Accepted Manuscript

Analysis of student behavior in learning management systems through a big data framework

Magdalena Cantabella, Raquel Martínez-España, Belén Ayuso, Juan Antonio Yáñez, Andrés Muñoz

PII: S0167-739X(17)32921-7
DOI: https://doi.org/10.1016/j.future.2018.08.003
Reference: FUTURE 4387

To appear in: *Future Generation Computer Systems*

Received date : 20 December 2017
Revised date : 14 July 2018
Accepted date : 2 August 2018

Please cite this article as:, Analysis of student behavior in learning management systems through a big data framework, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.08.003

Analysis of student behavior in Learning Management Systems through a Big Data framework

Magdalena Cantabella, Raquel Martínez-España[1], Belén Ayuso, Juan Antonio Yáñez, Andrés Muñoz

mmcantabella@ucam.edu, rmartinez@ucam.edu, bayuso@ucam.edu, juan.yanez@alu.ucam.edu, amunoz@ucam.edu

Department of Computer Science, Universidad Católica de Murcia, Murcia, Spain. ZIP 30007

Abstract

In recent years, learning management systems (LMSs) have played a fundamental role in higher education teaching models. A new line of research has been opened relating to the analysis of student behavior within an LMS, in the search for patterns that improve the learning process. Current e-learning platforms allow for recording student activity, thereby enabling the exploration of events generated in the use of LMS tools. This paper presents a case study conducted at the Catholic University of Murcia, where student behavior in the past four academic years was analyzed according to learning modality (that is, on-campus, online, and blended), considering the number of accesses to the LMS, tools employed by students and their associated events. Given the difficulty of managing the large volume of data generated by users in the LMS (up to 70 GB in this study), statistical and association rule techniques were performed using a Big Data framework, thus speeding up the statistical analysis of the data. The obtained results are demonstrated using visual analytic techniques, and evaluated in order to detect trends and deficiencies in the use of the LMS by students.

Keywords: Learning management systems, MapReduce, Apriori algorithm, E-learning analytics, Student behavior, Big Data

[1]Corresponding author

1. Introduction

A current tendency in higher education consists of the analysis and processing of data relating to the activity generated by users through the use of learning management systems (LMSs). The significant amount of data extracted from these platforms provide fundamental information that can aid both teachers and students in improving their educational goals. One of the main problems at present is the analysis of this information, owing to two main factors: the already mentioned large volume of data available, and the different formats of these data, particularly for the management of unstructured data.

According to several studies (see, for example, [1, 2]), there exists a need for analytical tools to help to interpret LMS data and provide new knowledge for improving and even designing new e-learning techniques and methodologies. Before manipulating such information, it is also important to explore and select the necessary data from the LMS, according to the goals to be achieved.

The main objective of this work is to design and implement a framework based on big data technologies to identify the behavior patterns of LMS users and illustrate them in an intuitive and intelligible manner. For this purpose, we define the following steps:

- Data preprocessing, by studying the data to be extracted from the LMS and its storage in a big data platform.
- Data analysis and identification of pattern recognition techniques that may provide value in the educational context.
- Presentation of the obtained results according to suitable visual analytics techniques and tools.

For developing these steps, we have considered data processing guided by e-learning analytics, in which the connections among educational techniques, learning concepts and educational data mining are studied [3, 4]. Within this

field, the areas most relevant to our work are learning analytics and visual analytics. The former aids us in data processing for discovering connections among students, teachers and the learning process, with the purpose of creating recommendations that improve the overall educational process. The latter uses visual interfaces to illustrate the results obtained from analytical reasoning, facilitating an understanding of the new knowledge and aiding the users in discovering new relations or possible irregularities [5]. Here, we take a further step forward in the use of e-learning analytics by integrating big data techniques into the educational data analysis. In this manner, both the trends and deficiencies in e-learning methodologies can be detected by analytical techniques applied to large volumes of data.

We propose an exploration and analysis of the LMS data extracted from user events generated during four complete academic years in all courses for the three learning modalities (namely on-campus, online and blended) available at our university, amounting to 70 GB of data. The aim of this proposal is to evaluate whether the results obtained by applying a big data framework to these LMS data aid in detecting tendencies and anomalies in the use of these platforms in any learning modality.

This study was performed at the Catholic University of Murcia (UCAM). Several on-campus degrees have been offered since 1996, and within the past five years, the university has consolidated its training offer with several degrees in online and blended modalities. The Sakai LMS[2] is used as a resource management and collaborative platform for all of the training modalities.

The remainder of this paper is structured as follows. Section 2 reviews several previous works relating to the analysis of educational data. Section 3 explains our general framework proposal based on big data technologies in order to analyze student data. Section 4 provides the results of our proposal from analyzing 70 GB of Sakai LMS data gathered during four academic years. Finally, conclusions and future work are outlined in section 5.

[2] https://sakaiproject.org/

2. Related Work

The inclusion of an LMS as an essential methodological tool in higher education is the standard at present [6], generating new needs and fields of study to aid in the design of novel learning models through the knowledge obtained from the LMS data. LMSs provide a large volume of data, while also generating the need for intelligent tools integrated within the LMS that aid in their interpretation and provide feedback of this information. A hot topic in this field is the identification of user behavior patterns with the use of data mining techniques, which is known as educational data mining [7]. The identification of user behavior patterns is aimed at developing new teaching methodologies that aid and improve both student and teacher performance by means of analyzing the data provided by the LMS and other tools such as surveys.

Romero et al. [8] performed a theoretical study exploring the application of data mining to the use of Moodle LMS. Their purpose was to provide guidance in initiating this discipline. Details of the main data mining techniques for e-learning were provided, and these were compared with a practical case evaluated in Moodle. Similarly, in [9], Moodle was used as an LMS to analyze the integration of data mining techniques with data warehouse tools and online analytical processing. The authors performed a classification of the activities that aid in improving the performance and the results of students enrolled in e-learning modalities.

In [10], the necessary requirements for integrating data mining services into the Sharable Content Object Reference Model (SCORM) (see [11] for details on this standard) compliant platforms were reviewed. The authors proposed analyzing the records based on Web server access logs to obtain student behavior. These records can provide massive volumes of data representing click-stream or click-flow data. The paper is concluded by stating that the data obtained from logs are very limited and do not provide the necessary requirements for generating the required information. A different procedure was illustrated in [12], where the authors took an innovative approach to course evaluation through

data mining techniques. They analyzed the learning behavior of students in the K-12 level through their activity records in the LMS, along with demographic data and end-of-course assessment surveys. The use of multiple data forms allows for more meaningful analysis of student behavior and identifies possible relationships.

It is well known that the data produced by LMSs have increased considerably in recent years. Therefore, the current analysis techniques for LMS data must evolve and adapt to the new challenges faced by higher education institutions. A recommended solution is the use of big data in e-learning as an emerging discipline. The need for the evolution from educational data mining to educational big data is a reality, as stated elsewhere [13, 14]. Big data offers us the opportunity to reach a higher level in the use of LMS, obtaining increased benefits from student experience by making decisions based on the strategic responses obtained from big data results. Thus, it is possible to convert complex, unstructured data into actionable information, thereby aiding in identifying useful data and transforming it into valuable information for higher education institutions [15, 16].

West [17] and Picciano [18] described an initial outline for the use of big data techniques in an educational context, but these works did not study specific techniques or methods in detail. Their purpose was to examine the evolving world of big data and analytics in higher education by means of LMSs. The two studies coincided in the expected benefits that will aid in determining new pedagogical techniques. This may represent a great advance in decision making and educational strategies, allowing for the analysis of large amounts of data and offering the possibility to extract further knowledge.

Furthermore, we found interesting studies that have demonstrated the benefit of applying big data techniques in higher education, such as [19], in which the student learning patterns were searched based on data extracted from forum tools integrated in massive open online courses (MOOCs). In the aforementioned work, an information model based on big data, known as topic-oriented learning assistance, was developed, which provides a ranking of forums in on-

line courses. In this manner, forum topics can be classified automatically, so that lecturers can make specific comments depending on the classification, while students can quickly find the required content. Likewise, the work in [20] presented the SAP HANA, a big data-based analysis and monitoring tool that implements a scoring system for students in the LMS. The score represents the student participation in learning activities, and low scores usually imply poor student achievements. Finally, [21] demonstrated that interactions with learning environments can be modeled and measured effectively. The authors defined an IoT-based interaction framework and analyzed the student experience of electronic e-learning. The framework evaluated the behavior of students attending videoconferences by measuring their level of attention based on face and eye observation, using an attention-scoring model.

Our work extends this research line on the adoption of big data for analyzing LMS data. We present a framework based on big data technologies to analyze large volumes of data from events generated in the use of every learning tool available in the Sakai LMS for the three main training modalities: on-campus, online and blended.

3. Framework based on big data for analyzing Sakai data

This section describes our proposal for a framework based on big data technologies, aimed at analyzing the Sakai data. It consists of three stages: data acquisition and storage, data analysis, and visualization of results.

3.1. Data acquisition and storage

The goal of this stage is to study the original working dataset stored in the Sakai LMS and extract it to a big data storage platform. Sakai data are stored in a relational database containing more than 100 tables. Following a deep study related to the relevance of such tables with respect to user behavior patterns, the following three tables were found to include the most important information for our study: Sakai_User, Sakai_Session and Sakai_Event. The

first contains basic information regarding Sakai users: id, name, email, and role (for example, student or instructor). Sakai_Session stores information regarding user logins on the platform: user id, and session start and end dates. Finally, Sakai_Event stores data of user events while using the platform. Among these data are session id, event id, event date, and context (the course in which the event occurred). Relationships exist among these three tables, so it is possible to query session and event data for any user.

Once the data sources have been selected, the data are anonymized in order to protect personal information such as names and emails. Next, the data need to be transferred from the Sakai database to big data storage. To this end, a big data solution based on Azure HDInsight[3] has been adopted, using its Hadoop distributed file system (HDFS) implementation. The tool used to transfer data from the Sakai database is Sqoop[4]. These data are stored in a Hive [22] data warehouse owing to their analytical features, as explained in section 3.2. Figure 1 illustrates the architectural schema for the data acquisition and storage steps, and the technologies involved.

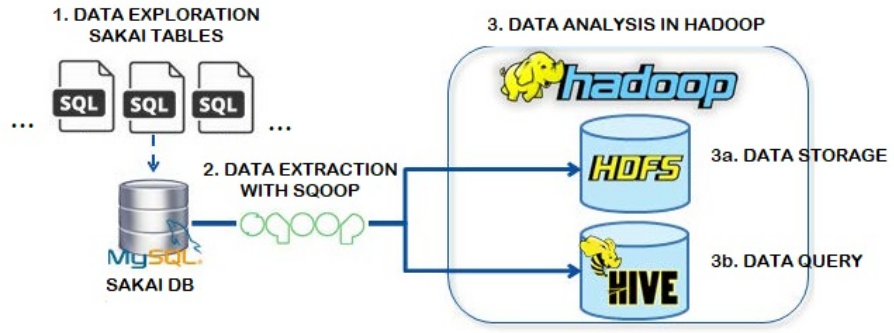


Figure 1: Big data architecture for acquiring and storing Sakai data.

The Azure cluster deployed for analyzing the Sakai data uses Hadoop v2.7.3, running on an Ubuntu 16.04 kernel. The cluster is composed of two head nodes

[3] https://azure.microsoft.com/en-us/services/hdinsight/
[4] http://sqoop.apache.org/

(four cores and 28 GB RAM each) and four worker nodes (eight cores and 28 GB RAM each).

3.2. Data analysis

After storing the data and before starting to obtain conclusions from them, we need to study these data with the purpose of identifying the most important aspects to consider. It should be noted that the data may contain noise or irrelevant features that could affect the results. In this situation, we have two options: we can perform data preprocessing in order to remove all unnecessary features, or we can employ a technique that can work properly with noise and irrelevant features. In this work, we design and implement several techniques, based on big data features, which can deal with noise and irrelevant data without affecting the quality of the results.

As stated in the previous section, Hive was selected for storing the student data, as the use of this data warehouse system within Hadoop enables us to apply different techniques, such as HiveQL queries or statistical analysis.

We firstly perform a quantitative analysis using HiveQL, which is the ad-hoc query system for Hive. With this analysis, we are able to calculate the following information items.

- Tool ranking: This analysis studies the tools that are used more by each student in each session. A session is defined as the time lapse for which a student is connected to the e-learning platform. For each tool, we analyze the most-used events per tool. Moreover, during this process, we analyze the correlation between events in a session. This correlation will be used and analyzed together with the Apriori algorithm. The purpose of this correlation process is to determine which events are related during the same session, in order to determine student behavior patterns. Therefore, the Pearson correlation coefficient [23] is calculated by means of a Hive function.
- Event ranking: This process analyzes the events carried out by a student

in each course, in order to identify not only the most frequent events in each course, but also the absence of certain events. It is aimed at detecting courses with high or low activity, along with a global ranking of events occurring in each training modality. Thus, this query may provide certain insights into the actions performed by each student in a specific course/modality as well as the possible lack of actions that may be relevant to the student training modality (for example, repeatedly not attending videoconferences in the online modality).

- Event trends: The intention of this query is to analyze the timeline related to events of interest in the e-learning platform (for example, an event for connection to Sakai), in order to identify certain significant cyclical patterns. By using the *time series analysis* technique offered by Hive, it is possible to identify periods with high or low activity in the e-learning platform.

- Connection trends: This process performs a statistical study to analyze the monthly and weekly trends in the connections to the LMS, and the mean number of visits to the LMS by students grouped according to academic course and training modality. This information may aid in detecting differences in the number of accesses to the LMS for a specific day of the week, month, year, and modality.

By using the information obtained in the previous steps, we need to define a technique for analyzing the associations and sequences among events. The desirable features for this technique are interpretability, robustness, and speed. Thus, after studying different possibilities and taking into account the amount of data, we selected the Apriori algorithm [24]. This is one of the most popular and widely used algorithms in both data mining and educational data [13, 25, 26, 27]. These literature references used the Apriori algorithm within the educational data field; in this case, we search for a pattern-seeking approach to analyzing student behavior.

The Apriori algorithm is an association rule data mining technique that can be implemented in a distributed and parallel manner [28]. Its robustness and interpretability make it possible to obtain reliable results that can be interpreted by non-technical personnel. This is one of the main reasons for its selection, with the added value of the possibility of implementing it in MapReduce in order to manipulate large amounts of data.

This association rule technique attempts to determine associations among items or frequent patterns in datasets. In order to parallelize the algorithm and be able to work with large amounts of data, we implemented this technique by following the Hadoop MapReduce framework [29]. This framework avoids the problem of grid computing, where potential opportunities always exist for a node to fail, and the task must therefore be executed again. In particular, we have implemented a version of the Apriori algorithm taking into account the characteristics of the educational context. This technique receives a set of items (in *attribute-value* format) as input and returns a set of association rules (item-rules). The Apriori technique is composed of two phases, as follows.

- In the first phase, the technique counts the frequency of each item and then the frequency of different item combinations. A combination of items (item-rules) that exceeds a certain threshold will be taken into account for the final item-rules set. This threshold is known as *support* and is calculated as the number of item repetitions divided by the number of transactions, with a transaction being an entry in the dataset containing different items. It is necessary to establish a minimum support to eliminate the less frequent item-rules.

- In the second phase, a set of item-rules is generated from the item sets occurring more frequently and exceeding a confidence threshold. The confidence is the conditional probability that a transaction containing item X also contains item Z.

These two phases are demonstrated in algorithm 1. The algorithm receives the minimum support value and D datasets to work with as input. Firstly, the

Algorithm 1: General procedure of Apriori technique

Input $Support, D$
L_1 ={Get-frequent-1-item-rules(D)}
for all ($k = 2; L_{k-1}! = \emptyset; k++$) **do**
  C_k = candidates generated from L_{k-1}
  for all ($transaction\ t \in D$) **do**
    $C_t = subset(C_k, t)$
    for all ($candidatec \in C_t$) **do**
      c.support++;
    end for
  end for
  $L_k = \{c \in C_i \mid c.support \geq Support\}$
end for
Output $\bigcup_k L_k$

frequency of all item-rules with one item is counted. Next, the algorithm uses this set to generate a subset of two elements, then a subset of three elements, etcetera, until no further subset combinations can be created. All possible pairs complying with the minimum support measures are taken. Finally, the L_k rules satisfying the confidence threshold are generated. This final step is considered as a prune, and is reflected in algorithm 1 in the code line $L_k = \{c \in C_i \mid c.support \geq Support\}$.

Algorithm 1 was adapted to the MapReduce framework in order to be executed in our Azure HDInsight configuration, as explained in section 3.1. Figure 2 depicts the first phase of the algorithm; specifically, the function *Get-frequent-1-item-rules(D)*, where counting of the most frequent items occurs. Here, 'K' represents the key and I_x represents the different item set (key, value). These values are obtained using the methods provided by the MapReduce API and the capability of HDFS and Hive to handle this algorithm in a distributed manner.

With respect to the problem addressed in this paper, an item is an LMS

event, an item-rule is a set of events, each occurring with a certain frequency, and a transaction corresponds to a student LMS session of. The input dataset is composed of all sessions and all events for all students. According to this technique, we can determine associations or recurring behaviors regarding the different events performed by students in the Sakai platform. In this paper, the Apriori technique is implemented for the proposed framework as a MapReduce process in order to take a very large number of events from numerous courses as input. Regarding the minimal support and confidence thresholds, the values

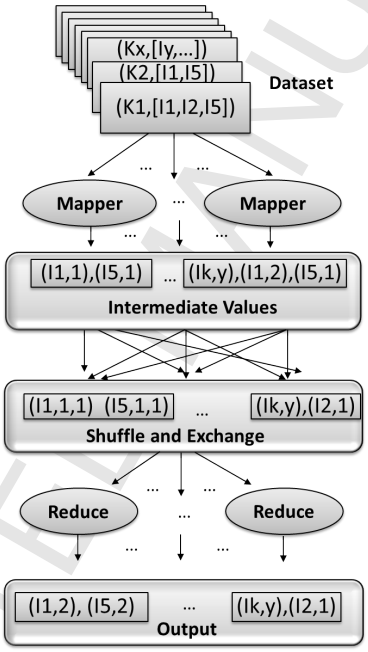


Figure 2: General scheme of function Get-frequent-1-item-rules

established for the support and confidence values are between [20% to 30%] and [80% to 70%], respectively. Following several assessment tests, we verified that the support and confidence values in these intervals obtain similar rules and conclusions. Therefore, for the study case presented in the following section, we fixed the support and confidence thresholds at 20% and 80%, respectively.

3.3. Data visualization

Data visualization, or the manner in which to illustrate the findings obtained following the data analysis phase appropriately, is an essential part of any big data-related project. The results must be presented in an intuitive and easy-to-understand manner, as they are usually discussed by people who are not data science specialists.

To this end, the tools selected in the proposed framework for the visualization of results are Tableau [30] and QlikView [31]. Tableau is one of the leading data visualization tools for charts, graphs, maps, and other visualization types. This tool allows for exporting the graphics in multiple formats and embedding the results in any web page. Moreover, Tableau manages large complex data stored in Azure, Hortonworks, MapR, and Amazon EMR distributions via Hive. We used the free version of Tableau, known as Tableau Public, for our framework. It is used to plot ranking event graphics and event correlation graphics.

QlikView is a business intelligence tool that handles large amounts of data from multiple sources, processing and presenting these in a very easy and intuitive manner. One of its main advantages is that its dashboard enables data integration in memory; therefore, it can operate while disconnected from the data source and provides very high performance. This tool is used in this study for plotting graphics related to the event timeline.

4. Case study

In this section, we present our case study, in which 70 GB of event data collected from Sakai are analyzed by means of our proposed big data framework. The data to be analyzed correspond to the Sakai events generated by all students in our university for the three learning modalities (online, on-campus and blended) during four academic years; that is, from 2012/2013 to 2015/2016. A total of 41 degrees and 93 master's degrees have been taken into account, as displayed in Table 1, grouped according to training modality. The total number of students registered during this time period amounts to 76,268, as indicated

Table 1: Number of masters and degrees by modality and areas of study.

Areas of study		Online	On-campus	Blended
Social sciences	Degree	1	6	3
	Master	6	3	2
Health	Degree	1	9	1
	Master	3	4	22
Sport	Degree		3	
	Master	2	2	12
Engineering	Degree	1	5	
	Master	3	3	4
Business	Degree	2	5	1
	Master	10	3	3
Juridical law	Degree	1	1	1
	Master	4	4	3

in Table 2. The events generated by the students during this period amount to 79,432,423, distributed by modality and academic year, as indicated in Table 3.

The following sections discuss the most relevant results and highlights of our case study, considering the most used tools and events, trends in the use of tools, including the detected use associations among the tools, and finally, log-in records are analyzed to determine potential connection patterns. In all of these sections, we have analyzed the behavior of student groups according to their training modality.

Table 2: Number of students grouped by academic year and modality.

Modality/year	2012/2013	2013/2014	2014/2015	2015/2016
Online	628	863	1526	2849
On-campus	12114	13483	15333	16960
Blended	2425	1885	3457	4745

Table 3: Number of events grouped by academic year and modality.

Modal./year	2012/2013	2013/2014	2014/2015	2015/2016
Online	335557	593423	2169552	4205723
On-campus	12410137	14689112	16420582	9745588
Blended	1467401	3637706	6373376	7384266

4.1. Tool ranking

According to Figures 3, 4, and 5, by using normalized data regarding the number of students, we highlight the following findings on the evolution during the four academic years in terms of the use of Sakai tools.

- For the online modality, as illustrated in Figure 3, there is a significant increase in the student activity for the academic year 14/15 in almost all tools. This is a direct consequence of applying regulations on teaching materials and methodologies that are obligatory for lecturers. This increase is highly significant in the use of Lesson Builder and Resource tools, and to a lesser extent but also outstanding in Assignments and Announcements. This increase is owing to the application of templates and content containers in the virtual campus, which makes the sequence of contents and activities more attractive and intuitive, particularly through the Lesson Builder's own tool, which defines the unit template. The only tool that suffers a decrease in the online modality is Forum, for academic years 13/14 and 14/15. Its justification is owing to the fact that, according to academic regulations, this tool was used for resolving doubts, many of which were not oriented with the subject contents, but rather organizational issues that, with the application of the measures discussed in the previous paragraph, have now been solved. With the aim of increasing participation in Forum, during the academic year 15/16 a new measure was established, namely the use of discussion forums in which the lecturer poses a challenge to students, which results in a considerable increase in the use of this tool.

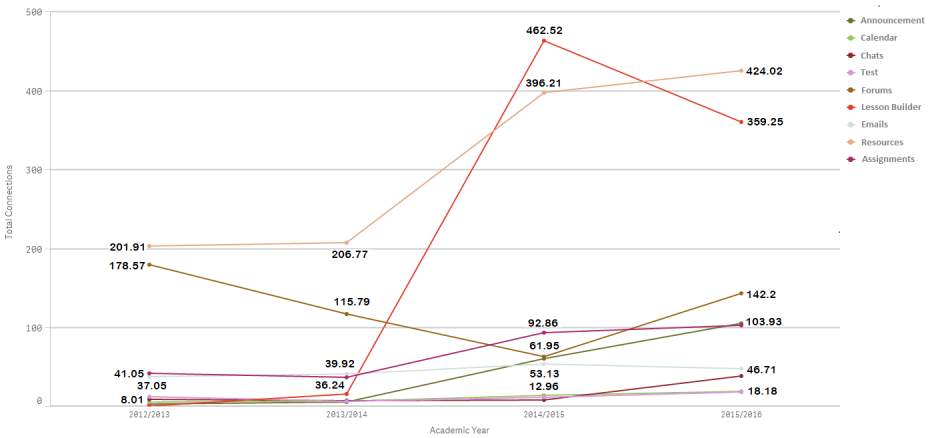


Figure 3: Evolution for academic years regarding use of Sakai tools in online modality.

- The blended modality presents similar behavior in the use of Sakai tools as for the online modality, with the exception of the Forum tool, as illustrated in Figure 4. Despite the fact that the same regulations regarding materials and templates apply in both modalities, this tool is more widely used in the blended modality. This may be attributed to the fact that such a small percentage of on-campus sessions occurring in this modality increase participation even in activities outside the classroom.

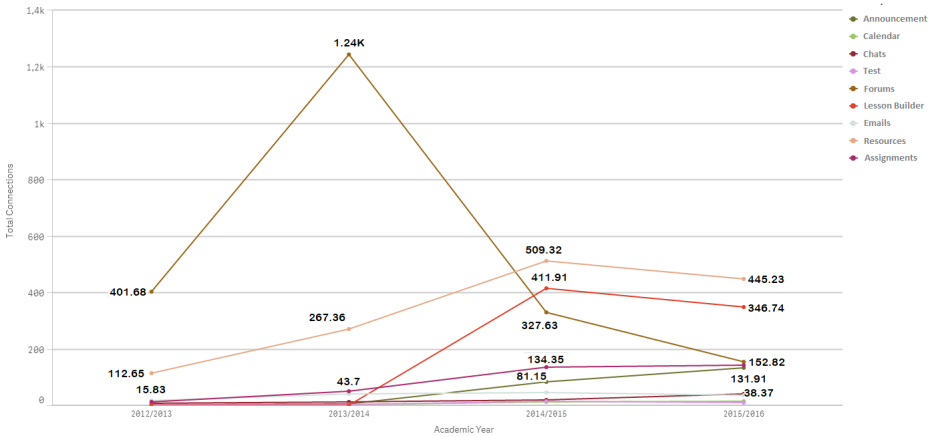


Figure 4: Evolution for academic years regarding use of Sakai tools in blended modality.

- For the on-campus modality, the majority of the tools exhibit a decreasing tendency of use, with certain exceptions, as can be observed in Figure 5. The Announcement and Assignment tools are the only ones with an increasing tendency, which is presumably motivated by the introduction of automatic notifications sent to the students' mail when certain information is published.

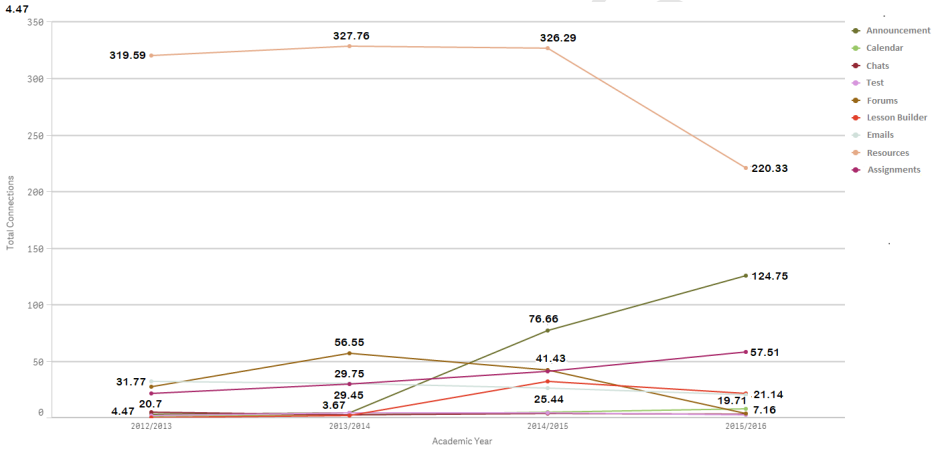


Figure 5: Evolution for academic years regarding use of Sakai tools in on-campus modality.

Comparing the normalized use of tools according to the number of students in each modality, we find that in most cases, the greatest use of tools is found in the online modality, followed by the blended modality with a similar value. There is a strong decrease in the use of the on-campus modality for the final academic year. The only tool with similar behavior in the three modalities is Announcements. Finally, the most used tool for all academic years and modalities is Resources.

We further analyze the use of tools according to the specific events generated by each Sakai tool. After studying Table A.6, as illustrated in Appendix A, we can add the following highlights related to the use of Sakai tools.

The Lesson Builder tool was not implemented in the academic years 12/13 and 13/14. In the year 13/14, it was implemented as a pilot experience in

certain grades. However, it was not until academic year 14/15 that it became operational for all degrees at the university. It can be observed that, after the Resources tool, the Lesson Builder tool and its event 'visit to unit' are the most used. This tool was very well received by the students, and the syllabuses and contents of the subjects were presented with greater clarity and ease. For the Chats tool, in the 12/13 and 13/14 courses, the students created many more chat messages than they read. In contrast, during 14/15 and 15/16 they read more chat messages than they created. Finally, the Forum tool also exhibits differences between the first and second couple of years. During the first two years, the students wrote in the forums more than they read posts. However, during the final two years, the students read more answers from their peers than they wrote.

In conclusion, we must emphasize the behavior changes in the students. In the academic courses of 12/13 and 13/14, the students were more active, creating more forums and chat messages. However, in the academic courses of 14/15 and 15/16, the students became more passive, reading more forums and chat messages as opposed to creating them. This passivity on the part of the students coincides with the implementation of the Lesson Builder tool.

As indicated in the results, the Lesson Builder tool is one of the most outstanding, being one of the most used by students. Figure 6 illustrates a screenshot of a learning unit of the web programming subject. Certain data have been intentionally blurred for the sake of privacy. Students can view the unit start and end schedule, as well as the estimated study hours. Moreover, the materials are organized into main and additional materials, and have videos, forums, and videoconference dates and times, as well as links to these.

4.2. Event rankings

In this section, we analyze the event rankings independently of their associated tools. Figures 7, 8, and 9 illustrate the top 10 event rankings grouped according to modality. The y-axis indicates the event identifier, while the x-axis shows the number of records for each event. These rankings were obtained

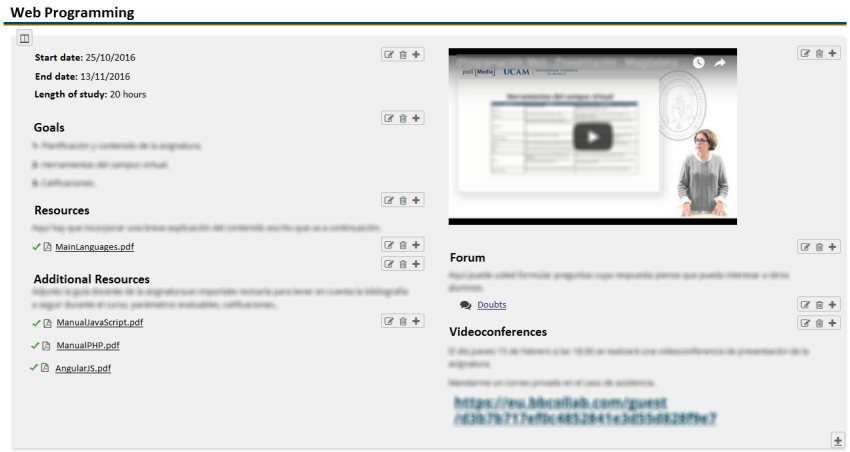


Figure 6: Example of typical organization for learning unit for Lesson Builder tool in Sakai

using the ad-hoc HiveQL queries explained in section 3.2. Comments on these rankings are provided as follows.

- Figure 7 presents the event ranking for the online modality. The most frequent events in this modality are the "Download resource" and "Visit unit (Lesson Builder)" events; moderate events are "Create resource" and "Customize site"; and finally, the events with the least activity are the "Read assignment" and "Read message" events.

- Figure 8 presents the events for the on-campus modality. The most frequent events are "Download resource" and "Read announcement"; moderate occurrences of the "Read post (Forum)" and "Update profile" events are observed; the events with the least activity are "Read message" and "Save draft (Assignment)".

- Figure 9 refers to the blended modality. The most frequently performed events are "Download resource" and "Visit unit (Lesson Builder)", while "Read assignment instructions" and "Read announcement" are considered as moderate; the events with the least activity are "Read chat" and "Customize site".

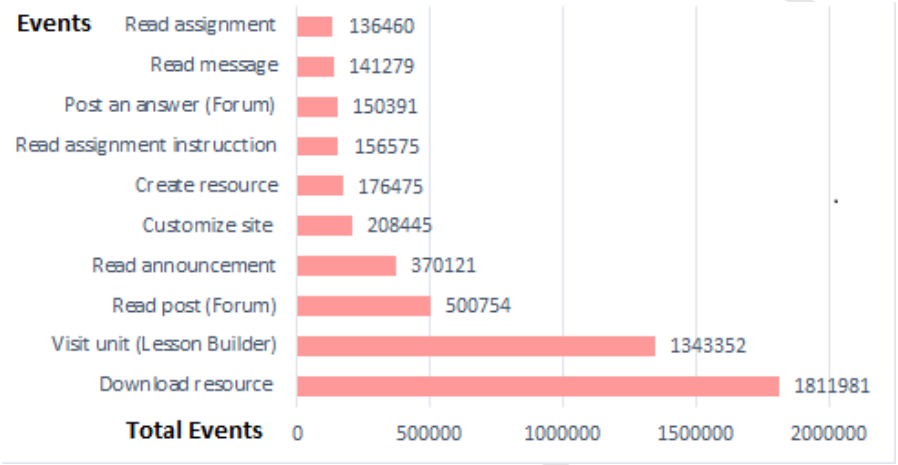


Figure 7: Event ranking for online modality

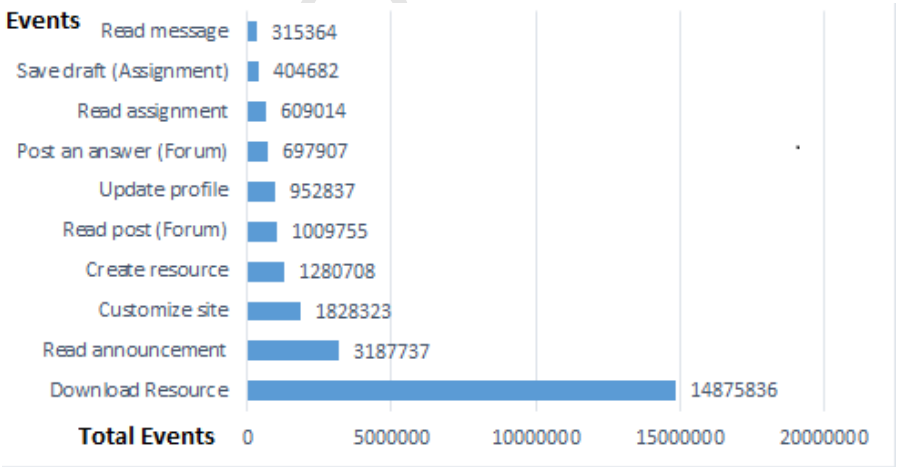


Figure 8: Event ranking for on-campus modality

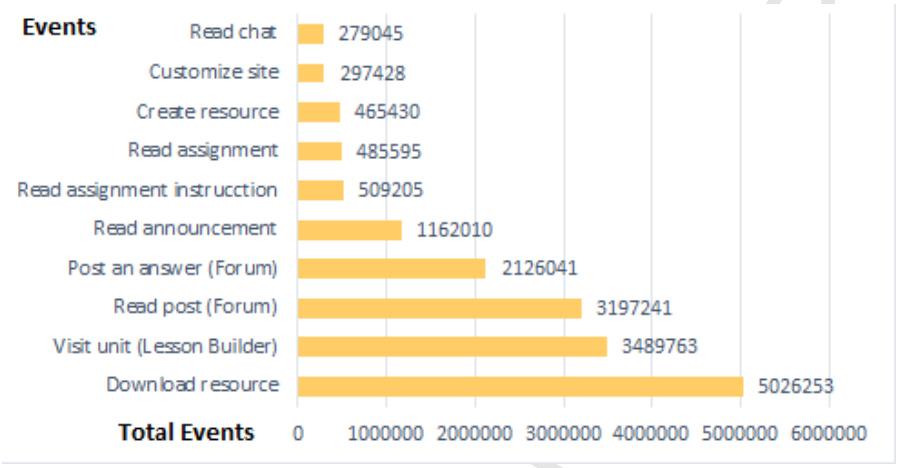


Figure 9: Event ranking for blended modality

After analyzing these rankings, we find that the event most repeated by students in the three modalities is the "Download resource" event. We must remember that Sakai is a content manager, the fundamental function of which is to provide different resource types to students, and this is confirmed by the event data gathered.

It can be observed that the online and blended modalities present the same behavior in the most used Sakai events, which is owing to the use of the Lesson Builder template as the access point to the remaining tools in both modalities. Furthermore, it should be noted that students in the online modality do not read messages from the private message tool integrated in Sakai. This fact is justified because these messages are associated with their academic emails, which they can read directly without needing to access the platform. However, the low activity for the event "Read assignment" in the online modality is surprising. This is because the instructions for each assignment are specified in an attached file. This activity corresponds to the event "Read assignment instructions", which offers further possibilities when defining assignments, such as using increased space in writing, integrating text with images, and formatting options, among others. In the remaining modalities, these instructions are provided in

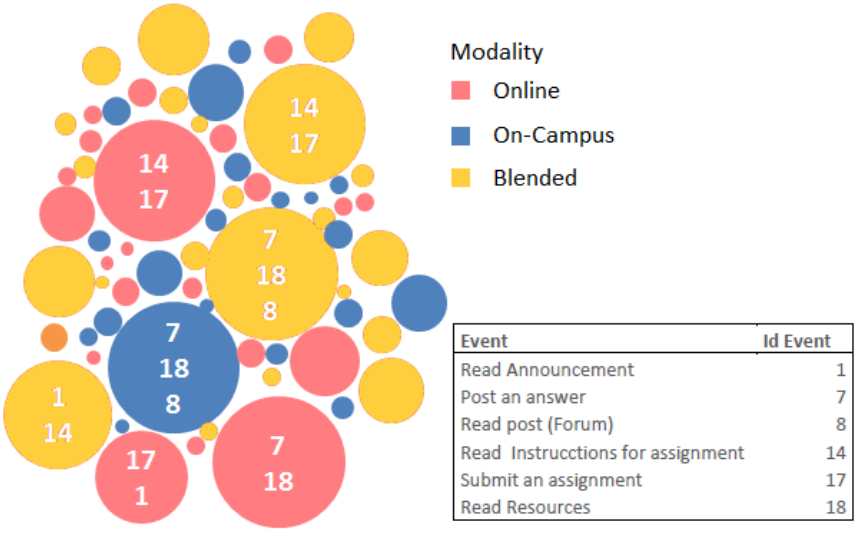
the classroom.

In the following section (section 4.3), an in-depth analysis is performed on the possible relationships among the most frequently used events.

4.3. Event relationships

By applying and HiveQL queries and Apriori algorithm presented in the previous section, we obtained several association rules that provide information regarding student behavior and the relation between the events they perform on the LMS platform. The event relationships obtained take into account the different modalities. Thus, Figure 10 uses a bubble chart to depict the probabilities for relationships among events in each modality, as obtained by the Apriori algorithm. Each modality is grouped by red, yellow, and blue colors for the online, blended, and on-campus modalities, respectively.

The analysis of the most representative associations between events performed in the same session, with a reliability index higher than 70%, is as follows.



Event	Id Event
Read Announcement	1
Post an answer	7
Read post (Forum)	8
Read Instrucctions for assignment	14
Submit an assignment	17
Read Resources	18

Figure 10: Combinations of events performed by students in same session grouped by modality

- For the online modality:
  - Students read the instruction for an assignment (event 14) before submitting it (event 17).
  - Students read an announcement (event 1) before sending an assignment (event 17), owing to the fact that lecturers provide the final indications by means of announcements.
  - Before posting an answer in the forum (event 7), students consult the necessary resource (event 18) to submit a correct answer. This behavior allows the teacher to ask more questions in the forums, so that students study indirectly when consulting resources study.
- For the on-campus modality:
  - In the same manner as online students, on-campus students access resources (event 18) before posting an answer (event 7) in the forum. However, the students also look at the resources (event 18) when they read a forum post (event 8). This behavior allows students to consult and study resources indirectly using the forums as a tool for discussion, despite being on-campus students and attending classes.
- For the blended modality:
  - Again, the students in the blended group access the resources (event 18) before reading (event 8) or posting an answer (event 7) in a forum.
  - As with online students, blended students read the instructions (event 14) before submitting an assignment (event 17). This behavior is owing to the fact that online and blended student are more meticulous and do not receive instructions in class, so they must read instructions to avoid errors in the submission of assignments.
  - Students read an announcement (event 1) and then read instructions for an assignment (event 14). This is owing to the fact that, on many

occasions, lecturers notify the activation of a pending assignment through announcements in this modality.

Further relationships among events performed in the same session, with a reliability index higher than 50%, are the following. The online modality students read an announcement and then read the instruction for an assignment. There are announcements indicating that new assignments are available, which is the reason of this relationship. For the on-campus modality, students read a post, and instead of posting an answer, they create a new thread in the forum for answering. This usually occurs when they answer the question of the lecturer, while they answer a classmate in the same thread. Finally, for the blended modality, students read an announcement and visit the unit in the Lesson Builder tool before submitting an assignment.

In addition to these relationships among events, it is important to note that on-campus students do not follow a behavior pattern when they access the platform. However, the usage pattern determined by the association rules differs between the online and blended modalities on the e-learning platform. When online students access the LMS platform, they first visit a unit using the Lesson Builder tool, then read the instructions for an assignment, finish reading the announcements, and finally submit the assignment. However, once blended students have accessed the platform, they start reading the instructions for an assignment, then visit a unit using the Lesson Builder tool to submit the assignment later. In conclusion, forums are a tool that students always use, and the pattern determined indicates that they must be fomented by lecturers because this provides a means for students to study, consult resources, and be made aware of assignments.

4.4. Connection trends

In this section, we search for connection trends according to the number of accesses to Sakai for each modality during all of the analyzed years. In order to achieve this, we analyzed the trends with respect to the number of student

connections to Sakai annually and weekly, grouped by modality, and the mean number of visits for each year and modality.

Firstly, Figure 11 illustrates the tendency with respect to the number of Sakai visits made by students during the period studied (namely, from 2012/13 to 2015/16). It can be observed that the period with higher activity corresponds to January and February, coinciding with the mid-term exam period at our university. Note that, for the final exam period of May and June, there is less activity compared to the mid-term period. We identified three main reasons for this difference: firstly, subjects requiring a more individual and autonomous workload and less contact with lecturers (for example, degree or master's final projects) are scheduled in the second term; secondly, subjects with external practicums (particularly in health and education degrees) are mostly placed in the second term; and finally, to a lesser extent, the number of drop-out students also has an effect on the number of visits. Moreover, in Figure 11, the course opening months of September to November and holiday term months of July to August are identifiable.

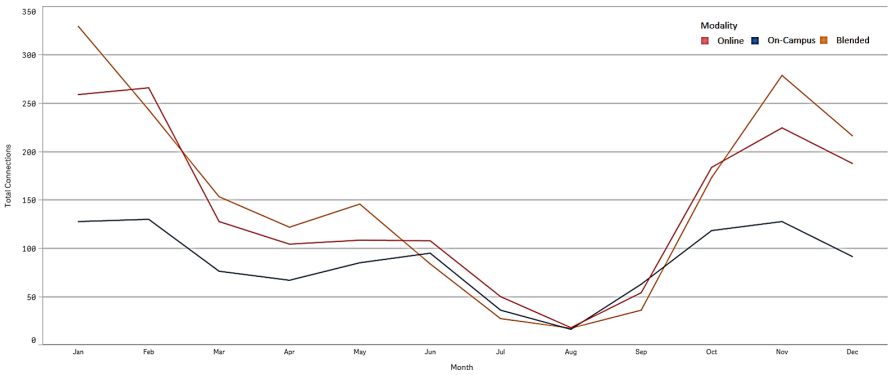


Figure 11: Monthly connection trends for each modality.

Secondly, Figure 12 presents the trends relating to the number of connections for each day of the week in the three modalities. A constant decrease is observed in the number of accesses throughout the week starting on Monday, with a more prominent decrease on Saturday and a small increase on Sunday. Unusually, this

pattern is shared by the three modalities, as higher activity was expected on weekends for online students. However, it is demonstrated that students prefer to follow a traditional organizational approach by working little by little during the week and resting over weekends.

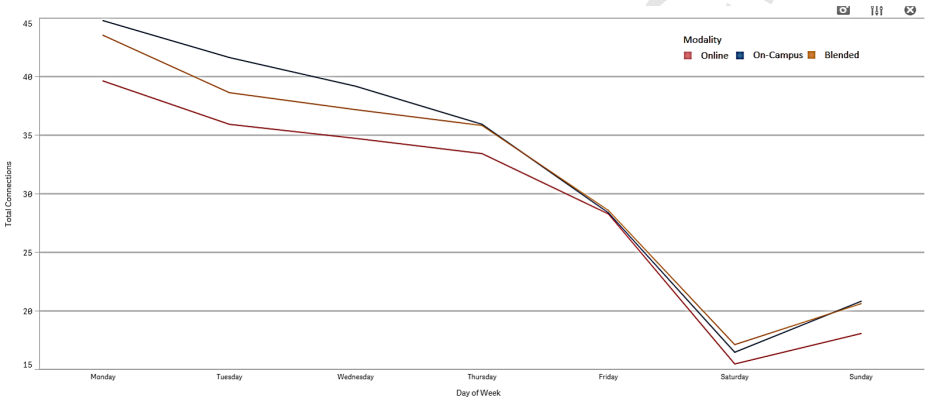


Figure 12: Weekly connection trends for each modality.

Finally, in Table 4, it is analyzed the mean number of visits for each modality and academic year. Owing to the very high standard deviation of such means, we have filtered the students with a number of visits between 25% (Q1) and 75% (Q3) of the total number of visits for each analysis, in order to rule out students with very low or high numbers of visits. However, there is still an item with a very high standard deviation for the blended modality in 2013/14.

The most relevant finding in this table is that the blended students exhibit the highest number of visits, while the online students exhibit a very low number of visits in comparison. Again, this demonstrates that having a low percentage of on-campus sessions results in higher student engagement. Focusing on each modality, the online students doubled their number of visits from the 2013/14 year to the next. This is owing to the inclusion of Lesson Builder in the 14/15 course, which motivates online students to visit the virtual campus more frequently. However, a very high decrease in the number of visits for on-campus and blended students is observed in the 15/16 year. This is owing to a change

Table 4: Mean number of accesses to Sakai grouped by modality and academic year. Students were selected between the Q1 and Q3 of the number of accesses in each analysis to reduce the standard deviation.

Modality		N	Q1	Q3	Mean	SD
Online	**2012/13**	236	44	731	215,00	172,093
	2013/14	363	45,00	868,00	271,50	215,040
	2014/15	860	49,00	1563,00	517,18	441,097
	2015/16	1542	33,00	1298,00	420,94	359,464
On-campus	**2012/13**	4754	303,00	1009,25	590,57	193,253
	2013/14	5138	278,00	854,25	524,05	160,786
	2014/15	5899	210,00	904,00	526,88	193,451
	2015/16	8106	24,00	438,00	164,48	124,033
Blended	**2012/13**	270	326,00	3428,25	1265,63	843,372
	2013/14	590	259,50	4649,00	1647,22	1176,263
	2014/15	1777	111,00	2630,50	988,65	700,200
	2015/16	2244	16,50	1123,75	364,19	338,107

to augment the inactivity time lapse during that year, reducing the number of expired sessions and the necessity of reconnecting again.

5. Conclusion and future work

The main goal of this paper resided in obtaining knowledge from data stored in an e-learning platform, such as the Sakai LMS. We have proposed the use of big data technologies and a framework to attempt to obtain student behavior patterns and be able to provide conclusions to increase student performance by improving their learning process. We selected a big data solution based on Azure HDInsight using its HDFS implementation. The tool used to transfer data from Sakai database was Sqoop, and these data were stored in a Hive [22] data warehouse. Moreover, we implemented the Apriori algorithm following the Hadoop MapReduce framework in order to obtain association rules for the

events performed by students in the Sakai LMS. Using these technologies and the big data framework, we studied and analyzed a database containing 70 GB of information regarding the behavior of UCAM students, including all of the available data on degrees and masters. The obtained results have been discussed, translated, and visually depicted visually in order to be easily interpreted by people who are not related to the big data field, such as degree coordinators, lecturers or students.

The results demonstrate that students for all modalities used the Forum tool before/after revising the resources and academic materials when posting and reading on it. Therefore, students reinforce their learning process indirectly by using the forums. This arrangement is surprising, as the forums are used even in the on-campus mode, where the use of certain tools is expected to be lower owing to the face-to-face interactions in the classroom. However, even for this modality, this is a reinforcement of the student learning process. Thus, lecturers should encourage the use of these and propose additional challenges to increase and encourage their use. Moreover, the Lesson Builder tool, specifically its event "Visit unit", is the most performed after the "Resource download" event. Therefore, lecturers should continue to use the templates provided by this tool, as students find the contents clearer and easier to use. This fact demonstrates that the organization of contents in the LMS, whether using the Sakai Lesson Builder or a similar tool in other platforms, could be a key factor in fostering student engagement. Finally, the blended students use the LMS Sakai more frequently, exhibiting many more accesses and thereby carrying out a more intense learning process. This result is surprising, as online students would be expected to use the tool most frequently because they do not have the opportunity to attend face-to-face classes.

The framework proposed in this paper may be used for further studies in this area; for example, to study lecturer behavior patterns as well. It can also be employed in other fields, such as smart cars, to identify good (or poor) driver behaviors, or smart homes to study the energy usage of inhabitants.

An immediate future line for this work is determining possible correlations

among student behavior patterns and their grades, in order to identify and promote behaviors that will aid in improving student qualifications. Moreover, we are studying the reasons why certain tools are more accepted in a training modality than in others, so as to upgrade these tools in such modalities with a lesser acceptance level.

Acknowledgments

This work is supported by the Spanish MINECO, under grant TIN2016-78799-P (AEI/FEDER, UE). The authors would like to thank members of the Online Department of this University for their participation in this paper. They also wish to thank the degree coordinators, lecturers, and students involved in the study.

References

[1] S. Ozkan, R. Koseler, Multi-dimensional students evaluation of e-learning systems in the higher education context: An empirical investigation, Computers & Education 53 (4) (2009) 1285 – 1296. doi:http://dx.doi.org/10.1016/j.compedu.2009.06.011.

[2] L. P. Macfadyen, S. Dawson, Mining LMS data to develop an "early warning system" for educators: A proof of concept, Computers & Education 54 (2) (2010) 588 – 599. doi:http://dx.doi.org/10.1016/j.compedu.2009.09.008.

[3] G. Siemens, P. Long, Penetrating the fog: Analytics in learning and education., EDUCAUSE review 46 (5) (2011) 30.

[4] J. M. Dodero, E. J. González-Conejero, G. Gutiérrez-Herrera, S. Peinado, J. T. Tocino, I. Ruiz-Rube, Trade-off between interoperability and data collection performance when designing an architecture for learning analytics, Future Generation Computer Systems 68 (2017) 31–37. doi:https://doi.org/10.1016/j.future.2016.06.040.

[5] 1st international conference on learning analytics and knowledge 2011, https://tekri.athabascau.ca/analytics/. Accessed 5 February 2017. (2011).

[6] Campus-Computing-Project, Campus Computing Survey, https://www.campuscomputing.net/ Accessed 15 January 2017 (2017).

[7] C. Romero, S. Ventura, Educational data mining: a review of the state of the art, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40 (6) (2010) 601–618.

[8] C. Romero, S. Ventura, E. García, Data mining in course management systems: Moodle case study and tutorial, Computers & Education 51 (1) (2008) 368–384.

[9] M. H. Falakmasir, J. Habibi, Using educational data mining methods to study the impact of virtual classroom in e-learning, in: In Proceedings of the 3rd international conference on educational data mining, 2010, pp. 241–248.

[10] Y. Psaromiligkos, M. Orfanidou, C. Kytagias, E. Zafiri, Mining log data for the analysis of learners behaviour in web-based learning management systems, Operational Research 11 (2) (2011) 187–200.

[11] J. Poltrack, N. Hruska, A. Johnson, J. Haag, The next generation of scorm: Innovation for the global force, in: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), Vol. 2012, National Training System Association Orlando, 2012.

[12] J.-L. Hung, Y.-C. Hsu, K. Rice, Integrating data mining in program evaluation of k-12 online education., Educational Technology & Society 15 (3) (2012) 27–41.

[13] K. Sin, L. Muthu, Application of big data in education data mining and learning analytics–a literature review., ICTACT journal on soft computing 5 (4). doi:10.21917/ijsc.2015.0145.

[14] B. Tulasi, Significance of big data and analytics in higher education, International Journal of Computer Applications 68 (14).

[15] B. Daniel, Big data and analytics in higher education: Opportunities and challenges, British journal of educational technology 46 (5) (2015) 904–920. doi:10.1111/bjet.12230.

[16] P. Ducange, R. Pecori, L. Sarti, M. Vecchio, Educational big data mining: how to enhance virtual learning environments, in: International Conference on EUropean Transnational Education, Springer, 2016, pp. 681–690. doi:https://doi.org/10.1007/978-3-319-47364-2_66.

[17] D. M. West, Big data for education: Data mining, data analytics, and web dashboards. governance studies at brookings., Brookings Institution.

[18] A. G. Picciano, The evolution of big data and learning analytics in american higher education., Journal of Asynchronous Learning Networks 16 (3) (2012) 9–20.

[19] J. Song, Y. Zhang, K. Duan, M. S. Hossain, S. M. M. Rahman, Tola: Topic-oriented learning assistance based on cyber-physical system and big data, Future Generation Computer Systemsdoi:https://doi.org/10.1016/j.future.2016.05.040.

[20] V. Kellen, A. Recktenwald, S. Burr, Applying big data in higher education: A case study, Cutter Consortium white paper 13 (8).

[21] M. Farhan, S. Jabbar, M. Aslam, M. Hammoudeh, M. Ahmad, S. Khalid, M. Khan, K. Han, Iot-based students interaction framework using attention-scoring assessment in elearning, Future Generation Computer Systemsdoi:https://doi.org/10.1016/j.future.2017.09.037.

[22] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, R. Murthy, Hive: a warehousing solution over a map-reduce framework, Proceedings of the VLDB Endowment 2 (2) (2009) 1626–1629.

[23] D. Ary, L. C. Jacobs, C. K. Sorensen, D. Walker, Introduction to research in education, Cengage Learning, 2013.

[24] R. Agrawal, R. Srikant, et al., Fast algorithms for mining association rules, in: Proc. 20th int. conf. very large data bases, VLDB, Vol. 1215, 1994, pp. 487–499.

[25] S. Ougiaroglou, G. Paschalis, Association rules mining from the educational data of esog web-based application, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2012, pp. 105–114.

[26] V. Murugananthan, B. ShivaKumar, An adaptive educational data mining technique for mining educational data models in elearning systems, Indian Journal of Science and Technology 9 (3).

[27] S. K. Verma, R. Thakur, S. Jaloree, Pattern mining approach to categorization of students' performance using apriori algorithm, International Journal of Computer Applications 121 (5).

[28] S. Singh, R. Garg, P. Mishra, Review of apriori based algorithms on mapreduce framework, arXiv preprint arXiv:1702.06284.

[29] J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters, Communications of the ACM 51 (1) (2008) 107–113.

[30] A. Nandeshwar, Tableau data visualization cookbook, Packt Publishing Ltd, 2013.

[31] M. García, B. Harmsen, Qlikview 11 for developers, Packt Publishing Ltd, 2012.

Appendix A.

Table A.5: Event acronyms

Acronym	Event Name
E1	Read announcement
E2	New message Chat
E3	Read Chat
E4	Exam started
E5	Exam revised
E6	Send exam
E7	Post an answer
E8	Read post
E9	Update Element
E10	Visit unit
E11	Read Mail
E12	New folder Mail
E13	Read assignment
E14	Read assignment instruction
E15	Save draft
E16	Download resource

Table A.6: The most generated events for each tool of Sakai LMS. For each academic year (rows) and for each tool (columns) it is shown the most used events by students per modality, where 'E_x' represents the event and the number represents the times that these events has been performed. For each academic course the modalities are represented by the acronym 'O','C' and 'B' for Online, On-Campus and Blended modalities respectively. The column named 'Announ.' refers to announcement tool. The event name with their acronyms are shown in Table A.5

Acd.Year	**Tool**	**Announcement**	**Chat**	**Exams**	**Forum**	**Lesson** Builder	**Internal** Mail	**Assignment**	**Resources**
2012/2013	**O**		E2 5002	E4 2693	E7 54487		E11 10793	E13 10316	E16 95342
	C		E2 53977	E4 6887	E7 159317		E12 282814	E13 86550	E16 3467309
	B		E2 15065	E4 2075	E7 484938		E12 16091	E13 10088	E16 234625
2013/2014	**O**		E2 5812	E4 2245	E7 47552	E9 4860	E11 15170	E13 13666	E16 136146
	C		E2 25576	E4 12292	E7 374865	E9 6237	E12 269721	E13 136365	E16 3886253
	B		E2 19737	E4 1192	E7 1165302		E11 34958	E13 34103	E16 430336
2014/2015	**O**	E1 84409	E3 6276	E5 5084	E7 50167	E10 874681	E11 45564	E14 43731	E16 507730
	C	E1 1125143	E3 45163	E5 12097	E8 450102	E10 395650	E11 273242	E14 171010	E16 4391931
	B	E1 267598	E3 42074	E6 14300	E8 640636	E10 1201058	E11 67086	E14 166321	E16 1526640
2015/2016	**O**	E1 8285712	E3 93303	E5 12978	E8 355962	E10 495685	E11 69752	E14 99121	E16 1072763
	C	E1 2062594	E3 44953	E5 9974	E8 42519	E10 267586	E12 257822	E14 281487	E16 3130343
	B	E1 609319	E3 148608	E6 9314	E8 642336	E10 1486544	E11 60919	E14 215965	E16 1826641

Magdalena Cantabella obtained her B.S. in Computer Science at the Catholic University of Murcia in 2008 and her M.S. in New Technologies in Computer Science applied to Biomedicine AT the University of Murcia in 2012. Since 2010 she is an associate professor in the Polytechnic School within the Department of Degree in Computer Engineering of the Catholic University of Murcia. Her areas of research include massive statistical analysis of data, e-learning and definition of user profiles.

Raquel Martínez-España is an associated professor in the Technical School at the Catholic University of Murcia (UCAM), Spain. She obtained her M.S. in Computer Science in 2009 and her PhD in Computer Science in 2014 at the University of Murcia. She has worked on several research projects in artificial intelligence and education. Raquel has participated in various academic and industry projects. His research interests include data mining, big data, soft computing, artificial intelligence and intelligent data analysis.

Belén López Ayuso obtained her M.S. in Computer Science from the University at Murcia and her PhD in Computer Science at the same University. She has 18 years of experience in teaching, both in Degree and Master courses at University Level, include e-learning methodology. She has participated in several educational innovation projects from which publications in the area of educational innovation have been obtained. At the moment she is the Dean of the Degree in Computer Engineering of the Catholic University of Murcia and Heat of the Online Department at this University. Her areas of research include teaching assessment and e-learning methodology evaluation.

Juan Antonio Yáñez obtained her B.S. in Computer Science at the Catholic University of Murcia in 2015, currently works as a computer consultant in a technology company and has started his doctoral studies in the area of research include massive statistical analysis of data.

Andrés Muñoz is a senior lecturer in the Technical School at the Catholic University of Murcia (UCAM), Spain. He obtained his PhD in Computer Science in 2011 at the University of Murcia. He has worked on several research projects in artificial intelligence and education. His main research interests include argumentation in intelligent systems, Semantic Web technologies and Ambient Intelligence and Intelligent Environments applied to education.

Magdalena Cantabella

Raquel Martínez-España

Belén Ayuso

Juan Antonio Yáñez

Andrés Muñoz

Highlights

- Big Data techniques have been applied in the Academic Analytics context.
- Implementation of Apriori algorithm by the Hadoop MapReduce framework.
- An analysis of student behavior patterns in e-learning platform has been performed.
- It is compared the activity and use of LMS tools according to learning methodologies.