Mercado-Varela, M. A., García-Holgado, A., García-Peñalvo, F. J., & Ramírez-Monoya, M. S. (2016). Analyzing navigation logs in MOOC: A case study. In F. J. García-Peñalvo (Ed.), *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'16) (Salamanca, Spain, November 2-4, 2016)* (pp. 873-880). New York, NY, USA: ACM. doi:10.1145/3012430.3012620

# Analyzing navigation logs in MOOC: A case study

Martín Alonso Mercado-Varela
Instituto de Investigación y Desarrollo Educativo
Universidad Autónoma de Baja California
Ensenada, México
martin_mercado44@hotmail.com

Alicia García-Holgado
GRIAL Research Group,
University of Salamanca
Paseo de Canalejas, 169
37008 Salamanca, Spain
+34 923294500 3433
aliciagh@usal.es

Francisco J. García-Peñalvo
GRIAL Research Group,
University of Salamanca
Paseo de Canalejas, 169
37008 Salamanca, Spain
+34 923294500 3433
Distinguished Professor of the Escuela de Humanidades y Educación, Tecnológico de Monterrey, México
fgarcia@usal.es

María Soledad Ramírez-Montoya
Escuela de Educación, Humanidades y Ciencias Sociales,
Tecnológico de Monterrey
Monterrey, Mexico
solramirez@itesm.mx



## ABSTRACT
Continued use of various technological devices has massively increased the generation of digital data, which are recorded as an opportunity for research. In the educational case, it is common to analyze data generated in Learning Management Systems which allows better understand the learning process of the participants and make informed decisions for better e-learning processes and situations in which develop. This paper analyzes participants' navigation logs in a MOOC hosted on the Coursera platform, for which a visual e-learning analytics process was performed. The results confirm that the videos of experts are an essential educational resource for learning in a MOOC, similarly, the discussion forums are an important resource which are recurrent social spaces in different navigation paths complementing other activities.

## CCS Concepts
• Information systems~Data analytics • Applied computing~Education • Applied computing~Interactive learning environments

## Keywords
Learning analytics; statistical analysis; MOOC; Coursera; log analysis.

## 1. INTRODUCTION
The increased use of Information and Communication Technology (ICT) in e-learning processes have provided opportunities to increase knowledge about the forms of learning in online environments. This has been heightened by the massive amount of data generated and which are mainly available when entering in massive learning environments as MOOC [11]. Data or logs produced by a student inside a virtual platform are important because of they allow get information about what happens in their learning processes [1; 12] as well as improving situations in which e-learning takes place on. The processes by which knowledge of these data is extracted is known as Data Mining [5]; when these processes are applied in education the Educational Data Mining (EDM) arises. The International Educational Data Mining Society defines the EDM as a discipline comprising methods to explore unique data types from educational contexts, which are used to better understand students and to learning contexts.

The development of the MOOC has generated research opportunities in the field of learning analytics [2; 6; 9; 13; 14]. Explain and reduce the high dropout rates presented by participants [12; 14] as well as the amount of digital data that is generated as an opportunity to investigate [2; 9] are factors for learning analytics use in the MOOC field. The arrangement of MOOC platforms (e.g. Coursera) to support research activities of its partner institutions through sharing the data produced by the participants, such as forums posts or the navigation logs, ensures continuity in the development of this discipline. We live in the era of big data, abundance of data as defined by McAfee et al. [7], is one of the factors that will shape the future of higher education [12], hence the importance of analyzing such data.

This paper describes the analysis of participants' navigation logs in a massive learning environment hosted on the Coursera platform. The "Educational Innovation with Open Resources" course developed by the Tecnológico de Monterrey (Mexico) is the learning environment under analysis. Analyzed logs are part of the second instance, which ran from September 1, 2014 to September 28, 2014.

Instructional proposals for massive courses integrated into their formats different types of components such as videos, peer review, discussion forums and social networks. The course also integrated Open Educational Resources (OER) in different formats (readings, videos, podcast, software, learning objects). It is considered the course as a hybrid MOOC (cMOOC + xMOOC), because it has a centralized structure to a platform (xMOOC) but it encourages self-learning and creating learning communities, sharing of intellectual work and open access to materials (cMOOC). Beyond pedagogy, the course is classified in accordance with 12 dimensions identified by Conole [3] that give best example of the nature of the course (See Table 1).

The analysis aims to identify teaching resources and most used socialization spaces, it is a first approach to the navigation patterns and factors associated with learning in a particular MOOC. The analysis presented in this study is considered exploratory. Moreover, this work is part of a thesis about the development of skills to mobilize OER in MOOC format, where different practices in the construction of knowledge, navigation patterns and factors associated with learning are described. The results presented in this paper are only part of a larger work still developing, which follows a mixed methodology, where quantitative results are exploratory and qualitative results have the function to expand and deepen.

**Table 1. Course classification. Adapted from Conole [3]**

| Dimension | Low | Medium | High |
|---|---|---|---|
| Open | | | X |
| Massive | | | X |
| Use of multimedia | | X | |
| Degree of communication | | | X |
| Degree of collaboration | | | X |
| Learning pathway | | | X |
| Quality assurance | | | X |
| Amount of reflection | | | X |
| Certification | X | | |
| Formal learning | | X | |
| Autonomy | | | X |
| Diversity | | X | |

The paper has been divided into five parts. The first part gives a description of the methodology. The second part explains the data pre-processing. The second part describes the sample to analyze. Fourth part presents the data analysis and the results. Finally, the last part closes the paper and provides the primary conclusions of this research.

## 2. METHODOLOGY

To carry out the analysis of participants' navigation within the MOOC we have adapted the visual e-learning analytics process (VeLA) proposed by Gómez-Aguilar et al. [4]. The VeLA model Figure 1 provides a framework to process the data provided by Coursera in order to prepare the information for extending the visual component in future works.

First we have transformed the dataset in order to prepare them to apply analytics models, particularly, statistical analysis. In parallel with data analysis, we have mapped a portion of the data to visualize navigation patterns between course resources pairs. The results have provided feedback to post-processing the dataset and repeat the process in order to discover knowledge. The following sections describe each phase after several iterations.



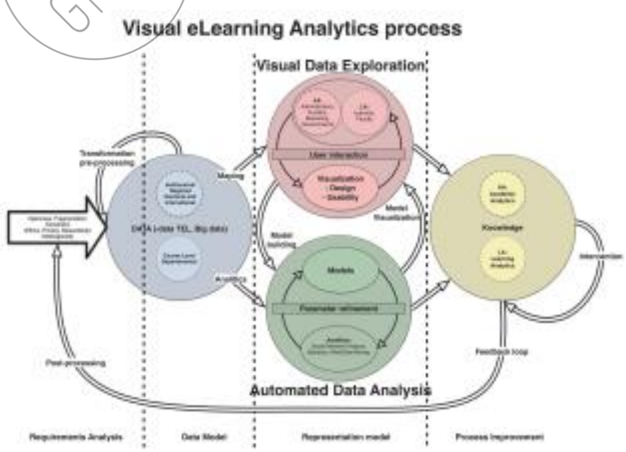


**Figure 1. VeLA process [4]**

## 3. DATA PRE-PROCESSING

Coursera is a large platform that keeps track of all students and staff activities details as they pertain to hosted courses [8]. For the purpose of this study, we have limited our analysis to data about participants' navigation behavior provided by the stream of click events they generated on the course webpages.

The used dataset is a file provided by Coursera that record information about click events that occur during the participants' visit to the course. This record is called clickstream log. Each click event is represented through a set of variables [10]: *key*, the type of user action; *value*, stores information related to the interaction that cannot be stored in the other variables such as metadata values about video playing; *username*; *timestamp*, a standard UTC timestamp containing the number of milliseconds since the UNIX epoch; *page_url*, the URL of the page on which the click occur; *client*; *session*, a cookie-stored value used to track individual client machines, it is not an individual browsing session; *language*, ISO language code that indicates the language to serve the site in; *from*, the URL of the page that the user navigated to the current page provided by page_url; *user_ip*; *user_agent*; *12*, the width and height of the screen used by the user; *13*, it indicates whether the interaction is performed through a GET or POST request; *14*, indicates the page from which the user initially navigated to class.coursera.org; *30*, UNIX timestamp in milliseconds in which the user initiated the interaction from the page indicated in from.

Clickstream log use the JSON format (JavaScript Object Notation), a standard format for exchanging data between applications (http://www.json.org). Each click event is represented by an individual serialized JSON array.

To perform statistical analysis of user click events it was necessary to process the file. This process is carried out in three cycles.

First we were analyzed the different variables contained in the clickstream log in order to delete those that did not provide relevant information for further analysis. The selected variables were:

- *Client* because it always takes the same value, 'spark'.
- *User_ip*, *user_agent* and *12* because the data analysis does not consider issues geolocation devices or use by the user.
- *Value*, it remains only when key variable indicates that the event is related to a video, in other cases it not provide relevant information, only duplicates the value of *page_url*.

The second cycle has focused on dump the entire contents of the file in a relational database in MySQL. The database is composed by two tables (Figure 2), the main table whose columns are the variables selected in the previous cycle except value (keyaction table), and a table related with the previous to store the metadata contained in value variable when the type of action taken by the user is viewing a video (videovalue table).

Some of the variables have been renamed to facilitate tasks in the next cycle, so that *type* corresponds to the *key* variable, *page_from* is *from*, *post_event* is 13, *initial_referrer* is 14 and *timestamp_end* is 30. Also included the variable *kaid* as the primary key of the table in the context of relational databases.


**Figure 2. Database scheme to save clickstream log of Coursera**

The third cycle relative to data pre-processing has enabled infer the derived variables from generic variables presented in previous cycles. To facilitate this process and support dynamic query about different participants, we have developed a web tool, (https://github.com/aliciagh/courseraparser) Coursera Parser, that provides a web form to indicate the anonymous identifier of one or more participants and provide the results in a CSV file separated by semicolons where each line represents a user and each column a derived variable.

The derived variables are:

- Clicks: total number of interactions that the user made within the course. It has been calculated counting the number of records associated with the user in the clickstream log.
- Week1: number of interactions that the user made within the course during the first week, from September 1, 2014 to September 7, 2014 inclusive. It has been calculated counting the number of records associated with the user in the time period indicated.
- Week2: number of interactions that the user made within the course during the second week, from September 8, 2014 to September 14, 2014 inclusive. It has been calculated counting the number of records associated with the user in the time period indicated.
- Week3: number of interactions that the user made within the course during the third week, from September 15, 2014 to September 21, 2014 inclusive. It has been calculated counting the number of records associated with the user in the time period indicated.

- Week4: number of interactions that the user made within the course during the fourth week, from September 22, 2014 to September 28, 2014 inclusive. It has been calculated counting the number of records associated with the user in the time period indicated.
- Clicks_homepage: total number of clicks that the user performs on the course home page.
- Clicks_auth: total number of clicks that the user performs on the login pages.
- Clicks_forum: total number of clicks that the user performs on the forums.
- Clicks_human_grading: total number of clicks that the user performs on peer assessment pages.
- Clicks_lecture: total number of clicks that the user performs on videos.
- Clicks_quiz: total number of clicks that the user performs on self-assessments.
- Clicks_wiki: total number of clicks that the user performs on pages that compose the course space.
- Clicks_search: total number of clicks that the user performs on search pages.
- Clicks_other: total number of clicks that the user performs on other course elements.
- Different_videos: number of different videos that the user viewed within the course.
- Different_selfassessments: number of different self-assessments that the user viewed or replied within the course.
- Total_different_days: total number of different days that the user made some interaction in the course.
- Days_between_first_last_connection: days since the user made the first interaction to the last recordered.
- First_connection: date on which the first user interaction occurred in the course.
- Last_connection: date on which the last user interaction occurred in the course.
- Total_real_seconds: total time in seconds that the user interacted with the course from the first interaction to the last recorded. It is calculated from the timestamps of each click, so it is considered that two clicks are consecutive when there is a difference between them less than one hour in order to cover those connections where videos are viewed or contents are read. The variable is transformed in format of day, hour, minute and second (dd:hh:mm:ss).
- Total_real_seconds_per_week: total time in seconds that the user interacted with the course each week (four variables).
- Total_sessions: number of sessions that the user made. There is no variable in the clickstream log that provides information about users' sessions so it has been considered that a new session starts when the time between two clicks is more than one hour.
- Total_sessions_per_week: number of sessions that the user made within the course during each week (four variables).
- Different_days_after_september: total number of different days that the user made some interaction in the course after the official course dates, September 28, 2014.
- Real_seconds_after_september: total time in seconds that the user interacted with the course after the official course dates.
- Total_sessions_after_september: number of sessions that the user made after the official course dates.

In order to calculate the number of clicks on the different types of existing resources within the course, it has been required performing a set of transformation in the URLs provided by *page_url* and *page_from* variables. These variables store the complete URL; before the transformation there were around 21,000 different URLs. This number is reduced to 9 different elements, each of which represents a different resource within the MOOC. The tool parses the URLs and distinguishes resources by the paths and parameters that have been used to access them.

First, the URL is divided in components following this scheme: [scheme]://[domain]/[path]?[variables]#[tags]

Then, we differentiate two types of URLs depending on the value of the domain, those refer to pages or resources within Coursera and those that do not. The URLs whose domain does not match with Coursera (www.coursera.org or class.coursera.org) are reduced to "/external". This allows identifying those clicks coming from outside the course environment. Otherwise, clicks that are made from the course to external resources cannot be recorded because the clickstream log does not register this kind of events.

The other URLs are reduced to the value of the path and processed to obtain the smallest possible number of resources, so that does not differentiate between videos or specific forums, but fall into a single element. These elements are: *human_grading*, which corresponds to the peer assessment pages; *lecture*, which covers the videos that compose the MOOC; *quiz*, self-assessments available each week of the course; *forum*; *wiki*, which represents the set of pages with the course contents; *auth*, access pages or login to the course space; *search*, pages related to searches doing inside the course; *other*, for those internal course links that do not represent a significant resource.

Finally, the web tool provides other CSV file with information about the navigation between resources of one or more participants. The file has three columns: *source*, which represents the resource where the navigation begins; *target*, which represents the resource where the navigation ends; and *weight*, that indicates the number of times that the navigation between two resources occurred.

## 4. SAMPLE SELECTION

In this paper, we analyze data obtained from Coursera for the second instance of the course "Educational Innovation with Open Resources". Coursera indicates that a total of 14,226 students enrolled for the course, of which 6,696 visited the course all time, 5,106 students committed to finish, and 1,197 submitted at least one exercise.

**Table 2. General information about participants**

| General information | Statistical data |
|---|---|
| **Profession** | |
| Teacher | 141 (49.3%) |

| | |
|---|---|
| Researcher | 64 (22.3%) |
| Student | 14 (4.9%) |
| Trainer | 9 (3.1%) |
| Academic management | 15 (5.2%) |
| Administrative | 11 (3.8) |
| Librarian | 5 (1.7%) |
| Businessman | 6 (2.1%) |
| Other profession | 21 (7.3%) |
| **Educational degree** | |
| **Baccalaureate** | 8 (2.8%) |
| **Technical career** | 15 (5.2%) |
| Bachelor's degree | 106 (37.1%) |
| Master's degree | 117 (40.9%) |
| Doctor's degree | 32 (11.2%) |
| Post-doctoral | 3 (1.0%) |
| **Teaching experience** | |
| Hybrid mode (face to face-virtual) | 97 (33.9%) |
| Face mode | 149 (52.1%) |
| Virtual mode | 13 (4.5%) |
| **Experience as a participant** | |
| Hybrid mode (face to face-virtual) | 128 (44.8%) |
| Face-to-face mode | 88 (30.8%) |
| Virtual mode | 45 (15.7%) |
| MOOC | 20 (7.0%) |



We recognize the diversity of participants who join the MOOC, however, for research, are only considered active participants, those have made different types of activities over the course. This imply that they were more exposed to the instruction of the course and its components; therefore, the analysis of their logs can provide a bigger picture of navigation within the course. The 286 participants included in the sample are mostly teachers, which have a higher educational level, more experience as teachers and participants in the hybrid and face-to-face modalities, but they have little experience in the MOOC mode. Table 2 shows general information about the participants' profiles.

## 5. ANALYSIS AND RESULTS

The analysis and presentation of the results is descriptive. The statistics for each variable are total, mean, standard deviation, minimum and maximum data. Although the data pre-processing provided 33 derived variables, this analysis is focused on variables related to clicks, connections and time periods:

- Total number of clicks: total number of interactions that the user made within the course. The variable is counted in each week of the course, to describe the trend over time.
- Number of connections: connections that the user made in the course. The variable is counted in each week of the course, to describe the trend over time.
- Connection time: total time in seconds that the user interacted with the course from the first interaction to the last registered. The variable is counted in each week of the course, to describe the trend over time.
- Total clicks on each course activity modules: number of clicks on each of teaching resources or social spaces that integrated the course.
- Post in forums: number of comments made by participants in the discussion forums.

- Forums comments: number of comments on the post made by participants in the discussion forums.

Also, other data were obtained by applying instruments that were defined within the course platform, with the help of the databases obtained from the Coursera dashboard to perform analyzes. The analyzed variables correspond to:

- Working connections: feedback from participants about work connections established with other participants and learning storage in non-human devices.
- Learning connections: feedback from participants about individual and collaborative learning, through forums and self-study groups.
- Socialization of knowledge: feedback from participants about the number of comments made on digital evidence of learning (portfolios).

The level of achievement (approved and unapproved) according to the instructions of course, is determined by the delivery of learning portfolios (four in total, 20 points each) evaluated by the pairs method (three participants evaluated) and the answer to the self-assessments (four in total, 5 points each); for these two activities, they must meet certain conditions, such as analyzing videos and readings and other interactions with the components of the course. Thus, the analysis focuses on knowing if a particular component (videos, discussion forums, peer evaluation and self-assessment), the participation (connections and time connections) or the socialization (working connections, learning connections, socialization of knowledge, post in forums and forum comments) is associated with the condition of being approved or not.

Mann-Whitney U test was applied to determine whether there were significant differences between groups using the level of achievement variable. This nonparametric test was used because the variables did not meet the criterion of normality. Within the course three levels of achievement were obtained:

- Approved with distinction (≥ 90 points)
- Normally approved (between 60 and 90 points)
- Unapproved (< 60 points)

It was decided to work only with approved with distinction and unapproved participants. This decision was made to polarize groups of participants (avoiding effects of the mean), although this involves data loss, allows work with two different groups.

They can be seen in Table 3 where the total connections after the course was 14,797, with a standard deviation of 81,108, total clicks 222.895, with an average of 779.35 and a standard deviation of 454,361, suggesting differences among participants. In addition, connections beyond the time limits of course, in the months of October and November were recorded.

**Table 3. Descriptions of variables associated with participation**

| | Total | Mean | Standard deviation | Min | Max |
|---|---|---|---|---|---|
| Clicks | 222,895 | 779.35 | 454.361 | 43 | 3052 |
| Connections | 14,797 | 51.74 | 81.108 | 2 | 972 |
| Time connection | 360:05:32:37 | 1:06:26:31 | 1:16:17:31.9 | 00:27:51 | 18:06:17:40 |
| Connection on different days | 6,471 | 22.63 | 6.596 | 2 | 57 |
| First and last connection | Among the first three days of September it registered the first connection (86.8%)<br>In October, 80.3 percentage of the last connections registered even when the course ended the last day of September; additional connections were recorded until late November although with lower percentages. | | | | |

Table 4 shows the modules in which the activities of the participants were grouped. The most consulted module was the video experts with 91.640 clicks, followed by the contents page with 38.123 clicks. Highlight there are participants with very low level of activity in the different modules, for example in discussion forums (0) or peer evaluation (1) which are spaces of socialization and where most activity would be expected by any participant.

The largest number of interactions occurred in the first week of the course with a total of 62,921 clicks (Table 5). A progressive decrease is observed in the number of registered interactions throughout the course.

**Table 4. Descriptions of variables associated with different activity modules**

| | Total | Mean | Standard deviation | Min | Max |
|---|---|---|---|---|---|
| Expert video | 91,640 | 320.42 | 288.616 | 2 | 2405 |
| Discussion forums | 34,901 | 122.03 | 177.897 | 0 | 1393 |
| Peer evaluation | 36,692 | 128.29 | 64.655 | 1 | 361 |
| Self-assessment | 8,270 | 28.92 | 10.567 | 12 | 74 |
| Homepage | 12,754 | 44.59 | 40.322 | 3 | 442 |
| Contents page | 38,123 | 133.30 | 61.426 | 19 | 414 |
| Search page | 80 | .28 | 1.641 | 0 | 24 |

**Table 5. Description of participation per week**

| | Clicks | | | Connections | | Time connections | |
|---|---|---|---|---|---|---|---|
| **Week** | **Total** | **Mean** | **Standard deviation** | **Total** | **Mean** | **Total** | **Mean** |
| 1 | 62,921 | 220.00 | 180.984 | 3,191 | 22.28 | 106:16:09:37 | 17:53:09 |
| 2 | 61,041 | 213.43 | 153.427 | 3,914 | 27,34 | 97:09:15:59 | 16:19:16 |
| 3 | 51,569 | 180.31 | 130.135 | 3,363 | 23,50 | 82:23:37:01 | 13:54:36 |
| 4 | 36,007 | 125.90 | 92.787 | 2,715 | 18,95 | 59:16:12:35 | 09:59:50 |

Regarding participants' navigation paths, we have visualized the trails between course resources pairs with a directed graph with weights. The higher weights (number of repetitions of pattern) are from the home page and the page content to each of the different resources: self-assessment, forums and peer evaluation; this is because they are the starting points to enter the course. However, they are more illustrative of knowledge construction practices the paths between resources and social spaces, so there are four paths with heavy weights. Forums displayed in three of the four identified paths, so it is considered important socialization space that accompanies other activities, as a first or subsequent activity.

(1) Discussion forums →Peer evaluation (weight=1740)

(2) Peer evaluation → Discussion forums (weight=1114)

(3) Self-assessment → Peer evaluation (weight =863)

(4) Self-assessment → Discussion forums (weight =200)

According to the Mann-Whitney U test only variables associated with socialization were significant (See Table 6). Observing the mean rank, there is a tendency that favors the approved participants, who had higher intensity in the variables.

**Table 6. Descriptions of variables associated with level of achievement**

| | Level of achievement | Mean rank | Rank-sum | Significance |
|---|---|---|---|---|
| Expert video | Approved | 145.05 | 31621.50 | p=.57 |
| | Unapproved | 138.52 | 9419.50 | |
| Discussion forums | Approved | 140.44 | 30616.50 | p=.26 |
| | Unapproved | 153.30 | 10424.50 | |
| Peer evaluation | Approved | 138.66 | 30228.00 | p=.07 |
| | Unapproved | 159.01 | 10813.00 | |
| Self-assessment | Approved | 142.46 | 31057.00 | p=.74 |
| | Unapproved | 146.82 | 9984.00 | |
| Connections | Approved | 144.04 | 31400.00 | p=.84 |
| | Unapproved | 141.78 | 9641.00 | |
| Time connections | Approved | 141.76 | 30620.00 | p=.78 |
| | Unapproved | 144.85 | 9850.00 | |
| Working connections | Approved | 146.68 | 31829.50 | p=.02 |
| | Unapproved | 124.21 | 8073.50 | |
| Learning connections | Approved | 145.75 | 31336.00 | p=.04 |
| | Unapproved | 125.53 | 8285.00 | |
| Socialization of knowledge | Approved | 151.46 | 32716.00 | p=.00 |
| | Unapproved | 106.23 | 6905.00 | |
| Post in forums | Approved | 157.77 | 34394.50 | p=.00 |
| | Unapproved | 97.74 | 6646.50 | |
| Forums comments | Approved | 155.34 | 33864.00 | p=.00 |
| | Unapproved | 105.54 | 7177.00 | |



## 6. CONCLUSIONS

The VeLA process has allowed to meet the research objectives. However, we recognize that further research to deepen the results is necessary because we can develop different types of visualizations in order to discover more complex navigation patterns and include the navigation logs of all MOOC participants. On the other hand, the lack of navigations log that lead to resources and spaces outside of the platform is a learning analytics problem when an online learning environment is analyzed. Thus, it is required exploring different ways to include those datasets.

Regarding the results, these confirms that videos are an essential resource in the learning proposals of a MOOC. In the analyzed course, a group of experts were invited to share their opinion, so it is considered that they were a resource that promoted reflection and contextualization of the topics in professional practice of the participants. In this context, after analysis, it verified that the videos have been the most consulted resources. About participants' navigation paths, the forums were a recurrent space in the identified paths; this implies that the forums were an important resource to carry out the learning activities. Also, participant's socialization is associated with the condition of being approved. This confirms that socialization is very important in a MOOC.

The variables related to the components of the course and participation were not related to the level of achievement. Thus, the intensity of these variables are not necessarily predispose the condition of being approved. This could be explained by the type of sample selection, where participants from both groups are not representative of its population, also the selection criteria "active participation" can agglomerate participants with the same characteristics regardless if they are approved or not. In addition, define a participant as approved or unapproved according with the criteria of the course does not imply a complete achievement, at least in learning terms. Measuring achievement in a MOOC course is complex, since the number of participants represents a huge diversity of specific objectives, so the achievement depends only on goals that the participant has set for himself and not on a label imposed. Therefore, it is necessary also, qualitative research that take into account the participants' expectations and satisfaction on completion of training, describing in detail both personal conditions and the training itself, for which expectations are met.

## 7. ACKNOWLEDGMENTS

The authors wish to thank the Research and Innovation Group in Education (GIIE) at Tecnológico de Monterrey (Mexico) http://sitios.itesm.mx/eehcs/iie/ and the Research Group in InterAction and eLearning (GRIAL) of the University of Salamanca (Spain) http://grial.usal.es for their contributions and support.

This research work has been carried out within the University of Salamanca PhD Programme on Education in the Knowledge Society scope (http://knowledgesociety.usal.es) and was supported by the Spanish Ministerio de Educación, Cultura y Deporte under a FPU fellowship (FPU014/04783).

## 8. REFERENCES

[1] Baker, R.S., 2014. Educational Data Mining: An Advance for Intelligent Systems in Education. *IEEE Intelligent systems 29*, 3, 78-82.

[2] Breslow, L., Pritchard, D.E., Deboer, J., Stump, G.S., Ho, A.D., and Seaton, D.T., 2013. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment 8*.

[3] Conole, G., 2013. MOOCs as disruptive technologies: strategies for enhancing the learner experience and quality of MOOCs. *Revista de Educación a Distancia*, 39.

[4] Gómez-Aguilar, D.A., García-Peñalvo, F.J., and Therón, R., 2014. Analítica visual en e-learning. *El profesional de la información 23*, 3, 1386-6710.

[5] Hand, D.J., Mannila, H., and Smyth, P., 2001. *Principles of data mining*. MIT press.

[6] Hollands, F.M. and Tirthali, D., 2014. *MOOCs: Expectations and Reality. Full report*. Columbia University, Center for Benefit-Cost Studies of Education, Teachers College, New York.

[7] Mcafee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., and Barton, D., 2012. Big data. *The management revolution. Harvard Bus Rev 90*, 10, 61-67.

[8] Mukala, P., Buijs, J.C.a.M., and Van Der Aalst, W.M.P., 2015. *Exploring students' learning behaviour in moocs using process mining techniques*. Eindhoven University of Technology, BPM Center Report.

[9] O'reilly, U.M. and Veeramachaneni, K., 2014. Technology for Mining the Big Data of MOOCs. *Research & Practice in Assessment 9*, 29-37.

[10] Online, R., 2014. Overview Of Coursera Data and Analytics RICE University.

[11] Sangrà, A., González-Sanmamed, M., and Anderson, T., 2015. Meta-analysis of research on MOOC along the period 2013-2014. *Educación XX1 18*, 2, 21-49. DOI= http://dx.doi.org/10.5944/educXX1.13463.

[12] Long, P.D. and Siemens, G., 2011. Penetrating the Fog: Analytics in Learning and Education. *Educause Review 46*, 5, 30-32.

[13] Valverde Berrocoso, J., 2014. MOOC: una visión crítica desde las ciencias de la educación. *Profesorado: Revista de curriculum y formación del profesorado 18*, 1, 93-111.

[14] Vázquez-Cano, E., 2015. El reto tecnológico para la sostenibilidad de los massive open online course (MOOC). *Panorama 9*, 17, 51-60.