International Journal of Computing Sciences Research (ISSN print: 2546-0552; ISSN online: 2546-115X)
Vol. 1, No. 3, pp. 1-16
doi: 10.25147/ijcsr.2017.001.1.09
https://stepacademic.net

Monitoring Class Activity and Predicting Student Performance Using Moodle Action Log Data

Rodolfo C. Raga Jr
Department of Computer Studies and Engineering,
Jose Rizal University
Mandaluyong City, Philippines
rodolfo.raga@jru.edu
(corresponding author)

Jennifer D. Raga
Management and Information Systems Department
Western Group of Companies,
Quezon City, Philippines
jenny.raga@western.com.ph

Date received: October 8, 2017
Date received in revised form: November 14, 2017
Date accepted: November 19, 2017

Recommended citation:

Raga, R. C. & Raga, J. D. (2017). Monitoring class activity and predicting student performance using Moodle action log data. *International Journal of Computing Sciences Research*, *1*(3), 1-16. doi: 10.25147/ijcsr.2017.001.1.09.*

**This paper is presented at the 2017 1st International Conference on Redesigning, Re-engineering Academic Direction for Global Competitiveness.*

Abstract

Purpose – This paper proposes a novel approach for processing course log data obtained from Moodle-based blended courses in order to visualize patterns of student activity within the online environment and to determine whether these log data can be used to predict student academic performance.

Method – Logs of student activities were summarized and processed using the Vector Space Model approach. This resulted in a novel vector-based form of representation which can be used to map students' activity in a latent activity space given a set of activity dimensions. An enriched form of this representation was also generated by processing the DateTime and IP address metadata for the purpose of developing classification/predictive model of students' performance.

Results – The activity space coupled with a one-hot vector representation for each unique activity dimension can be used to visualize the differences in level and type of activity preferences of students. Experiments using several machine learning algorithms indicate that the generated model can modestly distinguish between sets of activities that lead to High, Low, or Failed performances.

Conclusion – The development of easily interpretable graphics that can depict trends in student activity is a useful tool for instructors handling blended courses. It can provide constant monitoring of course progression with minimal effort and enable instructors to determine whether and how the environment actually affects student performance.

Recommendations – Further work on refining the process applied to the data is recommended. The log data should be time-sliced and processed to determine whether and how the student's level and type of activity changes over time. More powerful machine learning classification techniques shouldalso be tested to determine whether it can improve the classification accuracy.

Research Implications – These types of visualizations and predictive models could be used to monitor the student or class which requires immediate and specific pedagogical adjustments.

Keywords – action log analysis, blended learning, Moodle, performance prediction, student activity

---

INTRODUCTION

Blended Learning (BL) has become popular in the last few years, spurred by the widespread use of the web and the opportunities and conveniences that it provides (Norberg, Dziuban, & Moskal, 2011). Wikipedia defines BL as "an education program (formal or non-formal) that combines online digital media with traditional classroom methods. It requires the physical presence of both teacher and student, with some element of student control over time, place, path, or pace" (Wikipedia, 2017). By combining face-to-face with online learning techniques, BL is virtually considered as being capable of accommodating the various learning strategies and styles of students. This perceived benefit convinced many Higher Education Institutions (HEIs) in the Philippines to start adopting blended learning strategy in their curriculum. By providing students with the ability to independently access learning resources "anytime, anywhere", BL is assumed to create an advantageous environment that can remove many barriers to enhancing student performance, while promoting high-quality interactions between faculty and students (Nicdao, 2013; Kanuka, Brooks, & Saranchuck, 2009).

Learning Management Systems (LMS) is a key feature of blended learning (Dias & Diniz, 2014). Specifically, in the Philippines, the Modular Object-Oriented Developmental Learning Environment or Moodle LMS (Rice, 2006) is commonly used to support the blended learning setup not only because it is cost effective but also because it provides sufficient features to enable HEIs to create flexible online learning environments. These environments not only allow students convenient access to educational resources, it also provides HEIs the opportunity to collect vast quantities of data on students' activities. These data offer rich potentials in studying student behavior and can also help determine whether there are patterns that lead to better success in learning. An example of this is the action logs recorded in Moodle. This log maintains six data dimensions describing how students interact with the online environment (see Table 1).

Taking advantage of this data, however, is neither simple nor straightforward due to its massive volume and high rate of velocity. Most often than not, assistance from specialized tools is needed to extract useful information for tracking and assessing the activities performed by students, especially in cases where the faculty member is handling multiple classes. At the same time, although the Moodle system provides some reporting tools, it does not provide specific features which can enable educators to directly monitor and evaluate the activities of students in relation to the structure and contents of the course and how it affects the learning process (Zorrilla, Millan, & Menasalvas, 2005). Based on experience, this keeps instructors from making meaningful sense and use of this data (Estacio & Raga, 2017).

Table 1. Dimensions of Action Logs

Data dimension	Description
Course	Identification string of the course in which the action is related
Time	Date and time stamp of when action was executed
IP Address	Unique numerical label assigned to the device used by the user
User Full Name	The user who initiated the action
Action	Type of action initiated
Information	General information on learning activities

This paper describes a pilot study that discusses and illustrates the use of a novel approach in analyzing the log data generated by Moodle in a blended learning context. The proposed technique can be used to process and break down the multidimensional log data collected by the LMS in order to generate graphical representations that provide a profile of students' activities online, both individually and within a group. This can help to reveal some of the students' adopted self-regulatory learning strategies as reflected by the differences in the way they utilize features of the LMS. The study also attempts to determine whether the same data can be used to predict student course performance through the use of various data mining techniques. Experiments were conducted comparing several classification algorithms using a feature-enhanced version of the same data used in the previous task. The aim of this experiment will be to determine if there is an overall ideal set of data attributes that can be used to predict and anticipate student

academic performance in blended courses and which algorithm is best suited to process these attributes. The study is being conducted in support of a Course Redesign Program (CRP) of a University whose goal is to redesign instructional approaches by integrating e-learning technology into traditional classes to achieve quality enhancements.

THE PROPOSED APPROACH

The proposed approach for addressing the issues cited above combines techniques borrowed from the fields of Information Retrieval (IR) and Data Mining (DM). In particular, the concept of Vector Space Model (VSM) and Classification techniques are discussed in succeeding sections.

IR and Vector Space Model

VSM is a statistical model of representation often used in the field of Information Retrievalfor processing text documents (Singhal, 2001). The main idea behind VSM is to construct a vector of terms representation for documents and use these to compare the contents of documents in a latent semantic space. Recently, there has been some progress on utilizing VSM for purposes outside the field of IR. Sreeja and Mahalakshmi (2016), for example, explored the use of VSM to automatically detect emotions in English poems. Fraser and Hirst (2016) investigated using VSM to detect language impairments among people with Alzheimer's disease, and Younge and Kuhn (2016) used VSM as a measure to detect patent similarity. Salehi, Pourzaferani, and Razavi (2013), in an attempt to provide students with a tool that can be used to cope up with the ever-increasing numbers of learning materials in the web, also developed a hybrid recommender system that locates suitable learning materials and delivers them to learners based on their specific attributes. In the same manner, in this study, VSM is applied to activity data generated within the blended learning courses to determine whether it can enable instructors to overcome the voluminous amount of data and be able to use these as a guide in providing formative feedbacks and/or in adjusting pedagogical strategies.

In traditional VSM, if terms are represented using words, then every word in the document is treated as an independent dimension in the vector representation and the value assigned to each dimension is the number of occurrence of each unique word in the document. Using this approach, any document can be represented by a vector, and thereafter, plotted and compared in a multi-dimensional semantic space. To compute document similarity in this space, the angle produced between their representative vectors can be measured using the cosine distance formula (see Equation 1). This returns a value between zero and one. The higher the value, the more similar the documents are assumed.

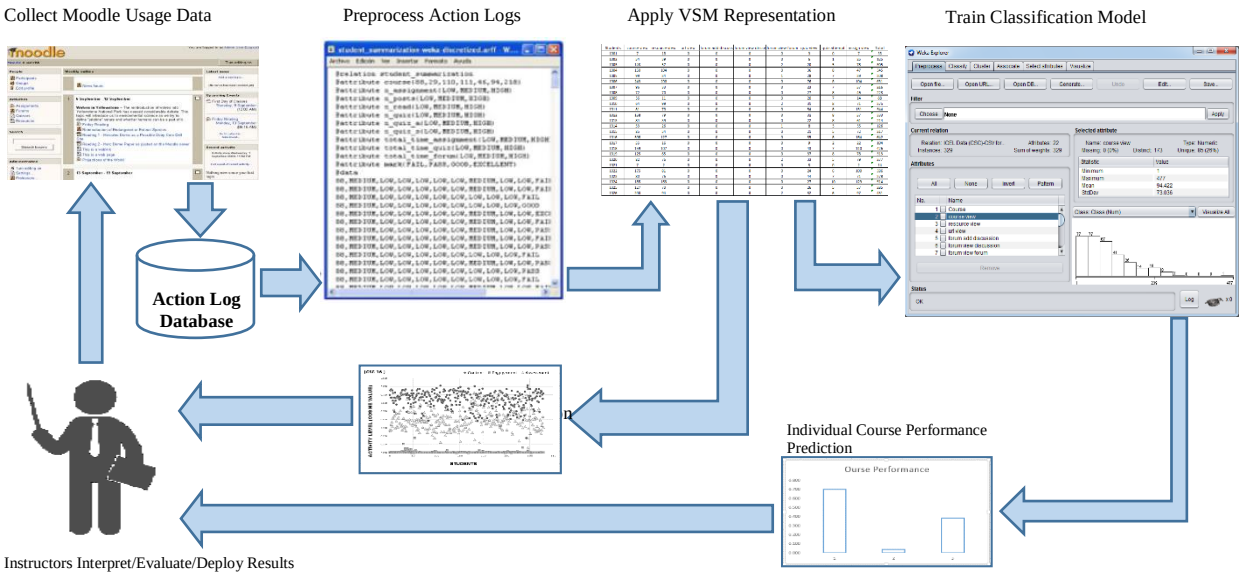


Figure 1. Data Processing Model

$$s(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum_{i=0}^{n-1} x_i y_i}{\sqrt{\sum_{i=0}^{n-1}(x_i)^2} \times \sqrt{\sum_{i=0}^{n-1}(y_i)^2}}$$

Equation 1

Classification Techniques

Classification is a data mining task used to analyze the attributes of data items in a collection with the end goal of assigning individual items to target categories or classes (Ahmed & Elaraby, 2014). The classification process involves two phases, the initial learning phase and the subsequent classification phase (Baradwaj & Pal, 2011).During the learning phase, a selected set of training data with corresponding class labels are first analyzed by a classification algorithm to develop a classification model. In the succeeding classification phase, sample test data (without their class labels) are inputted to the generated classification model to determine how well the model can predict the target categories or class labels of each item. The percentage of correct class labels outputted by the model is then used to estimate the accuracy of the model. If the accuracy is acceptable, the model is deemed fit to be applied to new and real-time data sets.

Applying classification techniques to predicting student performance is a challenging task. An initial key problem that needs to be addressed is identifying the most suitable method for predicting the performance (Shahiri & Husain, 2015). Following Kabakchieva (2013), several learning algorithms were used in this experiment. These are general-purpose learning algorithms covering different paradigms. They were selected because of their availability in the Scikit-learn machine learning library of Python (Pedregosa et al., 2011) and were used with default parameter settings:

(1) Logistic regression is a predictive modeling approach used to quantify the degree of relationship between a dependent variable and one or more independent variables.
(2) Linear discriminant analysis (LDA) is a method that can be used to classify a data object into one of several classes by finding the linear combination of features that characterizes the different classes. It is closely related to regression analysis (Xanthopoulos, Pardalos, & Trafalis, 2013).
(3) kNN is an instance based algorithm used to classify a data object by applying a majority voting mechanism among its nearest neighbors in a feature space.
(4) CARTis a regression-based predictive model that generates a decision tree.
(5) Random Forestis an ensemble type of learning algorithm that can classify data objects by constructing several decision trees during training time and then using the mean prediction of the individual trees as decision output.
(6) BayesNet is an algorithm that can be used to represent probability distributions using a network of nodes.
(7) SVM is a supervised machine learning technique often used for classification. It operates by finding a maximizedhyperplane that can segregate two classes effectively.

Analysis model

The proposed analysis model consists of three stages: (1) collection and pre-processing of action logs, (2) application of VSM representation to generate activity space visualization, and (3) training of classification model and producing course performance predictions (Figure 1). The analysis process focuses first on the action dimension. Table 2 shows the initial set of action types examined in this study. These actions were selected because they represent the various activities that the students most often engaged with inside Moodle. The collected records were pre-processed by anonymizing specific student information.

Table 2. Actions types and class activity

Action Type	Corresponding Class Activity
Course View Resource View URL View	Content Access
Forum Add Disc Forum View Disc Forum View Forum	Engagement
Quiz View Quiz Attempt Assign View	Assessment

To represent class activity using VSM requires the construction of vectors that represents the activity of each student. This activity vector can be defined as simply a list of action types with their corresponding values depicting how many times each action was initiated by the student. For instance, Figure 2 provides a sample matrix depicting a set of activity vectors for five students. Here, the values in each element represent the number of times each studentperformed such action. As such, a value of zero means that the action type was not performed at all.

Student	course view	resource view	url view	forum add disc	forum view disc	forum view forum	quiz view	quiz attempt	assign view	Total
1001	7	18	0	0	0	0	3	0	7	35
1002	54	39	0	0	0	0	6	1	25	125
1003	143	57	0	0	0	2	20	5	78	305
1004	150	104	0	0	0	0	36	8	47	345
1005	99	34	0	0	0	1	28	7	39	208

Figure 2. Student Activity Vectors

Following the semantic space analogy, each activity vector can serve as a coordinate that can be used to plot students in a 3-dimensional space where each dimension is notionally assigned to each type of activity. Figure 3 illustrates this space along with student vectors plotted in it.

This representation can be used to compare students' activity with each other and/or to measure how much students implicitly prefer a certain type of activity within the environment (e.g., engagement dimension). In this paper, the latter approach is explored. Notice that in Figure 2, the action types in the activity vector are ordered based on the type of activity to which the action type belongs (e.g., the first 3 columns belong to content access, the next set belongs to forum engagement, and so on). This coding enables one-hot encoding representation for each activity dimension to be constructed. This can be done by setting the action types for a specific activity to a non-zero value (i.e., one) while the rest of the action type values are set to zero. Thus, the representative vector for each activity dimension would be as shown in Table 3.

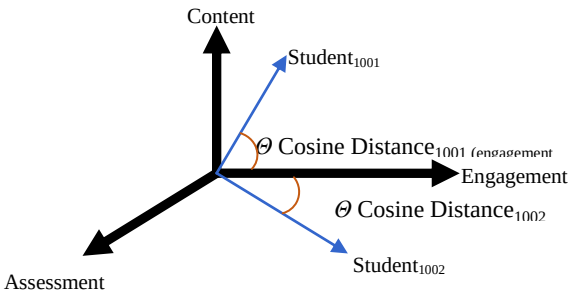


Figure 3. A 3-dimensional Activity Space

Measuring each student's level of activity relative to each dimension then simply requires applying the cosine formula between the students' activity vector and the one-hot encoding representation vector for each dimension. This process provides cosine scores for students representing the level of activity of each class for each dimension.

Table 3. Representative vectors for each activity dimension

Activity Dimensions	**Representative Vectors**								
Content	1	1	1	0	0	0	0	0	0
Engagement	0	0	0	1	1	1	0	0	0
Assessment	0	0	0	0	0	0	1	1	1

The final step applies classification techniques to analyze the data further and to test how good these algorithms can predict student course performance. For this purpose, the vector structure is first subjected to a data transformation process which aims to enrich it with additional data elements extracted from the *Time* and *IP Address* dimensions of the log data (see Table 1). The main idea for this enrichment process is to include as many attributes of the action log as possible into the vector structure, and then, later on, to test the strength of these attributes as predictors of student's academic performance.

The DateTime stamp was first processed by separating the Date and Time values. The date values were then used to compute the total_days_span (TDS) index. This was done by counting the total number of days elapsed between the first and last dates that the student logged an action in the system. Then, the total_access_days (TAD) index was measured by counting the total unique number of days that the student logged-in an action. The two index values were then combined to produce the access_density_score (ADS). The proposed formula used to define ADS is shown in Equation 2.

$$ADS = TanH(TDS/TAD) \qquad \textit{Equation 2}$$

The ADS expresses a scaled ratio of TDS and TAD. This scaled ratio is proposed to objectively rank the amount of effort exerted by each student in conducting activities within the online environment. For example, some students may incur the same number of access days, but if one student has a lower number of days spanned in using the system then the ADS will be assigning him a higher score value for his effort. ADS values are between -1 and +1, as the students exert more effort in accessing the system on a daily basis the ADS value approaches +1. Comparing the ADS of students serves to highlight the effort profile exerted by students in accessing the system.

The time metadata of all the actions incurred by the students, on the other hand, were grouped into four different categories, namely: (i) AM+, (ii) AM-, (iii) PM+, and (iv) PM-. The basis for assigning a particular time stamp in each category is provided in Table

4. This grouping serves to highlight the access time profile of students in accessing the system.

Table 4. Time metadata grouping

Category	Description	Criteria
AM+	Early morning	If timestamp is between 00:00 and 06:00
AM-	Morning	If timestamp is between 06:01 and 12:00
PM+	Afternoon	If timestamp is between 12:01 and 18:00
PM-	Evening	If timestamp is between 18:01 and 23:59

For processing the IP address stamp, another grouping based on the known Network ID (NID) of computers located inside the University was used. The NID is the leftmost numeric label in the IP address used to identify computers in a network. The designated NID of computers inside the target University is 168; therefore, any action whose NID is not 168 was initiated using devices outside the university premises. We grouped all actions between those incurred inside the university and those incurred outside the university. This grouping serves to highlight access location profile of students in accessing the system.

Finally, the final grade achieved by each student in the course was added as a final attribute. To generate a categorical class label for each student, the final grades were classified as to whether they are *High*, *Low, or Failed*. Figure 4 provides a preview of the complete and final set of data attributes used to train the classification model.

Activity Vector			Total	Time Specificity				Location Specificity		Access Specificity				Final	Class
Action 1	...	Action n	Actions	AM+	AM-	PM+	PM-	Within	Outside	TDS	TAD	TDS/TAD	ADS	Grade	Status
7	...	0	35	1	24	12	0	17	20	71	20	0.28169	0.274469	5	FAILED

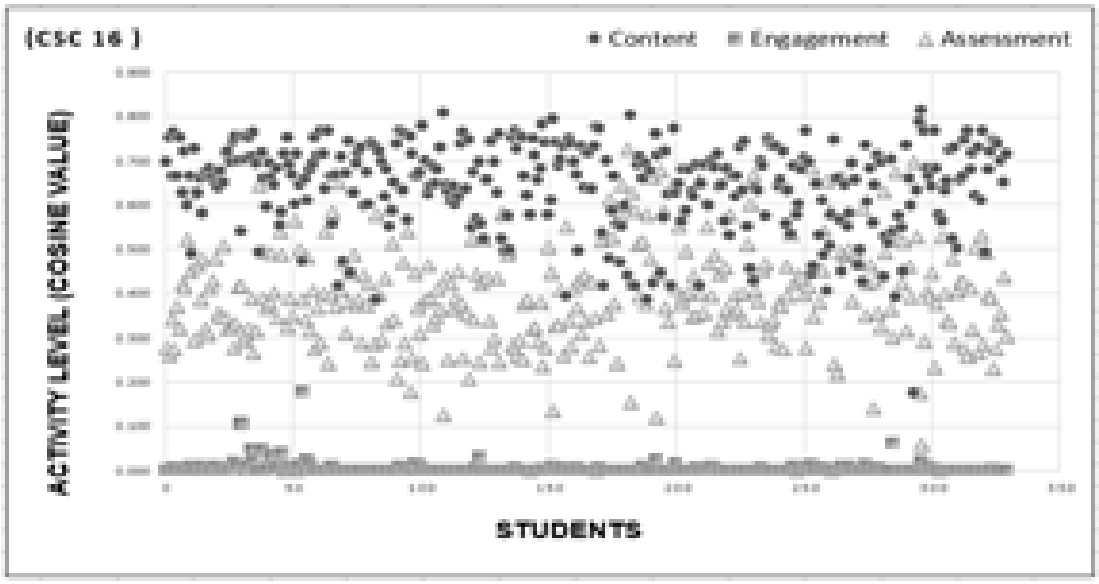
Figure 4. Classification model indicators

TEST AND RESULTS

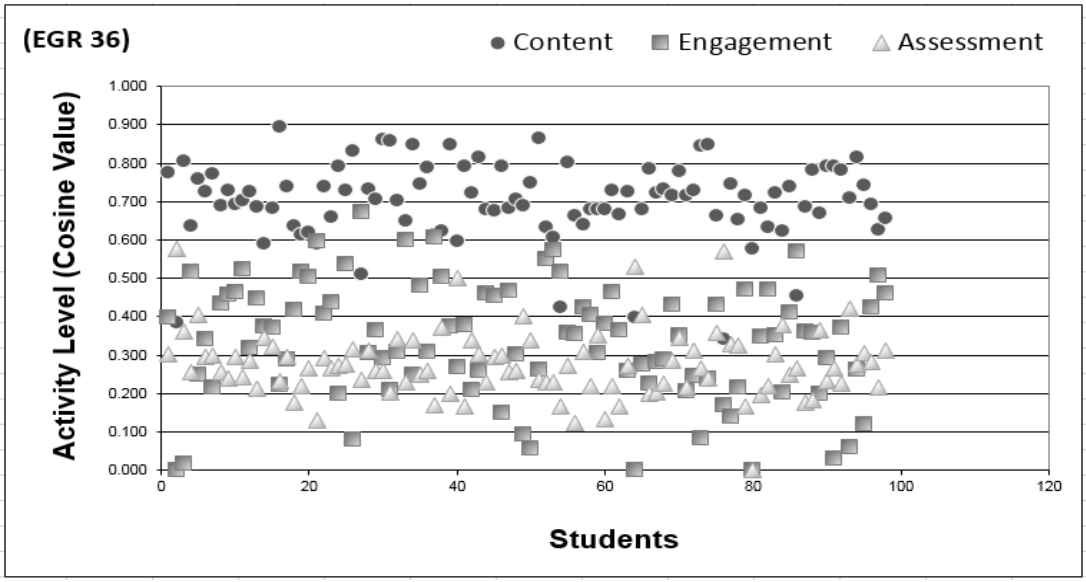
Activity visualization

Initial experiments were conducted exploring more than 285,000 action log data generated by 885 students in three different blended courses: Basic Computing (CSC16), Engineering Management (EGR36), and Elementary Statistics (MAT22). Figure 5 (graphs A-C) shows the results of applying VSM to activity vectors generated using this dataset. Each point in these graphs represents a student's cosine score for each activity dimension per class. Although students are anonymously depicted, the graph clearly indicates the overall degree of activity among the cohorts. The visualizations also indicate that different classes vary widely in how they utilize the tools provided within Moodle.

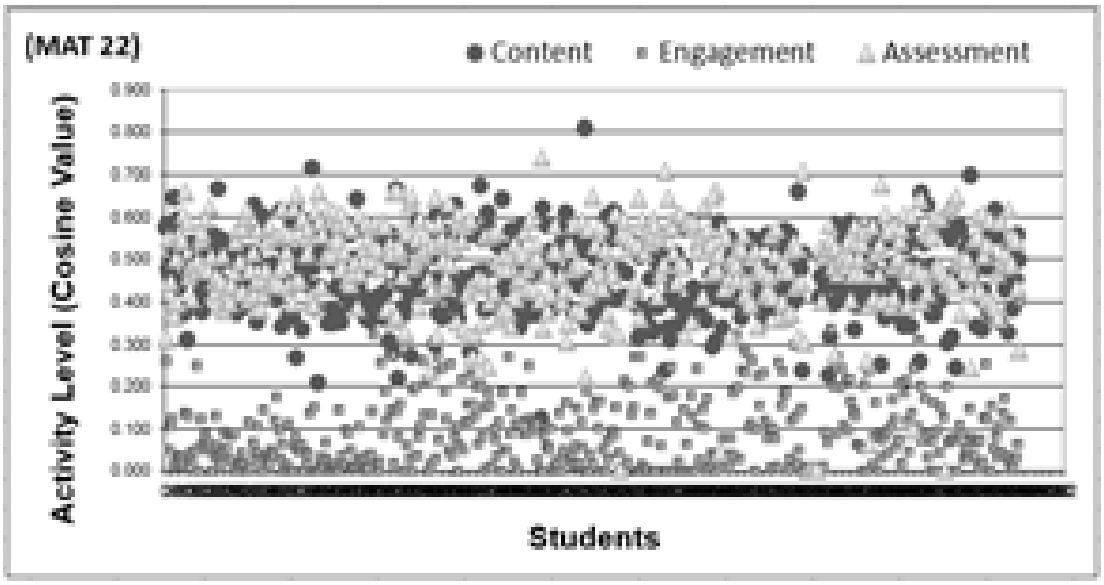
However, within a certain class, students take on a similar set of behavior in allotting time and effort between different tools and activities. Students from CSC16, for example, generally log-in into Moodle to access lecture materials with little regard for forum engagement. MAT22 students display an equal level of preference for accessing lecture materials and taking assessment activities with some forum discussions initiated, whereas EGR16 students seem to prioritize content access and forum engagement over access to assessment tasks. These visualizations can help course administrators in determining the type of strategic interventions that each class/course would need to ensure that student's activities are kept in line with the intended pedagogical outcomes.



(A)



(B)



(C)

Figure 5 (A-C). Graphs depicting student activity in different courses

Performance Prediction

In performing performance prediction the enriched data was first subjected to another round of pre-processing in order to identify and remove collinear attributes. First, attribute columns with all zero values was deleted from the data matrix, these include the "assign view submit assignment form", "quiz review", and "user view" attributes. Correlation analysis was then applied to every pair of attributes using Microsoft Excel. An absolute value of 0.25 was used as a threshold for identifying attributes with no significant correlation with the student's Final Grade (FG attribute). These attributes, which was subsequently removed from the dataset, includes the "course recent", "forum add discussion", "forum view discussion", "forum view forum", "URL view", "user view all", "AM+", and "TDS/TAD" attributes (marked in yellow in Figure 6).

Regression analysis was then applied to the resulting dataset to further process the attributes and identify strong predictors of the student's performance. This was done using Microsoft Excel, results of the analysis using are shown in Figure 7. A value of $p < 0.05$ was used to test for a significant relationship with the FG attribute. Out of the 16 remaining attributes, 9 were eliminated based on the regression results. These attributes include the following: "quiz attempt", "quiz close attempt", "quiz view", "resource view", "total", "AM-", PM+", PM-", and TAD. These attributes are highlighted in yellow in Figure 7.

Finally, using Python version 3.6.2, several classification algorithms available in the Scikit-learn machine learning library was applied to the enriched data representation to determine how well these algorithms can predict student performance given the

available amount of data. The remaining attributes used as predictors in this experiment with their corresponding p-values are shown in Figure 8.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
assign view	1.000																							
course recent	0.148	1.000																						
course view	0.570	0.200	1.000																					
forum add discussion	-0.251	0.137	0.130	1.000																				
forum view discussion	-0.185	0.187	0.175	0.713	1.000																			
forum view forum	-0.272	0.167	0.288	0.753	0.790	1.000																		
quiz attempt	-0.380	-0.024	0.263	0.268	0.199	0.422	1.000																	
quiz close attempt	-0.379	-0.025	0.263	0.272	0.203	0.425	0.999	1.000																
quiz continue attempt	0.386	0.044	0.339	-0.150	-0.111	-0.100	0.033	0.033	1.000															
quiz view	-0.267	0.029	0.410	0.249	0.222	0.470	0.905	0.903	0.090	1.000														
resource view	0.846	0.176	0.612	-0.149	-0.094	-0.177	-0.334	-0.334	0.406	-0.221	1.000													
url view	-0.354	0.074	0.258	0.557	0.500	0.692	0.569	0.568	-0.090	0.556	-0.259	1.000												
user view all	0.112	0.150	0.153	0.119	0.172	0.290	-0.002	0.000	0.007	0.046	0.131	0.107	1.000											
Total	0.575	0.191	0.867	0.131	0.192	0.283	0.307	0.307	0.705	0.439	0.637	0.245	0.160	1.000										
AM+	0.151	0.051	0.297	0.098	0.247	0.240	0.098	0.096	0.110	0.128	0.197	0.205	0.053	0.300	1.000									
AM-	0.590	0.075	0.521	-0.149	-0.103	-0.087	-0.011	-0.010	0.826	0.087	0.568	-0.114	0.060	0.755	0.088	1.000								
PM+	0.395	0.188	0.700	0.210	0.224	0.292	0.277	0.280	0.240	0.367	0.443	0.290	0.142	0.657	0.058	0.285	1.000							
PM-	0.190	0.143	0.576	0.232	0.259	0.390	0.417	0.416	0.233	0.500	0.271	0.356	0.143	0.592	0.246	0.119	0.231	1.000						
Within	0.527	0.101	0.507	-0.054	-0.025	-0.038	-0.029	-0.027	0.702	0.066	0.499	-0.052	0.008	0.679	0.040	0.809	0.479	0.015	1.000					
Outside	0.410	0.173	0.777	0.184	0.245	0.375	0.412	0.411	0.395	0.517	0.493	0.339	0.206	0.815	0.363	0.408	0.551	0.799	0.160	1.000				
TDS	0.020	-0.005	0.241	0.161	0.133	0.191	0.417	0.419	0.199	0.392	0.031	0.202	0.021	0.327	0.071	0.199	0.253	0.277	0.173	0.329	1.000			
TAD	0.587	0.194	0.896	0.150	0.201	0.299	0.276	0.276	0.574	0.400	0.664	0.268	0.186	0.954	0.313	0.669	0.680	0.637	0.580	0.862	0.310	1.000		
TDS/TAD	0.016	0.022	0.054	-0.011	-0.011	0.000	0.039	0.034	0.045	0.040	0.023	0.045	0.019	0.058	0.075	0.020	0.061	0.029	0.011	0.069	-0.366	0.063	1.000	
ADSS	0.348	0.150	0.593	0.172	0.184	0.246	0.317	0.315	0.353	0.367	0.375	0.274	0.103	0.644	0.215	0.444	0.479	0.445	0.404	0.587	-0.006	0.672	0.195	1.000
FG	-0.404	-0.098	-0.492	-0.076	-0.054	-0.116	-0.294	-0.295	-0.253	-0.304	-0.368	-0.156	-0.076	-0.533	-0.199	-0.375	-0.418	-0.353	-0.305	-0.520	-0.270	-0.553	-0.116	-0.471

Figure 6. Results of correlation analysis on the data matrix

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	4.758	0.111	42.890	0.000	4.540	4.976	4.540	4.976
assign view	-0.011	0.002	-7.007	0.000	-0.015	-0.008	-0.015	-0.008
course view	0.004	0.001	2.487	0.013	0.001	0.006	0.001	0.006
quiz attempt	0.010	0.064	0.159	0.874	-0.115	0.135	-0.115	0.135
quiz close attempt	-0.062	0.064	-0.969	0.333	-0.187	0.063	-0.187	0.063
quiz continue attempt	0.002	0.001	2.245	0.025	0.000	0.004	0.000	0.004
quiz view	0.003	0.002	1.493	0.136	-0.001	0.007	-0.001	0.007
resource view	0.000	0.002	-0.238	0.812	-0.003	0.003	-0.003	0.003
Total	0.001	0.001	0.985	0.325	-0.001	0.003	-0.001	0.003
AM-	0.001	0.001	0.708	0.479	-0.002	0.003	-0.002	0.003
PM+	0.001	0.001	1.214	0.225	-0.001	0.004	-0.001	0.004
PM-	0.002	0.001	1.697	0.090	0.000	0.004	0.000	0.004
Within	-0.003	0.002	-2.238	0.025	-0.006	0.000	-0.006	0.000
Outside	-0.004	0.001	-2.687	0.007	-0.007	-0.001	-0.007	-0.001
TDS	-0.002	0.001	-2.172	0.030	-0.003	0.000	-0.003	0.000
TAD	-0.001	0.001	-0.982	0.326	-0.004	0.001	-0.004	0.001
ADSS	-0.524	0.153	-3.433	0.001	-0.824	-0.225	-0.824	-0.225

Figure 7. Results of regression analysis on the resulting data matrix

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4.818	0.113	42.520	0.000	4.596	5.041	4.596	5.041
assign view	-0.005	0.001	-5.410	0.000	-0.007	-0.003	-0.007	-0.003
course view	0.005	0.001	3.935	0.000	0.002	0.007	0.002	0.007
quiz continue attempt	0.003	0.001	5.197	0.000	0.002	0.005	0.002	0.005
Within	-0.003	0.001	-4.746	0.000	-0.004	-0.002	-0.004	-0.002
Outside	-0.003	0.001	-6.612	0.000	-0.005	-0.002	-0.005	-0.002
TDS	-0.003	0.001	-4.804	0.000	-0.005	-0.002	-0.005	-0.002
ADSS	-0.820	0.150	-5.467	0.000	-1.115	-0.526	-1.115	-0.526

Figure 8. Final set of attributes used for predicting student performance

Table 5 provides the results of the prediction experiments for each group of students. Default parameter settings for the algorithms were used for these experiments along with a k-fold cross-validation where a value of k=10 was used. As shown, the overall best accuracy rating was obtained using the k-nearest-neighbor algorithm with an average accuracy of 72.8%. But the best accuracy per course was generated by the LDA algorithm for the Engineering course with an accuracy of 87.7% using the EGR36 course dataset.

Table 5. Classification Accuracy Ratings

	LR	**LDA**	**KNN**	**CART**	**RF**	**BayesNet**	**SVM**
CSC16	63.20%	60.70%	64.70%	62.00%	61.71%	53.80%	56.40%
EGR36	82.80%	87.70%	85.80%	80.80%	84.88%	77.70%	80.80%
MAT22	67.50%	67.70%	67.90%	57.40%	67.05%	67.20%	55.50%
Average	71.17%	72.03%	72.80%	66.73%	71.21%	66.23%	64.23%

These initial results indicate that the classification algorithms can modestly predict student performance. These modest performances can be due to the fact that there are other factors that can affect the performance of students beyond their study skills (Robbins et al., 2004). This is especially true in in blended courses, where students are immersed in two different learning environments. However, if investigated further, this sort of information could possibly provide a basis for identifying at-risk students and enabling instructors to provide effective formative feedbacks as early as possible.

DISCUSSION AND FUTURE WORK

This paper proposes a novel approach for processing course log data obtained from Moodle-based blended courses in order to visualize patterns of student activity and to determine whether these log data can also be used to predict and anticipate the academic performance of students. Logs of student activities were summarized and processed using the Vector Space Model approach. This resulted in a novel vector-based form of representation called an Activity Vector which can be used to map and understand the positioning of each student in a latent Activity Space. Results clearly indicate that the activity space coupled with a one-hot vector representation for each unique activity dimension can be used to visualize the differences in level and type of activity preferences of students both individually and per class. In the long run, these

types of visualizations could be used to monitor which student and/or class requires immediate and specific pedagogical adjustments.

Much work, however, needs to be done in terms of refining the process applied to the data. In particular, the log data should be time-sliced and processed on a per period basis in order to determine whether and how the student's level and type of activity changes over time. An extended approach for further enriching the VSM-based activity vector was also proposed by processing the datetime and IP address metadata of the log data. This enriched vector representation can be used as input to any classification/predictive model. An eventual application of this model is the immediate identification of at-risk students based on the actions they are exhibiting in the online environment.

Experiments testing the enriched representation on several machine learning algorithms using Python and the Scikit-learning library were also performed. The results indicate that classification algorithms can modestly predict a student's academic performance and, in particular, model the difference between high, low, and failed performances. This modest result indicates that there are more factors that need to be considered in predicting the performance of students in blended courses. More powerful machine learning classification techniques can be tested on the enriched vector representation to further isolate these factors and to determine whether the classification accuracy can still be improved.

ACKNOWLEDGEMENT

The authors wish to thank Mr. Jasper Vincent Alontaga of the Institute of Technology-based Learning (ITBL) for providing assistance in extracting the sample data from the Moodle database and for the insightful discussions. We would also like to thank the Research Office of Jose Rizal University for providing financial support to this study.

REFERENCES

Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, *2*(2), 43-47.

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, *2*(6), 63-69.

Dias, S. B., & Diniz, J. A. (2014). Towards an enhanced learning management system for blended learning in higher education incorporating distinct learners' profiles. *Journal of Educational Technology & Society*, *17*(1), 307-319.

Estacio, R. R. & Raga Jr, R. C. (2017). Analyzing students online learning behavior in blended courses using Moodle. *Asian Association of Open Universities Journal*, *12*(1), 52-68.

Fraser, K. C., & Hirst, G. (2016). *Detecting semantic changes in Alzheimer's disease with vector space models*. *Proceedings of LREC 2016 Workshop: Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016), Monday 23rd of May 2016* (No. 128). Linköpings Universitet: Linköping University Electronic Press. Retrieved from http://www.ep.liu.se/ecp/article.asp?issue=128&article=001&volume=

Kabakchieva, D. (2013) Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, *13*(1), 61-72.

Kanuka, H., Brooks, C., & Saranchuck, N. (2009). Flexible learning and cost effective mass offerings. Paper presented at the International Conference on Improving University Teaching, Vancouver, Canada.

Nicdao, J. M. (2013). Teaching literature through blended learning in the Philippines. In L. Gomez Chova, A. Lopez Martinez, I. Candel Torres (Eds.), *EDULEARN 13*. Paper presented at the 5th International Conference on Education and New Learning Technologies, Barcelona, Spain (pp. 2817-2822). Spain: IATED.

Norberg, A., Dziuban, C. D., & Moskal, P. D. (2011). A time-based blended learning model. *On the Horizon*, *19*(3), 207-216.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825-2830.

Rice, W. (2006). *Moodle e-learning course development*. Birmingham, UK: Packt Publishing.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*(2), 271-288.

Salehi, M., Pourzaferani, M., & Razavi, S. A. (2013). Hybrid attribute-based recommender system for learning material using genetic algorithm and a multidimensional information model. *Egyptian Informatics Journal*, *14*(1), 67-78.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422.

Singhal, A. (2001). Modern Information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, *24*(4), 35-43.

Sreeja, P. S., & Mahalakshmi, G. S. (2016). Comparison of probabilistic corpus-based method and vector space model for emotion recognition from poems. *Asian Journal of Information Technology*, *15*(5), 908-915.

Wikipedia. (2017). *Blended learning*. Retrieved from https://en.wikipedia.org/wiki/Blended_learning.

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. In *Robust Data Mining* (pp. 27-33). New York: Springer.

Younge, K. A., & Kuhn, J. M. (2016). Patent-to-Patent similarity: A vector space model. Retrieved from SSRN: https://ssrn.com/abstract=2709238. doi: 10.2139/ssrn.2709238.

Zorrilla, M., Millan, S., & Menasalvas, E. (2005). Data web house to support web intelligence in e-learning environments. 2005 IEEE International Conference on Granular Computing, (Vol. 2, pp. 722-727). IEEE.

AUTHOR'S BIOGRAPHY

Rodolfo C. Raga Jr. received his PhD in Computer Science degree from the De La Salle University, Manila, Philippines in 2013. He's an Associate Professor in the College of Computer Studies and Engineering of Jose Rizal University, Philippines. His research interests lie in natural language processing, educational data mining, learning analytics, and academic e-learning.

Jennifer D. Raga received her Master in Information Technology degree from the University of LaSalette, Santiago City, Philippines in 2009. She is currently the MIS Manager at Western Marketing Corporation and a part-time IT lecturer. Her research interests lie in knowledge management, business analytics, and corporate e-learning.