning performance prediction in the Learning Analytics workshop at ICCE2018. Firstly, we provide a definition of the task, the dataset, how the data was collected, and briefly introduce previous work that has analyzed similar data from the BookRoll system. We also give an overview of the various approaches that were adopted by the participating authors for learner performance prediction using BookRoll reading behavior logs.

Keywords: Learner Performance Prediction, Learning Analytics, Reading Behavior, E-Book Reading

1. Introduction

Learning environments are becoming digitized at an ever-increasing rate, with most systems storing data about the interaction and behaviors of learners as event logs (Verbert, 2012). The analysis of such data has been gaining attention not only through the sheer volume, but also because of the potential to analyze learning progress, personalization, and support more effectively by predicting learner behavior and outcomes. As the analysis of gathered data is playing an increasing role in Learning Analytics (LA) and Educational Data Mining (EDM), a joint activity was organized to prompt the prediction of student performance by analyzing reading patterns from logs of an e-book system. Anonymized reading log data was provided to participants before the workshop to create models that predict a learner's final score for a course.

Digital textbooks and e-books are being introduced into education at the government level in a number of countries in Asia (Ogata, 2015). This has prompted research into not only the use of such materials within the classroom, but also the collection and analysis of event data collected from the systems that are used for support and distribution. The data that was provided for the joint activity was generated using a digital learning material reading system called BookRoll (Ogata, 2015, 2017) which will be introduced in more detail in the following section.

1.1 BookRoll: Digital Learning Material Reading System

Digitized learning materials play a core role in modern formal education. They server not only as a learning material distribution platform, but also are an important source of data for learning analytics research into the reading behavior of students. As the materials are read by students using the system, the action events are recorded, such as: flipping to the next or previous page, jumping to different pages, memos, comments, bookmarks, and drawing markers to indicate parts of the learning materials that learners think are important or find difficult. Previous research into the reading behavior of students has been used in review patterns, visualizing class preparation, and investigate the self-regulation of learners (Yin et al., 2015; Ogata et al., 2017; Yamada et al., 2017). The analysis of reading behavior can be used to inform the revision of learning materials based on previous use, predict at-risk students that may require intervention from a teacher, and identify learning strategies that are less effective and provide scaffolding to inform and encourage more

effective strategies. The digital learning material reader can be used to not only log the actions of students reading reference materials, but also to distribute lecture slides.

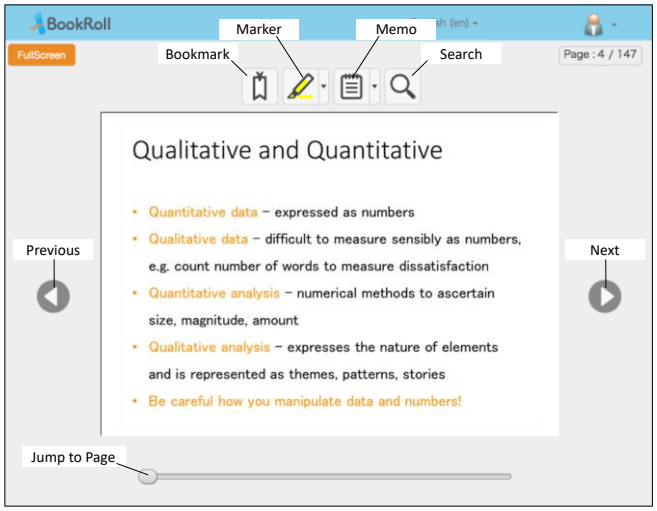

Figure 1. BookRoll digital learning material reader user interface.

As shown in Figure 1, the user interface supports a variety of functions, such as: moving to the next or previous page, jumping to an arbitrary page, marking sections of reading materials in yellow to indicate sections that were not understood, or red for important sections. Memos can also be created at the page level or with a marker to attach it to a specific section of the page. Users can also bookmark pages or use the full-text search function to find the information they are looking for later when revising. Currently, learning material content can be uploaded to BookRoll in PDF format, and it supports a wide range of devices, including: notebook computers, tablets, and smartphones, as it can be accessed through a standard web browser. When used in a standalone environment, the user behavior from BookRoll is logged in a local database and requires that analysis is performed by either connecting directly, or exporting data from the database.

2. BookRoll Dataset

2.1 Data Collection

The two datasets provided for analysis in the joint activity were collected using an LMS independent LA platform developed at Kyoto University as part of an ongoing project to establish fundamental LA infrastructure (Flanagan & Ogata, 2017, 2018). The platform enables the collection and analysis of data from behavior sensors, such as: BookRoll, LMS, and other learning systems with which learners directly interact. The user behavior events are sent by a xAPI interface and collected in a central independent Learning Record Store (LRS). The platform is designed to minimize the recording of personal information as much as possible by recording logs with identifiers that do not contain such data. The event logs were extracted from the LRS and transformed into CSV for ease of use and distribution to various researchers.

The BookRoll dataset released for the joint activity was collected from two different courses: a course that consisted of one 3-hour intensive lecture (dataset 1), and a course that spanned three 90-minute lectures (dataset 2). For each of these courses, there are two main types of files included in the dataset: the BookRoll clickstream log data and the final score of all students which is the target of the prediction task. The definition of the dataset columns for scores and clickstreams are shown in Table 1 and 2 respectively, with the details of the *operationname* column shown in Table 3.

Table 1

Details of the Score Data

Column	Description	Example
userid	Anonymized student id	"ds1001"
score	The final score that the student received for the course: this is the prediction target value	"80"

Table 2

Details of the Clickstream Data

Column	Description	Example
userid	Anonymized student id	"ds1001"
action	xAPI verb for the action that the student performed	"https://w3id.org/xapi/adb/verbs/read"
operationname	The BookRoll action that was performed by the student	*Please refer to Table 3 for details*
processcode	A grouping of actions by event type	*"23"*
devicecode	type of device used to view BookRoll	*"pc", "mobile", "tablet"*
contentsid	the id of the learning material that is being read	*"eccbc87e4b5ce2fe28308fd9f2a7baf3"*
markerposition	the position (x,y,w,h,onscreen w,onscreen h) of the marker added to a page	"367,34,28,20,714,504"
markercolor	color of the marker added to a page	important: "rgb(255,0,0)" not understood: "rgb(255,255,0)"
markertext	the text contained on the page where the marker was drawn	"Introduction to Elementary Informatics"
memotext	A comment or memo written by a student	"The concept of information entropy was introduced by Claude Shannon"
description	When operationname = {PAGE_JUMP | SEARCH_JUMP}: the page the user moved to by the jump	"5"
pageno	the current page where the action was performed	"2"
eventtime	A UTC timestamp of when the event occurred	"2017/05/19 4:02:24"

Table 3

BookRoll Operation Name Details

Value	Description
OPEN	Learning material was opened
CLOSE	Learning material was closed
NEXT	Next page button was clicked
PREV	Pervious page button was clicked
PAGE_JUMP	Jumped to a particular page
ADD BOOKMARK	Added a bookmark to current page
ADD MARKER	Added a marker to current page
ADD MEMO	Added a memo to current page
CHANGE MEMO	Edited an existing memo
DELETE BOOKMARK	Deleted a bookmark on current page

DELETE MARKER	Deleted a marker on current page
DELETE_MEMO	Deleted a memo on current page
LINK_CLICK	Clicked a link contained in the e-book current page
SEARCH	Searched for something within the e-book
SEARCH_JUMP	Jumped to a page from the search results

2.2 Data Characteristics

The basic characteristics of the provided datasets are described in this section. The sample size of each dataset can be seen in Table 4, and a graph of the differences in the distribution of scores that were used as the prediction target value is shown in Figure 2. It should be noted that the passing score for both courses was a score of 60 points or higher. As can be seen for the kernel density estimation for the scores of both datasets, approaching the prediction task as a binary classification with a cutoff of 60 points would result in pass biased prediction as the distribution is skewed toward high passing scores. Finally, each of the datasets contained three different learning materials that are identified by the contentsid field.

Table 4

Number of Samples in Both Datasets

Dataset	Number of students	Total Event Logs
1	53	28,827
2	55	36,930

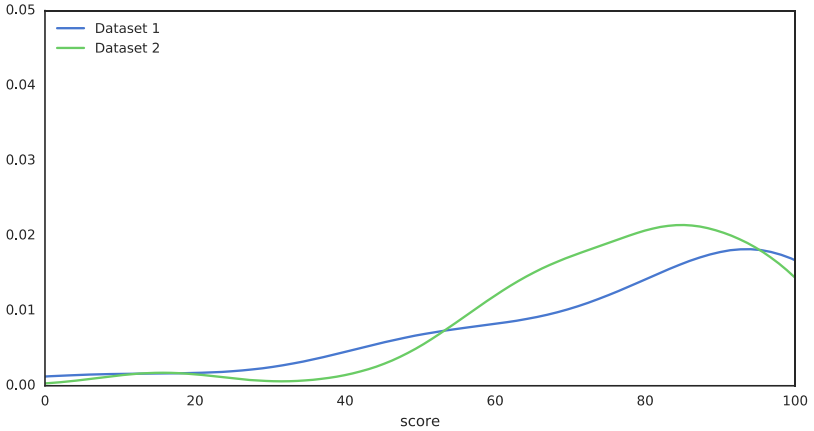

Figure 2. Histogram and kernel density estimation of the distribution of scores for Dataset 1 and 2.

3. Approaches

3.1 Preprocessing

As mentioned in the previous section, the final grade score was the target value of prediction and both datasets are skewed towards high passing scores. Some of the participants who approached the task as binary classification tried to address data imbalance by resampling the datasets. Dataset balancing was performed by two groups: resampling to was applied by Lu et al. (2017), and Hasnine et al. (2018) used the SMOTE algorithm for over-sampling. Askinadze et al. (2018), opted to divide the dataset into balanced groups by changing the cutoff threshold for passing scores. Most participants created new or aggregated features by analyzing the event data, with reading time being highlighted as an important feature for prediction by several groups. Hirokawa (2018), applied preprocessing to extract features that represent the transition sequence behavior of reading by

learners, and attributed this to a 25% increase in prediction accuracy when compared to only page access features. After observing that the features were hard to separate, Askinadze et al. (2018) transformed the features by representing them in the k-Means cluster-distance space.

3.2 Prediction Methods

As the task was mainly approached as a binary classification and/or regression task, many different machine learning techniques were employed for prediction. Some notable methods and the groups that utilized them are shown in Table 5, with Decision Tree (DT), Gradient Boosting (GB), k-nearest neighbors (kNN), Logistic/Linear Regression (LR), Naïve Bayes (NB), Neural Networks (NN), Random Forest (RF), Support Vector Machine/Multiple Linear Regression (SVM/MLR). SVM/MLR proved to be one of the most popular classification/regression models.

Table 5

Popular Prediction Methods used by Participants

Participant	**DT**	**GB**	**kNN**	**LR**	**NB**	**NN**	**RF**	**SVM/ MLR**
Askinadze et al.			•				•	•
Goh & Lo	•							•
Hasnine et al.			•	•	•	•	•	•
Hirokawa								•
Huang et al.	•	•		•	•	•	•	•
Kikuchi & Tezuka		•				•	•	
Lu et al.	•	•				•		•
Total	3	3	2	2	2	4	4	**6**

It was recommended that participants evaluate their predictions using Area Under the Curve (AUC) for binary classification, and Root-Mean-Square-Error (RMSE) for regression. However, as the format of the joint activity was not a formal data challenge, participants had the freedom to choose other evaluation techniques. Therefore, it is not possible to directly compare the results from all of the participants in the joint task.

3.3 Alternative Approaches

Some participants took approaches that did not involve the use of machine learning techniques to predict the final grade scores. Ono (2018), investigated focusing on the theories of reading comprehension, and analyzed the page-flipping history of learner using a small sub-sample of 10 students' data.

4. Conclusions

In this paper, we describe the datasets that were provided for the joint activity on learner performance prediction in the Learning Analytics Workshop at ICCE2018. A total of 8 research contributions were submitted and a range of various approaches to the problem of predicting student scores from reading log data were proposed. Several key problems were: dealing with imbalance and

bias datasets, the engineering and selection of effective features that represent the reading behavior of students, and the transformation or augmentation of data to improve the accuracy of prediction. As some participants mentioned, the size of the dataset shared in the joint task could have had an impact on the accuracy and range of analysis that could be applied to the task. Therefore in future work, we should examine the collection of larger datasets to encourage new approaches, or investigate methods that use transfer learning or generalizing across multiple datasets from different courses.

Acknowledgements

This research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304.

References

Askinadze, A., Liebeck, M., & Conrad, S. (2018). Predicting Student Test Performance based on Time Series Data of eBook Reader Behavior Using the Cluster-Distance Space Transformation. In *International Conference on Computers in Education (ICCE2018): Learning Analytics Workshop Joint Activity*.

Flanagan, B., & Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. In *International Conference on Computers in Education (ICCE2017)* (pp. 333-338).

Flanagan, B., & Ogata, H. (2018). Learning Analytics Platform in Higher Education in Japan, *Knowledge Management & E-Learning (KM&EL)* (in press)

Goh, K.S., & Ho, L.C. (2018). Predicting student performance using E-book clickstream data. In *International Conference on Computers in Education (ICCE2018): Learning Analytics Workshop Joint Activity*.

Hasnine, M.N., Akçapınar, G., Flanagan, B., Majumdar, R., Mouri, K., & Ogata, H. (2018). Towards Final Scores Prediction over Clickstream Using Machine Learning Methods. In *International Conference on Computers in Education (ICCE2018): Learning Analytics Workshop Joint Activity*.

Hirokawa, S. (2018). Good Students Look Back Previous Pages. In *International Conference on Computers in Education (ICCE2018): Learning Analytics Workshop Joint Activity*.

Huang, A.Y.Q., Weng, J.X., Huang, J.C.H., Lu, O.H.T., Jong, B.S., & Yang, S.J.H. (2018). Prediction of Students' Academic Performance based on Tracking logs. In *International Conference on Computers in Education (ICCE2018): Learning Analytics Workshop Joint Activity*.

Kikuchi, S., & Tezuka, T. (2018). Feature analysis for predicting students' performance from reading patterns in an e-learning system. In *International Conference on Computers in Education (ICCE2018): Learning Analytics Workshop Joint Activity*.

Lu, O.H.T., Huang, A.Y.Q., & Yang, S.J.H. (2018). Benchmarking and Tuning Regression Algorithms on Predicting Students' Academic Performance. In *International Conference on Computers in Education (ICCE2018): Learning Analytics Workshop Joint Activity*.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In *International Conference on Computer in Education (ICCE 2015)* (pp. 401-406).

Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., Wang, J., Hirokawa, S. (2017). Learning Analytics for E-Book-Based Educational Big Data in Higher Education. In *Smart Sensors at the IoT Frontier* (pp. 327-350). Springer, Cham.

Ono, Y. (2018). Can Page-Flip Predict Better Reading Comprehension? – A Preliminary Study. In *International Conference on Computers in Education (ICCE2018): Learning Analytics Workshop Joint Activity*.

Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-Driven Research to Support Learning and Knowledge Analytics. *Educational Technology & Society*, 15 (3), 133–148.

Yamada, M., Oi, M., & Konomi, S. I. (2017). Are Learning Logs Related to Procrastination? From the Viewpoint of Self-Regulated Learning. In *International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2017)*, (pp. 3-10).

Yin, C., Okubo, F., Shimada, A., Oi, M., Hirokawa, S., Yamada, M., Kojima, K., & Ogata, H. (2015). Analyzing the Features of Learning Behaviors of Students using e-Books. In *International Conference on Computers in Education (ICCE 2015)*, (pp. 617-626).