[未来一代计算机系统 90 (2019) 262-272](https://doi.org/10.1016/j.future.2018.08.003)

内容列表可在 [科学直通车](http://www.elsevier.com/locate/fgcs)

未来一代计算机系统

期刊主页: [www.elsevier.com/locate/fgcs](http://www.elsevier.com/locate/fgcs)

通过大数据框架分析学生在学习管理系统中的行为

[magdalena cantabella](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11), [raquel martinez-es帕尼亚](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11) [请使用](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page1), [贝伦阿伊索](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11) , [胡安·安东尼奥·亚涅斯](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11) , [安德烈斯·穆尼奥斯](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)

*西班牙 zip 30007年穆尔西亚天主教大学计算机科学系*

h "

· ·大数据技术已在学术分析环境中得到应用.

· ·由 hadoop mapreduce 框架实现 apriri 算法.

· ·对电子学习平台中的学生行为模式进行了分析.

· ·根据学习方法, 对 lms 工具的活动和使用进行了比较.

a r t t l l l o

*文章历史:*

收到2017年12月20日收到修订表格2018年7月14日接受 2018年8月2日2018年8月10日在线提供

*关键 字：*

学习管理系统

Mapreduce

apriri 算法

电子学习分析

学生行为

大数据

a b s t r a c t

近年来, 学习管理系统 (lms) 在高等教育教学模式中发挥了根本作用。在寻找改善学习过程的模式时, 已经开辟了一条新的研究线, 涉及对学生在 lms 中的行为的分析。目前的电子学习平台可以记录学生的活动, 从而能够探索使用 lms 工具所产生的事件。本文介绍了在穆尔西亚天主教大学进行的一项案例研究, 在该大学中, 考虑到访问的次数, 根据学习方式 (即校园、在线和混合) 对过去四个学年的学生行为进行了分析。lms, 学生使用的工具及其相关活动。鉴于难以管理 lms 中用户生成的大量数据 (本研究中的数据高达 70 gb), 使用大数据框架执行了统计和关联规则技术, 从而加快了对数据的统计分析。利用视觉分析技术对所得结果进行了演示, 并对其进行了评估, 以发现学生使用 lms 的趋势和不足。

© 2018年 elsevier b. v。保留所有权利。

**1. 导言**

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 fac-tors: 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11),[2](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]), there exists a need for analytical tools to help to interpret LMS data and

∗ Corresponding author.

*E-mail addresses:*[mmcantabella@ucam.edu](mailto:mmcantabella@ucam.edu)(M.Cantabella),[rmartinez@ucam.edu](mailto:rmartinez@ucam.edu)(R. Martínez-España), [bayuso@ucam.edu](mailto:bayuso@ucam.edu)(B. Ayuso), [juan.yanez@alu.ucam.edu](mailto:juan.yanez@alu.ucam.edu)(J.A. Yáñez), [amunoz@ucam.edu](mailto:amunoz@ucam.edu)(A. Muñoz).

<https://doi.org/10.1016/j.future.2018.08.003>

0167-739X/© 2018 Elsevier B.V. All rights reserved.

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 intel-ligible 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 tech-niques 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

|  |  |
| --- | --- |
| *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272* | 263 |

mining are studied [[3](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11),[4](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]. Within this field, the areas most relevant to our work are learning analytics and visual analytics. The for-mer 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]. 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 ex-tracted 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, amount-ing 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.

这项研究是在穆尔西亚天主教大学 (ucam) 进行的。自1996年以来, 该大学提供了几个校园学位, 在过去五年内, 该大学通过在线和混合模式的几个学位, 巩固了培训机会。酒井 lms[1](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page2)被用作所有培训模式的资源管理和合作平台.

本文的其余部分结构如下。 [第二](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page2)节回顾了 以前与电子数据分析有关的几项工作。 [第3部](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page3)分介绍了我们基于大数据技术的总体框架建议, 以便分析学生数据。 [第4](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page5)节提供了我们的建议的结果, 从分析 70 gb 的酒井 lms 数据收集在四个学年。最后, 结论和今后的工作概述在 [第5](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page10)节.

**2. 相关工作**

The inclusion of an LMS as an essential methodological tool in higher education is the standard at present [[6](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)], 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]. 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)], Moodle was used as an LMS to analyze the inte-gration 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)], the necessary requirements for integrating data min-ing services into the Sharable Content Object Reference Model (SCORM) (see [[11](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] for details on this standard) compliant plat-forms were reviewed. The authors proposed analyzing the records

1 <https://sakaiproject.org/>.

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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)], 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.

众所周知, 近年来, lms 产生的数据大幅增加。因此, 目前 lms 数据的分析技术必须不断发展, 适应高等教育机构面临的新挑战。推荐的解决方案是在电子学习中使用大数据作为一门新兴学科。正如其他地方所指出的那样, 从教育数据挖掘到教育大数据的演变是一个现实 [[13](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11),[14](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]。大数据为我们提供了在使用 lms 方面达到更高水平的机会, 通过根据从大数据结果中获得的战略响应做出决策, 从学生体验中获得更多好处。因此, 可以将复杂的非结构化数据转换为可操作的信息, 从而有助于识别有用的数据, 并将其转化为高等教育机构的有价值的信息 [[15](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)、[16](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)].

West [[17](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] and Picciano [[18](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] 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 demon-strated the benefit of applying big data techniques in higher edu-cation, such as [[19](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)], 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 online courses. In this manner, forum topics can be classified automatically, so that lecturers can make spe-cific comments depending on the classification, while students can quickly find the required content. Likewise, the work in [[20](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] 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 activ-ities, and low scores usually imply poor student achievements. Finally, [[21](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] demonstrated that interactions with learning envi-ronments can be modeled and measured effectively. The authors defined an IoT-based interaction framework and analyzed the stu-dent 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.

264 *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272*

**Fig. 1.**Big data architecture for acquiring and storing Sakai data.

**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[2](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page3)has been adopted, using its Hadoop distributed file system (HDFS) implementation. The tool used to transfer data from the Sakai database is Sqoop.[3](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page3)These data are stored in a Hive [[22](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] data warehouse owing to their analytical features, as explained in Section [3.2](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page3). [Fig. 1](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page3)illustrates the architectural schema for the data acquisition and storage steps, and the technologies involved.

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 (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

2 <https://azure.microsoft.com/en-us/services/hdinsight/>.

3 <http://sqoop.apache.org/>.

with noise and irrelevant features. In this work, we design and im-plement 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 algo-rithm. 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] 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 plat-form (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 possibleto 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]. This is one of the most popular and widely used algorithms in both data mining and educational data [[13](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11),[25](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)–[27](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]. 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 tech-nique that can be implemented in a distributed and parallel man-ner [[28](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]. 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

|  |  |
| --- | --- |
| *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272* | 265 |

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 associ-ations 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]. 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 min-imum 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 confi-dence threshold. The confidence is the conditional probability that a transaction containing item *X* also contains item *Z*.

**Algorithm 1:**General procedure of Apriori technique

**Input***Support*,*D*

*L*1={Get-frequent-1-item-rules(*D*)}**for all**(*k*=2;*Lk*−1! =∅;*k*+ +)**do**

*Ck*=candidates generated from*Lk*−1**for all**(*transaction t*∈*D*)**do**

*Ct*=*subset*(*Ck*,*t*)

**for all**(*candidatec*∈*Ct*)**do**

c.support++;

**end for**

**end for**

*Lk*= {*c*∈*Ci*|*c*.*support*≥*Support*}

**end for**

⋃

**Output** *Lk*

*k*

These two phases are demonstrated in algorithm [1](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page4). The algo-rithm receives the minimum support value and *D* datasets to work with as input. Firstly, the 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 *Lk* rules satisfying the confidence threshold are generated. This final step is considered as a prune, and is reflected in algorithm [1](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page4) in the code line *Lk* = {*c* ∈ *Ci* | *c*.*support* ≥ *Support*}.

Algorithm [1](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page4)was adapted to the MapReduce framework in order to be executed in our Azure HDInsight configuration, as explained in Section [3.1](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page3). [Fig. 2](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page4)depicts the first phase of the algorithm; specif-ically, 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.

**Fig. 2.**General scheme of function Get-frequent-1-item-rules.

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 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 inter-vals 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] and QlikView [[31](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)]. 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

266 *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272*

**Table 1**

Number of master’s and degrees by modality and areas of study.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Areas of study |  | Online | On-campus | Blended |  |
|  |  |  |  |  |  |
| Social | Degree | 1 | 6 | 3 |  |
| sciences | 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 | Degree | 1 | 1 | 1 |  |
| law | Master | 4 | 4 | 3 |  |

**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 |

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 advan-tages 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page5), grouped according to training modality. The total number of students registered during this time period amounts to 76,268, as indicated in [Table 2](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page5). The events gen-erated by the students during this period amount to 79,432,423, distributed by modality and academic year, as indicated in [Table 3](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page5).

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 as-sociations 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.

*4.1. Tool ranking*

According to [Figs. 3–5](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page6), 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 [Fig. 3](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page6), there is a sig-nificant increase in the student activity for the academic year 14/15 in almost all tools. This is a direct consequence of apply-ing 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 Announce-ments. 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.

• 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 [Fig. 4](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page6). 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.

• For the on-campus modality, the majority of the tools exhibit a decreasing tendency of use, with certain exceptions, as can be observed in [Fig. 5](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page6). The Announcement and Assignment tools are the only ones with an increasing tendency, which is presumably motivated by the introduction of automatic noti-fications sent to the students’ mail when certain information is published.

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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page10), as illustrated in [Appendix](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11), 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

|  |  |
| --- | --- |
| *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272* | 267 |

**Fig. 3.**Evolution for academic years regarding use of Sakai tools in online modality.

**Fig. 4.**Evolution for academic years regarding use of Sakai tools in blended modality.

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

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. How-ever, 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. [Fig. 6](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page7) 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

268 *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272*

**Fig. 6.**Example of typical organization for learning unit for Lesson Builder tool in Sakai.

**Fig. 7.**Event ranking for online modality.

**Fig. 9.**Event ranking for blended modality.

**Fig. 8.**Event ranking for on-campus modality.

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. [Figs. 7–9](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page7)illustrate the top 10 event rankings grouped according to modality. The *y*-axis indicates the event iden-tifier, while the *x*-axis shows the number of records for each event. These rankings were obtained using the ad-hoc HiveQL queries explained in Section [3.2](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page3). Comments on these rankings are provided as follows.

• [Fig. 7](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page7)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 fi-nally, the events with the least activity are the ‘‘Read assign-ment’’ and ‘‘Read message’’ events.

• [Fig. 8](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page7)presents the events for the on-campus modality. The most frequent events are ‘‘Download resource’’ and ‘‘Read an-nouncement’’; moderate occurrences of the ‘‘Read post (Fo-rum)’’ and ‘‘Update profile’’ events are observed; the events with the least activity are ‘‘Read message’’ and ‘‘Save draft (Assignment)’’.

• [Fig. 9](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page7)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’’.

After analyzing these rankings, we find that the event most repeated by students in the three modalities is the ‘‘Download re-source’’ 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

|  |  |
| --- | --- |
| *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272* | 269 |

**Fig. 10.**Combinations of events performed by students in same session grouped bymodality.

attached file. This activity corresponds to the event ‘‘Read as-signment instructions’’, which offers further possibilities when defining assignments, such as using increased space in writing, in-tegrating 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page8)), an in-depth analysis is performed on the possible relationships among the most fre-quently used events.

*4.3. Event relationships*

By applying and HiveQL queries and Apriori algorithm pre-sented 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, [Fig. 10](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page8)uses a bubble chart to depict the probabil-ities 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.

• For the online modality:

**–**Students read the instruction for an assignment (event14) before submitting it (event 17).

**–**Students read an announcement (event 1) before send-ing an assignment (event 17), owing to the fact that lecturers provide the final indications by means of an-nouncements.

**–**Before posting an answer in the forum (event 7), stu-dents consult the necessary resource (event 18) to sub-mit 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 stu-dents access resources (event 18) before posting an an-swer (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 theresources (event 18) before reading (event 8) or posting an answer (event 7) in a forum.

**–**As with online students, blended students read the in-structions (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 instruc-tions to avoid errors in the submission of assignments.

**–**Students read an announcement (event 1) and then readinstructions for an assignment (event 14). This is owing to the fact that, on many occasions, lecturers notify the activation of a pending assignment through announce-ments 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.

**Fig. 11.**Monthly connection trends for each modality.

270 *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272*

**Fig. 12.**Weekly connection trends for each modality.

In addition to these relationships among events, it is important to note that on-campus students do not follow a behavior pat-tern 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 assign-ment. 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 an-alyzed 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, [Fig. 11](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page8)illustrates the tendency with respect to the num-ber 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 lectur-ers (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 [Fig. 11](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page8), the course opening months of September to November and holiday term months of July to August are identifiable.

Secondly, [Fig. 12](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page9)presents the trends relating to the number of connections for each day of the week in the three modalities. A con-stant 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

**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 |  |
|  |  |  |  |  |  |  |  |  |  |
|  | 2012/13 | | 236 | 44 | 731 | 215,00 |  | 172,093 |  |
| Online | 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 |  |
|  | 2012/13 | | 4754 | 303,00 | 1009,25 | 590,57 |  | 193,253 |  |
| On-campus | 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 |  |
|  | 2012/13 | | 270 | 326,00 | 3428,25 | 1265,63 | | 843,372 |  |
| Blended | 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 |  |
|  | **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 | |  |  |  |  |

students prefer to follow a traditional organizational approach by working little by little during the week and resting over weekends.

Finally, in [Table 4](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page9), 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,

|  |  |
| --- | --- |
| *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272* | 271 |

**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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page9).

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

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 stu-dents 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 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](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page11)] data warehouse. More-over, 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 master’s. 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 class-room. However, even for this modality, this is a reinforcement of the student learning process. Thus, lecturers should encour-age the use of these and propose additional challenges to in-crease 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 demon-strates that the organization of contents in the LMS, whether using the Sakai Lesson Builder or a similar tool in other plat-forms, 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 stu-dents would be expected to use the tool most frequently be-cause 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 pat-terns 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

272 *M. Cantabella et al. / Future Generation Computer Systems 90 (2019) 262–272*

others, so as to upgrade these tools in such modalities with a lesser acceptance level.

**Acknowledgments**

This work is supported by the Spanish MINECO, Spain, 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.

**Appendix**

See [Tables A.5](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page9)and [A.6](https://www.translatoruser.net/bvsandbox.aspx?&from=en&to=zh-CHS&csId=a7dc6b60-5bae-4d91-ac21-d04f5b1dd3d2&usId=3c5c551d-f80f-47af-9cd3-6d239036a8f8&ac=true&bvrpx=false&bvrpp=&dt=2018%2F12%2F28%202%3A24#page10).

**References**

[1] S. Ozkan, R. Koseler, Multi-dimensional students’ evaluation of e-learning systems in the higher education context: An empirical investigation, Comput. Educ. 53 (4) (2009) 1285–1296, [http://dx.doi.org/10.1016/j.compedu.2009.](http://dx.doi.org/10.1016/j.compedu.2009.06.011) [06.011](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, Comput. Educ. 54 (2) (2010) 588– 599, <http://dx.doi.org/10.1016/j.compedu.2009.09.008>.

[3] [G. Siemens, P. Long, Penetrating the fog: Analytics in learning and education,](http://refhub.elsevier.com/S0167-739X(17)32921-7/b3) [EDUCAUSE Rev. 46 (5) (2011) 30.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b3)

[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 Gener. Comput. Syst. 68 (2017) 31–37, [http://dx.doi.org/10.1016/j.future.](http://dx.doi.org/10.1016/j.future.2016.06.040) [2016.06.040](http://dx.doi.org/10.1016/j.future.2016.06.040).

[5] 1st International Conference on Learning Analytics and Knowledge 2011, 2011, <https://tekri.athabascau.ca/analytics/>. (Accessed 5 February 2017).

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

[7] [C. Romero, S. Ventura, Educational data mining: a review of the state of the](http://refhub.elsevier.com/S0167-739X(17)32921-7/b7) [art, IEEE Trans. Syst. Man Cybern. C 40 (6) (2010) 601–618.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b7)

[8] [C. Romero, S. Ventura, E. García, Data mining in course management systems:](http://refhub.elsevier.com/S0167-739X(17)32921-7/b8) [Moodle case study and tutorial, Comput. Educ. 51 (1) (2008) 368–384.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b8)

[9] M.H. Falakmasir, J. Habibi, Using educational data mining methods to study the impact of virtual classroom in e-learning, 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](http://refhub.elsevier.com/S0167-739X(17)32921-7/b10) [analysis of learners’ behaviour in web-based learning management systems,](http://refhub.elsevier.com/S0167-739X(17)32921-7/b10) [Oper. Res. 11 (2) (2011) 187–200.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b10)

[11] J. Poltrack, N. Hruska, A. Johnson, J. Haag, The next generation of scorm: Inno-

vation for the global force, in: The Interservice/Industry Training, Simulation & Education Conference, I/ITSEC, vol. 2012, 2012, National Training System Association Orlando.

[12] [J.-L. Hung, Y.-C. Hsu, K. Rice, Integrating data mining in program evaluation](http://refhub.elsevier.com/S0167-739X(17)32921-7/b12) [of K-12 online education., Educ. Technol. Soc. 15 (3) (2012) 27–41.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b12)

[13] K. Sin, L. Muthu, Application of big data in education data mining and learning analytics–A literature review, ICTACT J. Soft Comput. 5 (4) (2015) [http://dx.](http://dx.doi.org/10.21917/ijsc.2015.0145) [doi.org/10.21917/ijsc.2015.0145](http://dx.doi.org/10.21917/ijsc.2015.0145).

[14] [B. Tulasi, Significance of big data and analytics in higher education, Int. J.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b14) [Comput. Appl. 68 (14) (2013).](http://refhub.elsevier.com/S0167-739X(17)32921-7/b14)

[15] B. Daniel, Big data and analytics in higher education: Opportunities and challenges, British J. Educ. Technol. 46 (5) (2015) 904–920, [http://dx.doi.org/](http://dx.doi.org/10.1111/bjet.12230) [10.1111/bjet.12230](http://dx.doi.org/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, [http://dx.](http://dx.doi.org/10.1007/978-3-319-47364-2_66) [doi.org/10.1007/978-3-319-47364-2\_66](http://dx.doi.org/10.1007/978-3-319-47364-2_66).

[17] [D.M. West, Big Data for Education: Data Mining, Data Analytics, and Web](http://refhub.elsevier.com/S0167-739X(17)32921-7/b17) [Dashboards. Governance Studies at Brookings, 2012,](http://refhub.elsevier.com/S0167-739X(17)32921-7/b17)

[18] [A.G. Picciano, The evolution of big data and learning analytics in american](http://refhub.elsevier.com/S0167-739X(17)32921-7/b18) [higher education., J. Asynchronous Learning Netw. 16 (3) (2012) 9–20.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b18)

[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 Gener. Comput. Syst. (2016) <http://dx.doi.org/10.1016/j.future.2016.05.040>.

[20] [V. Kellen, A. Recktenwald, S. Burr, Applying Big Data in higher education: A](http://refhub.elsevier.com/S0167-739X(17)32921-7/b20) [case study, Cutter Consortium White Paper 13 (8) (2013).](http://refhub.elsevier.com/S0167-739X(17)32921-7/b20)

[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 Gener. Comput. Syst. (2017) [http://dx.doi.](http://dx.doi.org/10.1016/j.future.2017.09.037) [org/10.1016/j.future.2017.09.037](http://dx.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,](http://refhub.elsevier.com/S0167-739X(17)32921-7/b22)

R. [Murthy, Hive: a warehousing solution over a map-reduce framework, Proc.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b22) [VLDB Endow. 2 (2) (2009) 1626–1629.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b22)

[23] [D. Ary, L.C. Jacobs, C.K. Sorensen, D. Walker, Introduction to Research in](http://refhub.elsevier.com/S0167-739X(17)32921-7/b23) [Education, Cengage Learning, 2013.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b23)

[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](http://refhub.elsevier.com/S0167-739X(17)32921-7/b25) [data of ESOG web-based application, in: IFIP International Conference on](http://refhub.elsevier.com/S0167-739X(17)32921-7/b25) [Artificial Intelligence Applications and Innovations, Springer, 2012, pp. 105–](http://refhub.elsevier.com/S0167-739X(17)32921-7/b25) [114.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b25)

[26] [V. Murugananthan, B. ShivaKumar, An adaptive educational data mining](http://refhub.elsevier.com/S0167-739X(17)32921-7/b26) [technique for mining educational data models in elearning systems, Indian](http://refhub.elsevier.com/S0167-739X(17)32921-7/b26)

J.[Sci. Technol. 9 (3) (2016).](http://refhub.elsevier.com/S0167-739X(17)32921-7/b26)

[27] [S.K. Verma, R. Thakur, S. Jaloree, Pattern mining approach to categorization of](http://refhub.elsevier.com/S0167-739X(17)32921-7/b27) [students’ performance using apriori algorithm, Int. J. Comput. Appl. 121 (5)](http://refhub.elsevier.com/S0167-739X(17)32921-7/b27) [(2015).](http://refhub.elsevier.com/S0167-739X(17)32921-7/b27)

[28] S. Singh, R. Garg, P. Mishra, Review of apriori based algorithms on mapreduce framework, arXiv preprint [arXiv:1702.06284](http://arxiv.org/abs/1702.06284), 2017.

[29] [J. Dean, S. Ghemawat, MapReduce: simplified data processing on large clus-ters, Commun. ACM 51 (1) (2008) 107–113.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b29)

[30] [A. Nandeshwar, Tableau Data Visualization Cookbook, Packt Publishing Ltd,](http://refhub.elsevier.com/S0167-739X(17)32921-7/b30) [2013.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b30)

[31] [M. García, B. Harmsen, Qlikview 11 for Developers, Packt Publishing Ltd, 2012.](http://refhub.elsevier.com/S0167-739X(17)32921-7/b31)

**Magdalena Cantabella**obtained her B.S. in ComputerScience at the Catholic University of Murcia in 2008 and her M.S. in New Technologies in Computer Science ap-plied 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 inthe Technical School at the Catholic University of Murcia (UCAM), Spain. She obtained her M.S. in Computer Sci-ence in 2009 and her Ph.D. 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 Sci-ence from the University at Murcia and her Ph.D. in Com-puter 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 method-ology. She has participated in several educational innova-tion projects from which publications in the area of edu-cational 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 On-line 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 Sci-ence at the Catholic University of Murcia in 2015, cur-rently 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 Schoolat the Catholic University of Murcia (UCAM), Spain. He obtained his Ph.D. 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 Intelli-gence and Intelligent Environments applied to education.