(m)Oodles of Data
Mining Moodle to understand Student Behaviour

Kevin Casey,
Griffith College Dublin.
kevin.casey@gcd.ie

Paul Gibson,
IT-Sud Paris.
paul.gibson@it-sudparis.eu

International Conference on Engaging Pedagogy 2010 (ICEP10) National University of Ireland Maynooth, September 2, 2010

Abstract

With 32 million users across approximately 3 million courses worldwide, Moodle has proven to be an immensely popular and important tool in education. One feature provided by Moodle is a rich source of information about student access to online material. While important, this information is presented in a raw format with little indication about how it might be used.

In this paper, we examine Moodle viewing statistics from an Irish third-level institution. We examine correlations between these statistics and student results on degree and masters courses. We identify the circumstances where the correlation can help predict poor performance (and hence allow early intervention). In the analysis we find some interesting associations between student behaviour on Moodle and their final grade. Some of these associations reinforce beliefs the authors already had about Moodle usage, and some results were surprising.

Keywords

Moodle, Data Mining, Student Behaviour, Early Intervention

1. Introduction and Motivation

Moodle, the popular Virtual Learning Environment (VLE), has attained enormous popularity over the last decade. The numbers are impressive with almost 37 million users across 3.7 million courses in 211 countries as of September 2010 (Moodle, 2010). Moodle has achieved this popularity for a number of reasons, not least for its open-source nature, extensibility and ease of install and administration.

Griffith College Dublin, currently uses Moodle for the majority of taught courses. In the Computer Science faculty, it is used intensively. For many courses in the faculty, Moodle is the main (if not exclusive) mode of lecture note distribution and coursework submission.

Given that students interact so intensively with Moodle, the question naturally arises: *What can Moodle tell us about our students?* There is a wealth of data stored in the Moodle system about the activities of users, both teachers and students. The kind of information recorded is typically of the who, what, when, where variety (Zhang *et al*, 2007). To be specific, Moodle records each action within the VLE, the user who initiated that action, when they initiated it and where they logged in from. A typical extract from a Moodle log is shown in Table 1.1. This short extract shows John Doe viewing the main page of a course at 9:28pm and viewing a resource a moment later. A few minutes later, the teacher begins to edit the course page on Moodle.

Course	Time	IP Address	Name	Action	Id
...	...	...	...	...	...
BSCH-SS/Dub/FT	14 July 2010, 09:28 PM	86.34.23.221	John Doe	course view	3935
BSCH-SS/Dub/FT	14 July 2010, 09:29 PM	86.34.23.221	John Doe	resource view	61091
BSCH-SS/Dub/FT	14 July 2010, 09:33 PM	192.168.1.136	Jane Doe	course editsection	1
...	...	...	...	...	...

Table 1.1 – Example of a Moodle log file

The level of such recorded data is quite fine-grained and, consequently, the volume of data can be overwhelming without some tools to help interpret it. For example, a one year Computer Programming course which we discuss in section 2 had over 33,000 actions recorded from approximately 40 active participants .

The idea of analysing student behaviour in Moodle is not new. The GISMO project (Mazza and Milani, 2005) offers visualisations of many of the statistics in the Moodle logs, but performs little automatic analysis. Moodog (Zhang *et al*, 2007), another Moodle log analysis tool, performs a similar role with an emphasis on visualisation of data rather than analysis. Other work (Superby *et al*, 2006) has examined data outside the VLE arena from sources such as questionnaires to classify students according to their likely levels of achievement. Some promising work has been carried out using the Keel framework (C. Romero *et al* 2008,C. Romero *et al* 2007) with advanced data mining techniques on Moodle logs. However, much of the existing analysis operates on a large number of variables, classifying students quite successfully, but without

any meaningful explanation about how or why the Moodle logs relate to the classification of students.

While the work in this paper concentrates on analysing data arising from Moodle usage, there are many other popular VLEs in use. Many of these such as Blackboard (Wells P *et al*, 2008) and StudyWiz (Nassau BOCES, 2007) maintain similar logfiles and the work here can easily be extended to them. Although this paper concentrates on the use of Moodle at the tertiary level, VLEs are also being employed elsewhere. As far back as 2006 in the UK (Becta 2006, Kitchen *et al* 2006), surveys have found that 22-30% of primary schools and 50-57% of secondary schools had some form of VLE deployed.

The authors have constructed an analysis tool to extract information from Moodle to examine issues relating to the data recorded there. The first issue revolves around finding correlations between student performance and Moodle usage. We do this to examine the potential to identify weak students though their Moodle activity (and inactivity). The second issue we examine is that of student usage patterns (independent of the students' final grades). As educators, we can use this information to guide us as to how to use Moodle more effectively.

2. Experimental Setup

The data for the analysis below was collected between September 2009 and May 2010 for a number of modules in different full-time courses. Three courses were involved: BSc I, a first-year course in Computer Science, BSc IV, a final year course in Computer Science and MSc DM, a taught MSc in Digital Media. All of the modules for BSc I were suitable for analysis, since Moodle was used extensively to deliver all of them. In addition, the number of students taking each module was sufficiently high (see Table 2.1). A selection of subjects from the BSc IV and MSc DM courses were chosen. Others were rejected, either because of smaller class sizes (a particular problem in the case of electives) or because the lecturer did not use Moodle enough to generate sufficient data for analysis.

Subject (level)	Name	Semesters	Students	Course Views	Resource Views
(BSc I)					
FC	Foundations of Computing	2	41	3110	3243
CP	Computer Programming	2	41	8930	5105
SS	Systems Software	2	41	5660	5622
BIS	Business and Information Systems	2	42	4641	2903
CH	Computer Hardware	2	42	3491	1786
ITS	IT and Society	2	43	3717	1763
(BSc IV)					
ANT	Advanced Network Theory	1	11	789	499
DM	Database Management	1	13	1414	1567
ET	Emerging Technologies	1	26	2091	1534
(MSc DM)					
ID	Interaction Design	1	24	1695	1804
IA	Internet Authoring	1	20	2237	1562
MP	Multimedia Programming	1	18	1928	2584
VC	Visual Communication	1	27	2248	1647

Table 2.1 – Meta-data for Moodle log files

All of our analysis has been performed on standard Moodle logs with no additional instrumentation required. The only additional data required for the analysis was the student grades themselves. Some repeat students appeared in the logs and in the grades data. We removed these repeat students from our analysis because their presence complicated analysis. We justify this decision by the fact that many of these students will have seen or downloaded much of the material from Moodle before, and therefore, will not be exhibit normal behaviour.

2.1 Resource view counting

Our first class of metric relates to counting the number of times students view resources and finding correlations with their final grades.

Does Moodle activity (total number of page views) relate to final grade?
A simple point to start is to examine if Moodle access (of any resource) can be correlated to a student's final grade. In order to do this, we first measured the total number of page views each student made throughout a semester. These values were then linked to student grades for the module in question. The ***Views*** column of table 2.2 shows the results from this analysis.

The correlation is a little erratic in BSc I, consistently positive in BSc IV and consistently negative in the MSc DM. By examining the scatterplots for BSc I, it was found that some students exhibited erratic behaviour, essentially repeating page requests to Moodle without any benefit to their final grade. On further analysis, these outliers tended to be students with particular difficulties outside the course (many had medical certificates on record).

Figure 2.1 shows the distribution of the average number of resource views across all first year subjects for each student who completed all subjects. It can be seen from the figure that some students view resources far more than is the norm. In first year, the subject that exhibited the highest correlation (a theme to be repeated later) was Computer Programming.

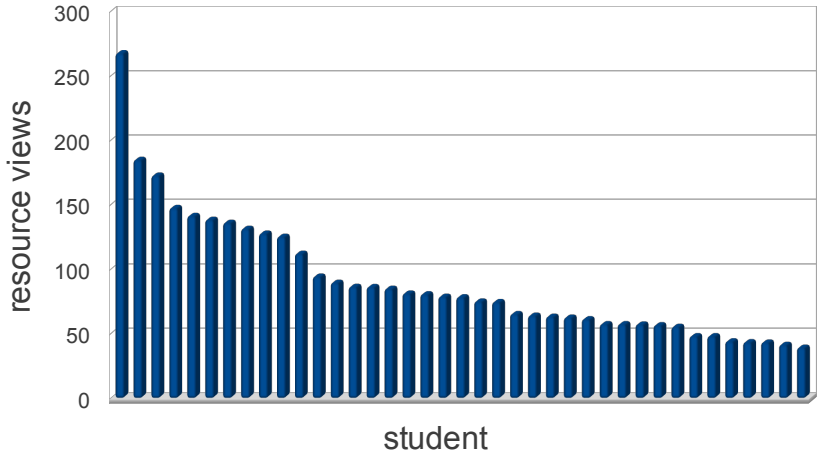

Figure 2.1 – Distribution of resource viewing among students in BSc I

BSc IV figures were better behaved, with all three selected subjects showing a mild positive correlation, as expected. However, the MSc DM figures are quite interesting and unexpected, showing a *negative* correlation between page views and final grade. We hypothesise that this negative correlation is due to the way that Moodle is used at MSc level to distribute material which is better suited to offline reading (ie printing hardcopies for reading and reflection). It would seem that repeated activity on such modules indicate that a student is not doing as much offline reading. We note that Visual Communication (VC), the subject with most of this kind of reading material has the most negative correlation between student activity and the final grade. We also note that the BSc I ITS has a large percentage of material, again designed for offline reading, which offers an explanation to the lower correlation observed in relation to other BSc I subjects.

Does Moodle activity (number of unique page views) relate to the final grade?
A variation of the analysis above is to filter out repeated views of the same page within a particular time period. We arbitrarily chose that time period to be within the same day. So, if a student visits a resource several times within one day, it will count as one view. The results of this analysis are shown in Table 2.2 under the ***Daily Views*** column. In most cases, the correlation under this new metric is similar or a little better that from before. An interesting variation of this time-period approach is to adjust the size of the window of analysis. While we currently consider views in the same day as being one view, further work will vary this window to other time periods, for example considering views within the same week to be one view.

Subject (level)	Views	Daily Views	Daily Course View
(BSc I)			
FC	0.07	0.12	0.26
CP	0.31	0.34	0.44
SS	0.02	0.06	0.12
BIS	0.20	0.22	0.28
CH	0.13	0.18	0.07
ITS	-0.12	-0.11	0.01
(BSc IV)			
ANT	0.36	0.40	0.30
DM	0.51	0.50	0.37
ET	0.52	0.51	0.41
(MSc DM)			
ID	-0.21	-0.09	0.20
IA	-0.23	-0.32	-0.09
MP	-0.17	-0.12	0.06
VC	-0.40	-0.41	0.02

Table 2.2 – Correlations using resource view counting

Do Moodle course logins (on a daily basis) relate to final grade?
A special case of the ***Daily Views*** analysis above relates to the course page for a particular module. Taking a similar strategy to before, we considered only the main

page for each module and measured the student views of it. Again, we considered multiple views on the same day to be just one view.

The results are shown in Table 2.2 under the ***Daily Course View*** column. There are some points of interest when comparing this metric to others. Most BSc I subjects display a higher correlation under this column, suggesting that students who check back at a course page regularly to see if there are updates, tend to achieve better results. The results under the MSc DM modules lose their negative correlation under this metric, as the process of viewing a course page to check for updates does not seem to have the negative associations that re-reading old material does.

Does it matter whether a student looks at resources on-campus or off-campus?
In table 1.1, a sample log from Moodle is shown. In this extract, you can see that the IP address for each action is recorded. This allows us to easily determine whether accesses were on-campus or off-campus. On-campus accesses come from one particular set of known IP addresses while off-campus accesses are from IP addresses outside that set. For this analysis we took a similar approach to the Daily Views and Daily Course Views measures above, eliminating multiple views of the same resource on the same day. However, we did count a view of the same resource on-campus and off-campus on the same day as different views. Table 2.3 shows the result of this analysis. An overall comparison of on-campus and off-campus activity (on the basis of resource views) is also shown. The information in the table suggests that the activity a student performs off-campus is a better indicator of their performance than on-campus activity.

Subject (level)	Daily Views			Daily Course View			Activity	
	Off Campus	On Campus	Overall	Off Campus	On Campus	Overall	Off Campus	On Campus
(BSc I)								
FC	0.03	0.15	0.12	0.23	0.18	0.25	63.98%	36.02%
CP	0.37	0.20	0.35	0.37	0.35	0.44	45.10%	54.90%
SS	0.14	-0.03	0.06	0.18	-0.05	0.12	56.73%	43.27%
BIS	0.32	0.02	0.22	0.27	0.14	0.27	59.83%	40.17%
CH	0.26	-0.01	0.19	0.12	-0.07	0.06	61.44%	38.56%
ITS	-0.02	-0.16	-0.12	0.05	-0.05	0.01	64.06%	35.94%
(BSc IV)								
ANT	0.46	0.13	0.42	0.39	0.04	0.31	83.69%	16.31%
DM	0.52	-0.21	0.52	0.36	-0.06	0.21	72.87%	27.13%
ET	0.67	-0.30	0.52	0.64	-0.28	0.45	45.32%	54.68%
(MSc DM)								
ID	-0.14	0.07	-0.11	0.11	0.17	0.17	57.92%	42.08%
IA	-0.24	-0.24	-0.31	-0.28	0.19	-0.15	47.63%	52.37%
MP	-0.18	0.13	-0.13	-0.12	0.31	0.06	55.76%	44.24%
VC	-0.49	0.07	-0.42	-0.20	0.34	-0.02	62.73%	37.27%

Table 2.3 – Correlations using on-campus and off-campus resource view counting

In certain cases, such as MScDM VC where activity in general is a negative indicator of student performance, a student's off-campus Moodle activity is an even stronger negative indicator. Therefore, it seems from this set of results that a student's off-campus Moodle activity is more significant than their on-campus activity (either positively or negatively) in relation to their final grade.

The relative effort students put into each subject in BSc I was also measured in terms of the number of resource views. We found that SS and CP had the most views (27% and 25% respectively) while CH and ITS had the least (around 8% each). There is a need to be careful when interpreting figure 2.1 which compares the student effort versus their final grade. The subjects on which they spent most effort (on Moodle) are ones which have historically proven to be difficult. Consequently many more resources are placed on Moodle for the students and they are encouraged to spend time looking at them.

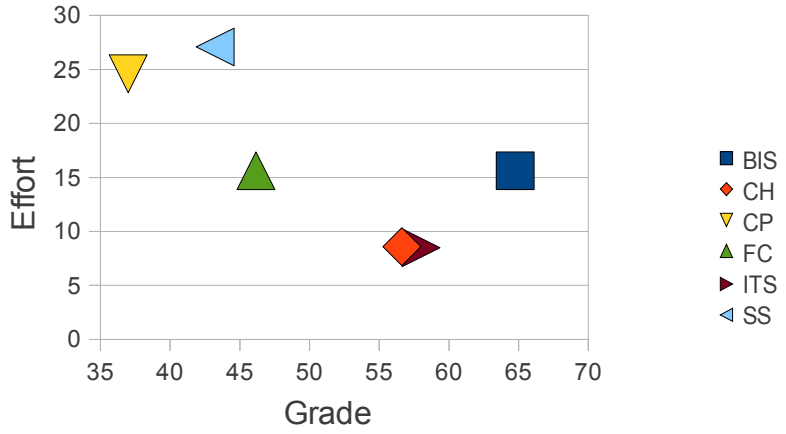

Figure 2.1 – Average effort vs average grade

2.2 A closer look at resources

In this section, we take a closer look at the individual resources that students access. We examine the results of analysis pertaining to the amount of available material a student reviews and take a fine-grained look at the individual resources that better students view on Moodle.

Does Moodle coverage (amount of material reviewed at least once) relate to the final grade? Many resources are made available on Moodle during the course of a module. However, this does not mean that all students will read all resources. Nor does it mean they should (as we explain later). In this metric, we compare the per-student coverage (percentage of material read at least once) against the student grade. Table 2.4 shows the results for a selection of subjects from years one and four of the BSc degree and from the MSc in Digital Media.

While the table shows some interesting figures in relation to coverage, there is little to show consistently in terms of correlation to a students final grade. What is interesting is how selective students are when it comes to coverage of material. Some of this can be explained by the instructor's teaching style. For example, in BSc I, the lower coverage figures come from BIS and ITS, which are both taught by the same lecturer. In BSc IV, the ET lecturer adopted a strategy of making substantial and optional material available on Moodle to students. In summary, this metric tells us more about a lecturer's mode of Moodle usage rather than offering any insight into student behaviour.

Subject (level)	Coverage	Correlation
(BSc I)		
FC	64.46%	0.00
CP	62.60%	0.32
SS	61.59%	0.20
BIS	51.52%	0.33
CH	63.64%	0.28
ITS	57.46%	0.10
(BSc IV)		
ANT	87.35%	-0.20
DM	46.94%	0.56
ET	35.94%	0.53
(MSc DM)		
ID	50.33%	0.00
IA	67.58%	-0.18
MP	82.10%	0.01
VC	56.27%	-0.38

Table 2.4 – Moodle coverage and associated correlation

Are some resources better indicators than others? In a variation of the ***Daily Views*** metric above, we consider individual resources and correlate student behaviour in relation to that resource to their final grade. The purpose here is to identify if student behaviour in relation to a particular resource (or class of resource) is more important than others. Table 2.5 shows the resources with the highest correlation to student grades from the subject SS on the BSc I programme. In this table, we show the correlation between student views of the resource and their grade, the readership (the percentage of students who accessed that resource at least once), and the title and description of the resource. The main result of note from the table is that none of the top five correlating resources are core materials, eg lecture notes. Instead they are optional resources to assist in understanding topics covered in lectures, and in the case of the 'Review' resource, to assist revision. Apart from the latter, a minority of students accessed each resource. The correlation figures suggest that this minority were the better performing students.

Correlation	Readership	Resource	Description
0.40	19.51%	Race conditions and hacking	A good example of race conditions
0.37	26.83%	HAL- wikipedia definition	Clarification of a term used in a lecture
0.34	34.15%	Cross Compilers	Examples of a concept presented in lectures
0.33	78.05%	Review	End of year course summary
0.28	29.27%	Race Conditions	A discussion on race conditions

Table 2.5 – Resources with highest correlation to final grade for the BSc I SS module

Table 2.6 shows similar information for the BSc IV course, ET. This is a 100% assessed group project subject where individual members within a group generally took on specific responsibilities. The pattern is similar to that observed in the BSc I SS module. This is that viewing of non-core but relevant resources is a good indicator of student performance. In this case, it is noted that the resource 'Rails-Ajax' could be considered a core resource since it is a set of slides from a lecture given to the students. However, given the nature of the roles assumed within groups, this resource was of greatest interest to those who chose the most challenging roles with the project

teams. Readership for 'All About Story Maps' was quite high because it involved an activity which cut across the various roles in each team.

Correllation	Readership	Resource	Description
0.66	23.08%	In Place Editor – example	Sample code
0.53	26.92%	ActionMailer tutorial	Sample code
0.48	30.77%	Rails - Ajax	Lecture slides
0.48	61.54%	All About Story Maps	How to build story maps
0.46	15.38%	Sample SCRUM meeting	A sample of how to write up a meeting

Table 2.6 – Resources with highest correlation to final grade for the BSc IV ET module

Table 2.7 shows the resources with the highest correlation to student grades from the subject MP on the MSc DM programme. In general, the readership for these resources was much higher with students attaching some significance to them. The resource where continued viewing was associated with higher grades (eg 'Useful Javascript tips') was in a state of constant flux, being updated at least once per week during term in response to student requirements. The rest of the resources with highest correlation to student performance were again, optional resources such as examples, revision notes and code solutions.

Correllation	Readership	Resource	Description
0.36	100.00%	Useful Javascript tips	A reference resource for programming
0.34	88.89%	Worksheet solution	Solutions to selected worksheets
0.31	88.89%	Puredata - revision	Revision notes
0.30	77.78%	Puredata – example	Sample code
0.29	83.33%	Bookmarklet example	Sample code

Table 2.7 – Resources with highest correlation to final grade for the MSc DM MP module

2.3 Time-based metrics

Apart from eliminating repeated views of resources within the same day in earlier analysis, so far the issue of time has been ignored. In this section we examine if the times when students access resources can yield any useful information.

How does activity on Moodle vary across the week (and across courses)? To get an insight into the days of the week where students are most active on Moodle, we examined the Moodle logs on a per student basis, measuring and separating out activity on a weekday basis. Table 2.8 shows the student activity across the week for 6 first-year courses on a per-course basis. There are no real surprises or significant variations among subjects that cannot be explained by timetabling. Weekdays are quite busy with a tail-off in activity on Fridays. Weekends are quieter with Sunday being the busier of the two days.

Subject	Sun	Mon	Tue	Wed	Thu	Fri	Sat
CH	9.57%	21.09%	18.12%	22.02%	13.25%	8.26%	7.68%
BIS	10.47%	19.48%	16.90%	15.01%	14.74%	16.07%	7.33%
CP	7.61%	13.48%	26.49%	18.29%	17.27%	10.17%	6.68%
FC	7.93%	13.05%	26.57%	19.30%	16.17%	8.54%	8.44%
ITS	8.22%	16.66%	14.83%	16.56%	19.48%	17.41%	6.83%
SS	6.82%	13.49%	24.67%	15.87%	20.02%	12.48%	6.64%
Overall	8.20%	15.59%	22.12%	17.52%	17.24%	12.18%	7.14%

Table 2.8 – First year BSc Moodle activity

Is activity on a certain day a better indicator of the final grade than on other days? Although the information from table 2.8 does not tell us anything new, we can correlate activity on certain days with students' final grades. In table 2.9 we show this correlation. There are a few points of interest here. Firstly, of all the days, activity at weekends seems to be associated with higher grades in general. Secondly, we note the peak in correlation between activity on a Friday and higher grades in the subject Computer Programming. We explain this by the observation that, for this particular subject, a special revision session took place on Fridays. Students who attended these tutorials did better overall. Again, we note the consistently low correlations for the ITS subject. As noted in section 2.1, these are related to the fact that much of the ITS material is better suited to offline reading and extensive online activity is a neutral to negative indicator.

Subject	Sun	Mon	Tue	Wed	Thu	Fri	Sat
CH	0.35	-0.08	0.10	0.08	-0.04	0.09	0.31
BIS	0.15	0.07	-0.02	0.28	0.28	0.32	0.39
CP	0.38	0.19	0.24	0.42	0.27	0.45	0.39
FC	0.20	0.18	-0.16	0.10	0.33	0.13	0.35
ITS	0.14	0.03	-0.11	-0.03	-0.06	-0.17	0.02
SS	0.24	-0.03	-0.21	0.15	0.11	0.12	0.35

Table 2.9 – Correlation between activity and final grade

Is there a link between student 'eagerness' (how quickly they view a published resource) and their final grade? Our analysis tool currently generates statistics relating to when a resource is first viewed by all students. We calculate the average first-view for each resource and then, for each student record the number of seconds from the average it was before they viewed the resource for the first time. We use the term first-view from average (FVFA) to describe this metric. If the FVFA is negative, the student viewed the resource before other students (on average). If positive, the student viewed the resource later than average. We then calculate the average of all FVFAs per student across all resources, to get an indication as to whether the student is an eager viewer or late viewer of resources. In this case we would expect a negative correlation between grades and the average FVFA per student.

Overall, the results shown in table 2.10 display this negative correlation. There are some notable results in the table. The BSc IV subject ANT stood out as one such notable exception. There are two reasons to explain this. The first is that the sample size is quite small in this class. The second is that, after a little background research into the anomaly, it transpired that many students in the class had Moodle access problems and could not access published resources in a timely fashion. At the other extreme, strong negative correlation for the subject MScDM MP is suspiciously high. Examination of the scatterplot highlighted that a single outlier (a student who accessed everything at the absolutely last moment – and unsurprisingly did quite badly in the exam) contributed significantly to the negative correlation. Removal of this outlier dropped the correlation back to a more conservative -0.37. This susceptibility of the Pearson Correlation to outliers at the extremes is a known issue and, as we mention in section 3, Spearman Correlation may be a better choice for future work.

An interesting issue with respect to the FVFA measurements is that of choosing a penalty for non-viewing of a resource. For the figures presented in table 2.10, there was no penalty. We simply used average FVFA times where the student accessed resources. However by introducing a two-week penalty, included in the overall FVFA average for each non-viewed resource gave us an improved correlation between the FVFA measurements and final grades for the Computer Programming subject in BSc 1. Better techniques for choosing a suitable penalty are an interesting subject for future research. Even better correlation was achieved on the same subject by limiting FVFA analysis to resources with a minimum of 75% viewership. Using the two-week penalty and the new limit, we were able to achieve a correlation of -0.56 on CP.

Subject (level)	Correlation
(BSc I)	
FC	-0.08
CP	-0.44
SS	-0.27
BIS	-0.12
CH	-0.27
ITS	0.00
(BSc IV)	
ANT	0.28
DM	-0.51
ET	-0.25
(MSc DM)	
ID	0.01
IA	-0.09
MP	-0.72
VC	0.07

Table 2.10 – Correlation between FVFA and final grade

3. Conclusions and Future Work

Although our work is in the early stages, there are quite a few lessons apparent in the basic analysis performed so far. There were two interesting findings from section 2.1. The simple metric ***Daily Course Views*** proved to be a surprisingly good indicator for student performance. Also, high student activity levels on Moodle for certain subjects such as the MScDM VC is sometimes a negative indicator for student performance. We found this latter result somewhat surprising, particularly given the hype surrounding VLEs, but understandable given how Moodle was being used in the circumstances. We subsequently interviewed the lecturer on the VC module about the unusual behaviour, and his response was illuminating:

> *"For this subject, I use Moodle mostly to refer students to external resources. They are not actually supposed to be spending their time on Moodle, but on these external sites. In the laboratory, I can usually spot the weaker students. They are the ones flipping endlessly through my course pages."*

Section 2.1 also compared students' off-campus Moodle activity against their on-campus activity. In subjects where Moodle activity had a positive correlation with grades, we found that off-campus activity had an even more positive correlation. In

subjects with a negative correlation between activity and grades, we found that off-campus activity had an even stronger negative correlation with grades. In general, we can say that how a student uses Moodle off-campus is a better indicator of performance than how they use Moodle on-campus.

In section 2.2, we took a more resource-centered view of the Moodle statistics. We found that there is a large variation of coverage (or readership) of material between subjects. Some subjects, such as the BSc IV subject ET, had quite a lot of optional material and students were quite strategic about which resources they accessed. For many subjects we found that coverage was associated with better student grades, but there were exceptions for some subjects such as MScDM VC.

In section 2.3, we examined the time aspect of the Moodle logs more closely. We first broke down student activity on a weekday basis per subject. We found, almost uniformly, that Moodle usage over a weekend is associated with higher grades. We also discussed the First View From Average (FVFA) metric and examined how it was associated with student grades. We found a reasonable negative correlation using this metric, but we feel that by addressing such issues as which resources to include in the metric, how best to penalise non-viewing of a resource, and how to treat edited resources, stronger correlations can be found.

Our work can be improved significantly from a statistical standpoint, most importantly by obtaining larger data samples. We also need to consider other forms of correlation than Pearson Correlation. We plan to move our analysis in the immediate future to Spearman Correlation to avoid the sensitivity Pearson has to outliers in the tails of samples. Longer term we plan to use other techniques to discover complex relationships in the data. These include such techniques as neural networks, statistical classifiers, rule induction and fuzzy rule learning.

The most promising aspect to our preliminary work is that, of all of the modules analysed in this paper, BSc I Computer Programming exhibited one of the strongest correlations between student behaviour and final grade. Science courses are notorious for their low retention and first-time progression rates, particularly at the first year level (Seymour E and Hewitt N, 1997). Addressing this low retention rate is one of the main motivations behind our work. If we can identify associations between student behaviour on Moodle and their ultimate grades at the end of the year, we can use this information to our advantage through early intervention. In the case of the institution we examined, the opportunity for early intervention is greater due to the recent restructuring of the first-year course on the BSc programme, where the semesterised structure has been dropped in favour of (academic) year-long courses.

Although this affords new students the advantage of being given time to adjust to college life before having to sit exams, it also means that lecturers do not get feedback from student exams after the first semester. Steps have been taken to address this lack of information in terms of summative assessment during the academic year. However, if analysis of Moodle data can provide us with a measure of student engagement (or lack thereof), it could prove to be an invaluable supplement to the above.

Finally, we noted in section 2.1 that a handful of students on the BSc I course exhibited unusual behaviour in their usage of Moodle. In this particular case, we found that many of these students who viewed Moodle significantly *more* than their peers, had difficulties where medical certificates had been lodged with our institution at the end of the year. The fact that we were able to see this in the Moodle data raises the intriguing possibility of being able to detect students with likely difficulties ahead of time.

4. References

Kitchen S, Butt S and Mackenzie, H. (2006), Curriculum Online Evaluation: Emerging findings from the third survey of schools. Coventry: Becta.

Mazza R and Milani C (2005) Exploring Usage Analysis in Learning Systems: Gaining Insights from Visualisations, In AIED Workshops (AIED'05).

Moodle, (2010). Moodle.org: Moodle Statistics @ http://moodle.org/stats/

Nassau BOCES (2007) StudyWiz Overview, The Board Of Cooperative Educational Services of Nassau County @ www.nassauboces.org/dln/studywiz/Studywiz_Overview_04.00.pdf

Romero C, Ventura S, Garcia Salcines E. (2007) Data mining in course management systems: Moodle case study and tutorial. Computer & Education 49 (2007) 1-17

Romero C, Ventura S, Hervás C, Gonzales P. (2008) Data mining algorithms to classify students. I International Conference on Educational Data Mining (EDM). Montreal (Canada, 2008) 8-17

Seymour E. and Hewitt N. (1997) Talking about leaving. Boulder: Westview Press.

Superby, JF, Vandamme JP, Meskens N. (2006) Determination of Factors Influencing the Achievement of the First-year University Students using Data Mining Methods. Workshop on Educational Data Mining, 2006. pp.37-44.

Wells P, de Lange P, Fieger P (2008) Integrating a virtual learning environment into a second-year accounting course: determinants of overall student perception, *Accounting and Finance*, 2008, vol. 48, issue 3, pages 503-518

Zhang H, Almeroth K, Knight A, Bulger M, Mayer R (2007), in C. Montgomerie & J. Seale (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007 (pp. 4415-4422). Chesapeake, VA: AACE.