# ARTICLE | #6

**Title**
**Data Mining for Learning Analytics; does lack of engagement always mean what we think it does?**

**Author(s)**
Ed de Quincey
Theocharis Kyriacou
Mark Turner

**Contact**
e.d.quincey@keele.ac.uk

**School**
School of Computing and Mathematics

**Faculty**
Faculty of Natural Sciences

**Abstract**
**Context and Objectives:** Learning Analytics (LA) has the potential to utilise student data to further the advancement of a personalized, supportive system of HE (Johnson et al., 2013). A number of LA systems are now being developed but there have been few studies that have analysed the usage of Virtual Learning Environments (VLE) in order to identify which analytics techniques and sources of data accurately reflect student engagement and achievement. **Methods:** The interactions of 66 students with a Level 4 programming module on a VLE have been analysed via the simple K-means clustering algorithm to identify classes of behaviour and their characteristics. **Results:** Two prominent classes were found with students achieving higher marks attending the lectures and tutorials more regularly and accessing all types of material on the VLE more frequently than students in the lower achieving cluster. However, there were a number of exceptions that had low levels of engagement that gained high marks and vice versa. **Discussion:** A student's prior experience and characteristics of their degree programme need to be taken into account to avoid incorrectly interpreting high and low levels of engagement. **Conclusions:** The number of times students view online module materials will be an important factor for inclusion in any predictive LA models but must be able to take into account the differences in student backgrounds, delivery styles and subjects

## Context and Objectives

Traditionally a student's progress and level of engagement has been measured by assessment and physical attendance. However, in a student's day-to-day interactions with a university, other real-time measures are being generated and stored e.g. Virtual Learning Environment (VLE) interaction, Library and Online Journal usage. The analysis of this data has been termed Learning Analytics (LA) and defined as a method for "deciphering trends and patterns from educational big data ... to further the advancement of a personalized, supportive system of higher education." (Johnson et al., 2013). Higher Education (HE) has traditionally been inefficient in its data use (Siemens & Long, 2011) but LA has the potential to identify at-risk learners and provide intervention to assist learners in achieving success (Macfadyen & Dawson, 2010).

Examples of systems that support elements of LA include the University of Southampton's "Student Dashboard"; the Open University's Anywhere app; the University of Bedfordshire's student engagement system; London South Bank University's partnership with IBM (Perry, 2014); Purdue University's Course Signals (Arnold and Pistilli, 2012) and the Student Success System (Essa & Hanan, 2012). A detailed review of systems has been published by JISC (Sclater et al., 2016), who are currently in collaboration with 50 universities to build a learning analytics service for the UK HE sector (JISC, n.d.).

In order to build a predictive Learning Analytics system, a behavioural model built from an example training set of input observations e.g. previous student VLE interaction data, is needed. However, there have been few studies that have analysed the usage of pre-existing VLEs in order to identify which analytics techniques and sources of data accurately reflect student engagement and achievement. The work presented in this paper follows on from the study by de Quincey and Stoneham (2015) and analyses the VLE interactions of students for a Level 4 module using a clustering algorithm to identify potential groups of students with similar learning behaviour and to study the characteristics of these groups.

## Methodology

The intranet within the School of Computing and Mathematical Sciences (CMS) at the University of Greenwich has been incrementally developed since 2002 and contains the key information and supports the main tasks that a student needs in order to complete their modules(1) (Stoneham, 2012). This includes digital versions of coursework specifications, previous exam papers, screencasts and podcasts of some lectures, book lists, common teaching material, final year project documentation and relevant forms such as those for requesting extenuating circumstances, applying for ethical approval and for making general enquiries. Very few paper-based handouts are given to students so learning materials are only accessible to them via the intranet.

All student interaction with the CMS intranet is recorded in the form of server logs. When a user requests a file from a web server, an entry is recorded in a log file i.e. by loading a web page via a web browser, a user is making a request for a HTML file along with other files that are embedded components of that page such as images and videos; each of these file requests make an entry in a log. These server log entries contain information such as the name of the file that was requested, the address of the page that referred the user to the requested page, the IP address of the device that requested the file (this can indicate the location of the user), the time the file was requested and the username of the person requesting the file. As part of a previous study (de Quincey and Stoneham, 2015), functionality has been developed that takes this server log information and inserts it into a database, facilitating easier querying and analysis.

Server log data generated by 2,634 students across the School has been collected during the 2012-13 Academic Year with 2,544,374 interactions being recorded. Previous analysis (de Quincey and Stoneham, 2015) has suggested significant correlations between pairs of attributes on a Level 4 module called “COMP1314: Digital Media, Computing and Programming”. COMP1314 is a 30 credit introductory module to computers and programming, delivered via weekly 2 hour lectures and 1 hour practical tutorial sessions across both semesters by 2 different lecturers. It is assessed by 2 pieces of coursework and an exam.

During the running of this module in 2012-13, there were 14,467 interactions with resources and pages on the CMS intranet related to the module by the 53 students who were still enrolled by the end of the module. Significantly high correlation was found between a student’s final module mark and overall attendance at tutorial and lab sessions (r=0.64(1)) and a similar correlation between the final mark and their interactions with COMP1314 resources and pages on the CMS intranet (r=0.63).

Here, a more holistic approach has been used, with the simple K-means clustering algorithm (MacQueen, 1967) being applied to a subset of the data generated by the programming component of the module delivered in the second semester (over 10 weeks), including all 66 students who were originally enrolled on the module. The K-means algorithm attempts to find k clusters in a set of observations/samples. Once the algorithm is run and clustering is completed each sample is assigned to the cluster with the nearest centroid (cluster centre). The centroid of a cluster is one that best represents the cluster. The centroid’s attributes are computed by finding the means of the attribute values of the cluster’s members.

## Results

For the programming component of the module there were a total of 2,622 views of related materials (mean=39.7 views per student). The following table shows the breakdown of views for the different material types.

| Resource Type | Number of Files | Total Views | Avg. Views per Student (n=66) |
|---|---|---|---|
| Tutorial Instructions | 11 | 1559 | 23.6 |
| Lecture Slides | 231 | 825 | 12.5 |
| Coursework Specification | 1 | 127 | 1.9 |

**Table 1:** Breakdown of views per resource type

In order to determine clusters of student behaviour on the module, the following features were then considered:

- the student’s degree programme (a code comprised of “P” followed by a set of numbers)
- their coursework mark for the programming component of the module
- their physical attendance percentage in lectures and tutorials
- the number of times they have viewed module related programming materials such as lecture slides, tutorial instructions and the coursework (CW) specification.

Running the simple K-means algorithm on this set of data revealed the two most prominent classes of students with the following centroids (average values of the attributes considered):

| Attribute | Full Data (66 students) | Cluster 0 (40 students) | Cluster 1 (26 students) |
|---|---|---|---|
| programmeID | P11361 | P11361 | P03657 |
| CW Mark | 48% | 34% | 70% |
| Attendance | 61% | 55% | 70% |
| Total File Views | 40 | 24 | 64 |
| Tutorial Views | 24 | 15 | 37 |
| Lecture Views | 13 | 6 | 22 |
| CW Spec. Views | 2 | 1 | 3 |

**Table 2:** Returned clusters from the K-means algorithm

The above two descriptors of the two classes show a clear distinction between the performance of students within each cluster (according to their coursework mark). The better performing students in Cluster 1 (i.e. those who have achieved a 70% average mark) attended the lectures and tutorials more regularly and accessed all types of material on the CMS intranet more frequently than the students in Cluster 0.

Of greater interest however are “the exceptions” to the above inferences. The figure below shows the distribution of student marks compared to their degree programme (represented by the “P” code on the x-axis). Each point, representing a student, has been assigned a colour that relates to one of the 2 clusters detailed in Table 2 above.

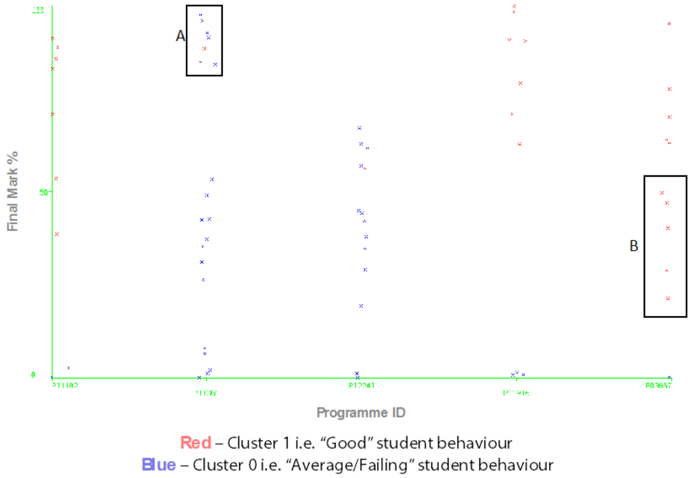

**Figure 1:** Clusters of behaviour related to degree programme and final mark %

It can be seen that for P11102, P11916 and P03657 there are examples of students that have high levels of engagement and are achieving high marks. It can also be seen that in P11361, P12241 and P11916 that there are examples of students who have low levels of engagement and have achieved low marks. This is perhaps to be expected and is in line with the correlation co-efficients detailed in the previous section.

However, there are a number of students on P11361 that had low levels of engagement and have achieved high marks (Box A) and conversely on P03657 showing high levels of engagement, whilst achieving low marks (Box B).

It seems that there are a minority of students on particular degree programmes who have achieved a high coursework mark but share the attributes of those who have not and vice versa i.e. students that have performed well on the module but have similar behaviour to those that have not. Potential reasons for this finding are discussed in the following section.

## Discussion

A number of previous studies have demonstrated the importance of attendance and the effect this has on final grades e.g. (Schmidt, 1983; Park & Kerr, 1990; Ryan et al., 2010). From the findings of this study, it is clear that attendance for this module was important but that online engagement i.e. viewing module resources, is equally important for the majority of students. For future LA systems, views of materials related to modules will therefore be an important factor for inclusion in any predictive models.

However, the use of unsupervised learning algorithms such as simple K-means needs further investigation as it is how exceptions to typical student behaviour are identified that may determine the success of LA. For this module, when looking at the particular students that are demonstrating the opposite behaviour to what is expected, a number of potential explanations can be suggested.

For students on P11361 who showed low levels of engagement but achieved high marks, one explanation could be that these students did not actually do the work themselves and achieved the high marks by collusion or plagiarism. However, P11361 is a Games and Multimedia Technologies degree which attracts students who have a pre-existing interest in programming and quite often prior experience in the subject. The more likely explanation (and having seen the students' progress through their degree) is that these students did not necessarily need to attend this set of lectures and tutorials and were not as reliant on the module resources as other students to complete the coursework.

The group that have shown the same level of engagement as the high achieving students but getting low marks is perhaps more worrying however. One explanation is that these students have not understood the materials and perhaps needed further support. P03657 is a Multimedia Technology degree which tends to attract students who are not sure which of the more specialised degree programmes to take and have lower levels of technical expertise when they start. These could be students who are engaged but their engagement is not being translated into higher levels of achievement. These students ideally would be highlighted by an LA system but perhaps would not be if their previous experience and motivation for doing the module was not taken into account. Another possible explanation is that some of these students have lower levels of digital literacy. Observing students interact with the CMS intranet in the tutorials(3) revealed that some will open the same file multiple times purely because they do not understand how to view separate pages in tabs or know how to download and save files in their own file stores. This repeated opening of files therefore is not increasing engagement, it is just the manifestation of their pre-existing digital practices or perhaps the increased reliance of having files available on demand online and not stored locally.

## Conclusions

It is clear therefore that although interactions with digital resources can represent engagement for students, there are other factors such as a student's prior experience and characteristics of their degree programme that need to be taken into account. For computing in particular this will become an increasingly important issue with the increased focus on programming within the National Curriculum and students coming into degrees with expected higher levels of experience and knowledge.

It is also important to note that this study has been performed on one Level 4 module with a particular structure both in face-to-face delivery and in the resources that are provided. Models that LA systems use to measure engagement and progress must be able to take into account the differences in delivery styles across modules, degree subjects, teaching teams and universities. Currently, we are using the same method to analyse student behaviour on a Level 5 programming module at Keele and will be producing classification models in the form of decision trees that will indicate the likely trajectory of a learner, given their activity on the VLE. In the first instance this will allow us to determine how generalisable our method is across different modules and institutions. The longer term hope however is that such models will help us to identify early on those students that can be supported further and offer that support to them.

## References

(1) In effect, the intranet is a bespoke VLE

(2) i.e. the more a student attends, the higher the mark they achieve

(3) As part of regular teaching activity

Arnold, K. E. & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. Proceedings of the 2nd International Conference on Learning Analytics & Knowledge. New York: ACM.

de Quincey, E., Stoneham, R. (2015) Student Engagement: An evaluation of the effectiveness of explicit and implicit Learning Analytics. In Proceedings of ePIC2014, the 12th International ePortfolio and Identity Conference. ADPIOS, Poitiers, France.

Essa, A. and Hanan. A. (2012) Improving student success using predictive models and data visualisations, Research in Learning Technology, [S.l.], v. 20, aug. 2012. ISSN 2156-7077.

Macfadyen, L.P. & Dawson, S. (2010) Mining LMS Data to Develop an ‘Early Warning System’ for Educators: A Proof of Concept, Computers & Education, vol. 54, no. 2, pp. 588–599.

MacQueen, J.B. (1967) Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

JISC (n.d.) Effective learning analytics Helping further and higher education organisations to analyse and understand their data. [online] Available at: < https://www.jisc.ac.uk/rd/projects/effective-learning-analytics > [Accessed 8th July 2016]

Johnson, L., Adams Becker, S., Cummins,M., Estrada,V., Freeman,A., and Ludgate, H. (2013) NMC Horizon Report: 2013 Higher Education Edition, Austin, Texas:TheNew MediaConsortium.

Macfadyen, L.P. & Dawson, S. (2010) Mining LMS Data to Develop an ‘Early Warning System’ for Educators: A Proof of Concept, Computers & Education, vol. 54, no. 2, pp. 588–599.

Park, K.H. & Kerr, P. (1990) Determinants of Academic Performance: A Multinomial Logit Approach. Journal of Economic Education 21: 101-111.

Perry, K. (2014) New IT system will identify failing university students. [online] Available at: < http://www.telegraph.co.uk/education/educationnews/10805717/New-IT-system-will-identify-failing-university-students.html > [Accessed 9th June 2014 ].

Ryan, M., Delaney, L., & Harmon, C. (2010). Does lecture attendance matter for grades?: Evidence from longitudinal tracking of Irish students. Belfield: University College Dublin.

Schmidt, R.M. (1983) Who maximises what? A study of student time allocation, The American Economic Review, 73:23-28.

Sclater, N., Peasgood, A. & Mullan, J. (2016) Learning Analytics in Higher Education: A review of UK and international practice. JISC, Bristol.

Siemens, G. & Long, P. (2011) Penetrating the Fog: Analytics in Learning and Education, Educause Review 46 (5), 30-32

Stoneham, R. (2012) Managed Learning Environments in Universities: Are they Achievable?, Compass (the Teaching and Learning Journal of the University of Greenwich), vol. 6, pp. 45-54.

(Footnotes)

[1] Each lecture had 2 pdf versions of the slides “handout” and “full”.