Trends in Digital Education:
Selected papers from EC-TEL 2015 Workshops CHANGEE, WAPLA, and HybridEd

Exploring the effectiveness of video viewing in an introductory x-MOOC of algebra

Joan Triay[1], Julià Minguillón[2], Teresa Sancho-Vinuesa[2], Vanesa Daza[1]

[1]Universitat Pompeu Fabra, Tànger, 122-140. 08018 Barcelona, Spain
[2]Universitat Oberta de Catalunya, Rambla del Poblenou, 156. 08018 Barcelona, Spain
juanfrancisco.triay01@estudiant.upf.edu
{jminguillona, tsancho}@uoc.edu vanesa.daza@upf.edu

Abstract. The huge amount of gathered data in a MOOC allows providing professors and course managers with insightful information about real course usage and consumption. The main aim of this work is to explore how efficient the video viewing is for completing and passing the first course offered by UCATx.cat platform, "Decoding Algebra", in order to improve its design and resources. The statistical method used is the principal component analysis but using polychoric correlation matrix between the binary variables involved in each group. The main result suggests that the participants' behavior is polarized in two extremes: they view all videos and pass de course or, on the contrary, they do not watch any one and they do not pass the test either. This information can be used by course managers to provide learners with better strategies for achieving their learning goals.

Keywords: video, learning analytics, course design, MOOCs.

1 Introduction

MOOCs (Massive Open Online Courses) have just started shaking higher education in a global scale. Now it is feasible to reach courses from top universities worldwide in a free and open way, threatening both the traditional and online higher education systems. These courses are supported by web-based learning management systems that keep track of all the navigation and interaction between course participants and the course elements (resources, activities, etc.). As thousands of participants take part in these courses, the large amount of gathered data make very interesting to analyze such courses from a participant perspective, providing teachers and course managers with insightful information about real course usage and consumption. Several recent works have been tackled the effectiveness in MOOCs through the analysis of these data ([1], [2], [3]).

In this sense, and quoting George Siemens, "Learning Analytics is the use of intelligent data, learner-produced data, and analysis models to discover information and social connections for predicting and advising people's learning" [4]. Learning Analytics can be used to better understand how participants in an online course learn as

well as to help them to achieve their learning goals, while improving the course each edition by detecting bottlenecks regarding teaching plan or interaction among participants and even misplaced or unused course elements.

In Europe, the main stakeholders in higher education have slowly started moving towards adapting the initial MOOC phenomenon in order to meet the educational needs in a more diverse, flexible, open and transversal way. Initiatives such as Future Learn in UK, Iversity in Germany, FUN in France, MiriadaX and UCATx in Spain show how massive online education evolves by both targeting complementary markets and strengthening the internal higher education systems building joint strategies [5].

Within the Catalan Programme UCATx, the Platform UCATx.cat[1], based on open edX, has been developed. The first MOOC in this platform, was named "Descodificando Algebra" (in English, "Decoding Algebra"). From the very beginning, the main aim of the course [6] was to take advantage of this new educational format to fill the gap between High School and University regarding basic notions of Algebra. At the same time, the course must remain appealing to students who do not fill this profile (in transition between school and university). Decoding Algebra was designed in such a way that despite its global outreach, it also allows prospective students of engineering or science to tackle first year Linear Algebra competently. To capture the students' interest, concepts from cryptography and coding theory were introduced.

The main aim of this work is to explore how to analyze the data of the "Decoding Algebra" in order to improve several aspects of its design and resources. In particular, we are interested in exploring how efficient is the video consumption (i.e. viewing) for completing and passing the course.

The rest of the paper will be structured as follows: In Section 2 we describe the main issues of the course, Section 3 is devoted to how analyze data of the course, while results obtained are showed in Section 4. Finally we conclude in Section 5, pointing out some future lines.

2 Description of the course

In a nutshell, the course aims to introduce some basic algebraic concepts. Problems related to communications (cryptography and coding theory) are used as a motivating factor.

The course is structured in 5 different modules/lessons spanning 5 weeks, with a weekly average dedication of 3 to 5 hours. Each module is about a different topic. Topics covered are: number sets (structure and properties), basics of modular arithmetic, matrices and polynomials, introduction to vector spaces and finally, complex numbers.

With the exception of the first module, at the beginning of all the other ones, it is introduced what it is called a challenge, which is basically a simple real problem related to the theory of communication stated in a challenge-style. We refer to those

[1] UCATX.cat platform: www.ucatx.cat

videos stating challenges as challenge videos. Thus, each module allows to go through enough mathematical notions to understand the solution to the challenge at the end of each week. Each challenge can be formulated in mathematical terms, so by applying the concepts of each module, students should be able to understand the proposed solution (we refer to these videos as challenge resolution videos) or even solve it by themselves.

All modules share the same structure. All of them have an ordered set of videos (we refer to them as conceptual videos) where the concepts of each module are developed. These concepts are accompanied by numerous illustrative examples. The duration of the videos ranges from 5 to 15 minutes, with 10 minutes being the average. Each one covers a single idea/concept so that students can watch it as many times as necessary to understand it before moving on to the next one.

There is still a final type of videos that should be taken into account, those ones that contain the resolution of the exercises proposed in the conceptual videos. We will refer to them as exercise resolution videos.

At the end of each module, students take a quiz consisting of 8 or 10 questions. The main objective is the self-evaluation of each student, so they can check if they understand the main concepts proposed in the videos that make up the module. Feedback is provided for each of the questions and, when a wrong answer is provided, the student is referred to the particular section/video of the course the students needs to work on. To pass the course, students are expected to obtain a 50% mark on each module. Following this structure, the MOOC assumes an individual participant activity and minimal interaction with both the professor and other colleagues.

Regarding the data gathered by the UCATx platform, during six weeks between 25th of August 2014 and 5th of October 2014, around 400000 events were generated for a total of 194 course participants. Use 10-point type for the name(s) of the author(s) and 9-point type for the address(es) and the abstract. For the main text, please use 10-point type and single-line spacing. We recommend the use of Computer Modern Roman or Times. Italic type may be used to emphasize words in running text. Bold type and underlining should be avoided.

3 Analysis methodology: learning analytics

Learning analytics is the methodology used to answer questions that we cannot be solved in a fairly straightforward way. What is the efficacy of video for passing a MOOC? What is the weekly connection pattern of students? Students who participate more on forum are those students who pass the MOOC? These are some of the questions we might try to answer with the aid of learning analytics. In this paper we focus on analyzing the relationship between video consumption and evaluation, using data derived from Decoding Algebra MOOC in UCATx platform.

Several years of research [7] have shown that using video in education can impact on teaching and learning and provide some benefits: increasing motivation [8] and as a necessary tool in a flipped classroom model[2].

[2] http://www.uq.edu.au/tediteach/flipped-classroom/index.html

With respect to data, Learning Analytics methodology can be divided in four different phases: data collection, data pre-processing, data analysis and data visualization. In this work we describe the first three stages, as follows.

3.1 Main entry barrier: still a very novel approach

UCATx platform automatically collects all data generated by the students when interacting with the course and span a log file in JSON format that registers all participants' activities, ranging from the enrollment action to the final MOOC action. JSON format is a type of text format for structured information based on key-value pairs. This file may contain from several hundreds of thousands lines of information up to several millions. Each line describes an event3. Each event has different fields of information, such as username, time, IP address, session and event type, among others. There is a special field named context that is very important as it contains specific information about the event and, depending on this, it may take different values. For instance, when a participant presses the pause button when watching a video, "context" is used to store the exact time when it occurs. The access to data structured in this way simplifies further processing and analysis.

3.2 Data pre-processing

In this second phase, we developed some scripts in Python language. Python was chosen because of all the functionalities that offer to interact with .JSON files as well as to extract the data from the log file. Our main goal is to obtain a "plain" structured file that describes the activity of each student of the course by means of aggregating and summarizing all the interaction available for each one of them. By plain we mean that we have the same information (columns) for each course participant (rows), that is, there are no missing fields or different length.

In this paper, we focus on those lines of the log file related to the interaction with the videos. These lines correspond to four events, namely: *play_video*, *stop_video*, *seek_video*, and *pause_video*. We also extract those lines based on the grades of the first course module, corresponding to the event called problem_check, in order to establish the relationship with the previous ones.

The plain structured file is built as follows. The result of the execution of the Python scripts, one for each event part of the analysis, is a set of new files, each one of them corresponds to a variable which values are the data that we want to extract from original log data according to such event or group of events. These variables can be a vector (containing a variable number of values, i.e. all the activity around a given video) or simple indicators, mostly numeric or binary. Finally, we join all these files into another one, that is, a matrix where each row contains the data of a course participant; and each column or a set of columns is a variable (corresponding with the different events we want to analyze). Once this process is finished, we can proceed with analyzing this structured file with a statistical package. For instance, if there are M videos in the course, obviously not all the N course participants watch the M videos; this process creates an N x M matrix containing a binary variable describing whether participant i (1…N) has seen video j (1…M) or not.

Concretely, with all this information, we created a variable called Videos Module One (VM1) composed by a vector of all videos used in such module (23 out of 98 course videos). The elements of this vector correspond to the videos related to the first module, containing binary values, either 1 or 0, according to whether the student has seen the video or not. To keep the information of those students passing the first module, we created a variable called Pass First Module (PFM). This variable is another matrix with only one column, that is, the result of taking the maximum grade of the three attempts available for the first module evaluation test. The minimum grade is 0 and the maximum grade is 8, because this module has only 8 questions. If the student did not take the test, we specify it by -1.

3.3 Data analysis

In this phase we processed the plain file obtained in the previous phase with a statistical package, namely R. As mentioned before, this file contains data from 194 course participants related to the consumption of the 23 videos used in the first course module and the final students' mark. According to course syllabus, students pass the test if their final grade is at least 4. To describe the result in the first test we generate a binary variable PFM (1 PASS, 0 FAIL). Notice that we have 23 binary variables (VM1_1 … VM1_23, one for each video) and only 194 samples, which is not a good ratio for prediction purposes.

Therefore, we need to explore how to reduce the number of variables according to the characteristics of each video in order to reduce dimensionality, and being able to compare categories, instead of individual videos, as well as analyzing the relative importance of each video within each category. For doing so, we classified the videos of module 1 in two different ways. First, according to topic, they were classified into 4 different categories: natural numbers (4 videos), integer numbers (13), rational numbers (3), and real-complex numbers (3). On the other hand, we classify them according to their activity type. Therefore, videos were classified as theory or conceptual (12 videos) and exercise videos (11). We will proceed as follows:

a) Create an indicator G1, G2, G3 and G4 for each one of the four groups. As we are just exploring the nature of the gathered data, we will use principal component analysis for summarizing how course participants consume the videos within a group.
b) Create two more indicators, GT and GE, for theory and exercise videos, respectively, using also PCA with the same goal.
c) Build two different generalized linear models (logistic), M1 and M2, one for each group of indicators abovementioned, trying to predict whether a student passes the first module test or not with respect to such group of videos.

4 Results

As we mentioned in Section 3, we compute a component summarizing the consumption of the videos for each group. We use principal component analysis but using the polychoric correlation matrix between the binary variables involved in each group. Table 1 shows the percentage of variance explained by the first component, which is reasonable for all of them. Furthermore, all these components also show a very interesting behavior: they have a large kurtosis, which means that most of the distribution mass is not centered on the mean and it follows a quite asymmetrical distribution. Notice that the maximum value is larger than the minimum one (in absolute value) but for G1. Table 2 shows the weights for each variable taking part in the component.

Table 1. Explained variance and range for each computed component.

Group	Number of Videos	Explained variance	Range
G1- Natural Numbers	4	44.1 %	[-1.43,0.89]
G2- Integer Numbers	13	39.8 %	[-0.93,1.41]
G3- Rational Numbers	3	45.0 %	[-0.79,1.47]
G4- Real/Complex Numbers	3	41.2 %	[-0.90,1.32]
GT- Theory	12	40.4 %	[-1.07,1.30]
GE-Exercises	11	47.0 %	[-0.93,1.53]

Table 2. Relative video weights for each computed component.

Group	Weights
G1	[0.390, 0.624, 0.694, 0.861]
G2	[0.454, 0.512, 0.444, 0.464, 0.553, 0.617, 0.605, 0.722, 0.651, 0.666, 0.773, 0.810, 0.774]
G3	[0.532, 0.805, 0.647] G4 [0.773, 0.518, 0.584]
GT	[0.299, 0.421, 0.471, 0.506, 0.557, 0.591, 0.663, 0.734, 0.719, 0.819, 0.850, 0.825]
GE	[0.399, 0.356, 0.576, 0.566, 0.651, 0.662, 0.793, 0.788, 0.808, 0.857, 0.856]

Using these components, we build two different generalized linear models, one for explaining the importance of each topic (G1 … G4) and another one to explain the importance of each kind of video (GT and GE), with respect to attempting (and passing) the first test of the course. In order to obtain positive β coefficients for all components, we force a Varimax rotation, so we can compare only magnitudes.

Table 3 shows the computed logistic model that tries to predict whether a student will attempt (and pass) the test according to the videos the student has viewed. This model has a (pseudo) R2 of 0.668, quite high. Notice that we are not trying to generalize these results, so we are only interested in the magnitudes of the β coefficients. As the intercept is negative (so students not watching videos or only a few are predicted to not pass the test), it is necessary to have large values in one or more components in order to pass the test.

Table 3. Generalized linear model for predicting which is the most important group of videos.

	Coef β	S.E.	Wald Z	Pr(>|Z|)
Intercept	-0.4346	0.2334	-1.86	0.0625
G1	-0.1877	0.3320	-0.57	0.5719
G2	1.0140	0.4708	2.15	0.0313
G3	0.2392	0.4792	0.50	0.6176
G4	1.2122	0.4189	2.89	0.0038

Table 4. Generalized linear model for predicting which kind of videos are most important.

	Coef β	S.E.	Wald Z	Pr(>|Z|)
Intercept	-0.4442	0.2259	-1.97	0.0492
GT	0.9925	0.4625	2.15	0.0319
GE	1.2267	0.4885	2.51	0.0120

5 Discussion

In the light of these results, and taking into account the exploratory nature of the analysis, we can draw some interesting conclusions about how course participants are consuming the videos.

First, the computed components summarizing the consumption of videos for each group show that most course participants watch all the videos within each group. The distribution of each component, once normalized (MEAN = 0, SD = 1), shows that the majority of students either do nothing or do everything, taking almost always extreme values of the range in Table 2.

In fact, for each group in Table 2, we can observe that the weights increase. This means that the more videos they watch, the better results they obtain. Therefore, those students that see all the videos accumulate more knowledge. This fact happens both for groups by topic (except perhaps the artificial group real / complex) and for the theory and exercises groups. It is also remarkable that within each topic, exercise videos have larger weights than theory videos, in general.

Table 3 shows that both G1 and G3 are irrelevant, since the beta coefficient multiplied by the maximum values of its range (the positive one) does not allow the model to predict who will succeed with the test. However, G2 and G4 are indeed relevant. In fact, that G1 is negative it may be caused by the fact that what it is really important is G2 (as natural numbers are just briefly presented compared to integers), so G1 consumption is subsumed by those students who see the videos in G2. This could be stated as if you "study" integers will make understand natural numbers. Moreover, perhaps the first model in Table 3 shows only that the exam is biased towards a particular type of exercise.

Finally, Table 4 also shows that whenever both theory and exercise videos are watched jointly, the chances to pass the test increase. It is important to remark that both must be watched, since the weight of each block is similar. Given the distribution of these components, it is necessary to do both things. Otherwise, weights are cancelled out and the model does not predict passing the test.

In summary, even a preliminary exploratory analysis can be very helpful for determining if course participants are using the proposed resources (i.e. videos) as expected. Principal component analysis combined with logistic regression can be used to determine how videos are watched, the relative importance of each video within a group and the relative importance of each group of videos with respect to the evaluation test. In fact, evaluation itself can be analyzed to detect whether course participants are skipping parts of the course or not, as well as test biases towards some topics rather than others.

Acknowledgements

This work was supported by research grants from Generalitat of Catalonia (2014 SGR 1271) and the interuniversitary programme UCATx.

References

1. Muñoz-Merino, P.J., Ruipérez-Valiente, J.A., Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado Kloos, C.: Precise Effectiveness Strategy for analyzing the effectiveness of students with educational resources and activities in MOOCs. Computers in Human Behavior, vol. 47, 108--118 (2015)
2. Milligan, S.: Crowd-sourced learning in MOOCs: learning analytics meets measurement theory. In: Fifth International Conference on Learning Analytics And Knowledge LAK'15, pp. 151--155 (2015)
3. Whitmer, J., Schiorring, E., James, P., Miley, S.: How Students Engage with a Remedial English Writing MOOC: A Case Study in Learning Analytics with Big Data. Educause Learning Initiative (2015)
4. Siemens, George. "What Are Learning Analytics?" Elearnspace, August 25, 2010. http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/
5. Sancho-Vinuesa, T., Oliver, M., Gisbert Cervera, M.: Moocs en cataluña: un instrumento para la innovación en educación superior. Educación XX1: Revista de la Facultad de Educación, vol. 18, 2, 125--146 (2015)
6. Daza, V., Rovira, C., Makriyannis, N. MOOC attack: closing the gap between pre- university and university mathematics. Open Learning: The Journal of Open, Distance and e-Learning, vol. 28, 3, 227--238 (2014)
7. Clearance Center: Video Use and Higher Education: Options for the Future. Technical report (2009)
8. Bravo, E., Amante, B., Simo, P., Enache, M., Fernandez, V. Video as a new teaching tool to increase student motivation. In: Global Engineering Education Conference (EDUCON), 2011 IEEE (2011)