通过大数据框架分析学生管理系统中的学生行为

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

强调

- 大数据技术已应用于 Academic Analytics 环境中。
- 通过 Hadoop MapReduce 框架实现 Apriori 算法。
- 对电子学习平台中的学生行为模式进行了分析。
- 它根据学习方法比较了 LMS 工具的活动和使用。

摘要

近年来，学习管理系统（LMSs）在高等教育教学模式中发挥了重要作用。已经开设了一个新的研究方向，涉及分析 LMS 中的学生行为，寻找改善学习过程的模式。目前的电子学习平台允许记录学生活动，从而能够探索使用 LMS 工具时产生的事件。本文介绍了在穆尔西亚天主教大学进行的一项案例研究，根据学习方式（即校内，在线和混合）分析了过去四个学年的学生行为，考虑到学生的访问次数。 LMS，学生使用的工具及其相关活动。由于难以管理 LMS 中用户生成的大量数据（本研究中高达 70 GB ），因此使用大数据框架执行统计和关联规则技术，从而加快统计分析速度的数据。使用视觉分析技术证明获得的结果，并进行评估以检测学生使用 LMS 的趋势和不足。

关键词

学习管理系统，MapReduce，Apriori 算法，电子学习分析，学生的行为，大数据

1 、介绍

高等教育目前的趋势包括分析和处理与用户通过使用学习管理系统（LMS）产生的活动有关的数据。从这些平台中提取的大量数据提供了基本信息，可以帮助教师和学生改善他们的教育目标。目前的主要问题之一是对这些信息的分析，这是由于两个主要因素：已经提到的大量可用数据，以及这些数据的不同格式，特别是对非结构化数据的管理。

根据一些研究（例如，参见 [1]，[2] ），需要分析工具来帮助解释 LMS 数据，并为改进甚至设计新的电子学习技术和方法提供新知识。在操纵此类信息之前，根据要实现的目标，从 LMS 中探索和选择必要的数据也很重要。

这项工作的主要目标是设计和实施基于大数据技术的框架，以识别 LMS 用户的行为模式，并以直观和可理解的方式对其进行说明。为此，我们定义以下步骤：

- 数据预处理，通过研究从 LMS 中提取的数据及其在大数据平台中的存储。
- 可以在教育背景下提供价值的模式识别技术的数据分析和识别。
- 根据合适的可视化分析技术和工具呈现所获得的结果。

为了开发这些步骤，我们考虑了以电子学习分析为指导的数据处理，其中研究了教育技术，学习概念和教育数据挖掘之间的联系 [3]， [4] 。在此领域内，与我们工作最相关的领域是学习分析和视觉分析。前者帮助我们进行数据处理，以发现学生，教师和学习过程之间的联系，目的是创建改善整体教育过程的建议。后者使用可视化界面来说明从分析推理中获得的结果，有助于理解新知识并帮助用户发现新的关系或可能的不规范 [5] 。在这里，我们通过将大数据技术集成到教育数据分析中，在使用电子学习分析方面向前迈出了一步。通过这种方式，可以通过应用于大量数据的分析技术来检测电子学习方法的趋势和缺陷。

我们建议对我们大学提供的三种学习方式（即校内，在线和混合）的所有课程中四个完整学年期间生成的用户事件中提取的 LMS 数据进行探索和分析，总计 70 GB 的数据。该提案的目的是评估通过将大数据框架应用于这些 LMS 数据获得的结果是否有助于检测在任何学习模式中使用这些平台的趋势和异常。

该研究在穆尔西亚天主教大学（UCAM）进行。自 1996 年以来，已经提供了几个校内学位，并且在过去的五年中，该大学通过在线和混合方式的几个学位巩固了其培训课程。Sakai LMS 1 用作所有培训模式的资源管理和协作平台。

本文的其余部分的结构如下。第 2 节 回顾了以前与教育数据分析有关的几项工作。第 3 节 解释了我们基于大数据技术的总体框架提案，以便分析学生数据。第 4 节 通过分析在四个学年内收集的 70 GB Sakai LMS 数据，提供了我们的建议结果。最后，第 5 节 概述了结论和未来的工作。

2 、相关工作

将 LMS 作为高等教育中必不可少的方法工具包含在内是目前的标准 [6] ，通过从 LMS 数据中获得的知识，产生新的需求和研究领域，以帮助设计新的学习模型。LMS 提供大量数据，同时还需要集成在 LMS 中的智能工具，以帮助解释并提供对此信息的反馈。该领域的一个热门话题是使用数据挖掘技术识别用户行为模式，这被称为教育数据挖掘 [7]。用户行为模式的识别旨在开发新的教学方法，通过分析 LMS 提供的数据和其他工具（如调查）来帮助和改善学生和教师的表现。

罗梅罗等人 [8] 进行了一项理论研究，探索数据挖掘在 Moodle LMS 中的应用。他们的目的是为启动这一学科提供指导。提供了用于电子学习的主要数据挖掘技术的详细信息，并将这些技术与 Moodle 中评估的实际案例进行了比较。同样，在 [9]中，Moodle 被用作 LMS 来分析数据挖掘技术与数据仓库工具和在线分析处理的集成。作者对有助于改善学生参与电子学习方式的表现和结果的活动进行了分类。

在 [10]中，审查了将数据挖掘服务集成到可共享内容对象参考模型（SCORM）（参见[11]有关此标准的详细信息）兼容平台的必要要求 。作者建议基于 Web 服务器访问日志分析记录以获得学生行为。这些记录可以提供大量数据，表示点击流或点击流数据。本文最后指出，从日志中获取的数据非常有限，并未提供生成所需信息的必要要求。[12]中说明了不同的程序 作者采用创新方法通过数据挖掘技术进行课程评估。他们通过 LMS 中的活动记录，人口统计数据和课程结束评估调查分析了 K-12 级学生的学习行为。多种数据形式的使用允许对学生行为进行更有意义的分析并识别可能的关系。

众所周知，LMS 产生的数据近年来已大幅增加。因此，LMS 数据的当前分析技术必须发展并适应高等教育机构面临的新挑战。推荐的解决方案是将电子学习中的大数据用作新兴学科。正如其他地方[13]，[14] 所述，从教育数据挖掘到教育大数据的演变需求已成为现实 。大数据为我们提供了在 LMS 使用中达到更高水平的机会，通过根据从大数据结果获得的战略响应做出决策，从学生体验中获得更多益处。因此，可以转换复杂的非结构化数据转化为可操作的信息，从而有助于识别有用的数据并将其转化为有价值的信息，供高等教育机构使用 [15]，[16]。

West [17]和 Picciano [18] 描述了在教育环境中使用大数据技术的初步概要，但这些工作没有详细研究具体的技术或方法。他们的目的是通过 LMS 检查高等教育中不断发展的大数据和分析世界。这两项研究恰逢预期的好处，有助于确定新的教学方法。这可能代表了决策和教育战略的巨大进步，允许分析大量数据并提供进一步提取知识的可能性。

此外，我们发现有趣的研究证明了在高等教育中应用大数据技术的好处，例如 [19] ，其中根据从大规模开放在线课程（MOOCs）中集成的论坛工具中提取的数据搜索学生学习模式。在上述工作中，开发了基于大数据的信息模型，称为面向主题的学习辅助，其提供在线课程中的论坛排名。通过这种方式，可以自动对论坛主题进行分类，以便讲师可以根据分类进行具体评论，同时学生可以快速找到所需内容。同样，[20]中的工作 介绍了 SAP HANA，这是一项基于数据的大型分析和监控工具，为 LMS 中的学生实施评分系统。分数代表学生参与学习活动，低分通常意味着学生成绩不佳。最后， [21]证明了可以有效地模拟和测量与学习环境的相互作用。作者定义了基于物联网的交互框架，并分析了学生的电子学习经验。该框架使用注意力评分模型，通过基于面部和眼睛观察来测量他们的注意力水平来评估参加视频会议的学生的行为。

我们的工作扩展了该研究线，采用大数据来分析 LMS 数据。我们提出了一个基于大数据技术的框架，用于分析使用 Sakai LMS 中可用的每种学习工具生成的事件的大量数据，用于三种主要的培训方式：校内，在线和混合。

3 、使用大数据框架分析 Sakai 数据

本节描述了我们针对基于大数据技术的框架的提议，旨在分析 Sakai 数据。它包括三个阶段：数据采集和存储，数据分析和结果可视化。

3.1 、数据采集和存储

此阶段的目标是研究存储在 Sakai LMS 中的原始工作数据集，并将其提取到大数据存储平台。Sakai 数据存储在包含 100 个以上表的关系数据库中。在对这些表与用户行为模式的相关性进行深入研究之后，发现以下三个表包含了我们研究中最重要的信息：Sakai_User，Sakai_Session 和 Sakai_Event。第一个包含有关 Sakai 用户的基本信息：id，姓名，电子邮件和角色（例如，学生或教师）。Sakai_Session 存储信息关于平台上的用户登录：用户 ID，会话开始和结束日期。最后，Sakai_Event 在使用平台时存储用户事件的数据。这些数据包括会话 ID，事件 ID，事件日期和上下文（事件发生的过程）。这三个表之间存在关系，因此可以查询任何用户的会话和事件数据。

选择数据源后，数据将被匿名化，以保护个人信息，如姓名和电子邮件。接下来，需要将数据从 Sakai 数据库传输到大数据存储。为此，已采用基于 Azure HDInsight 2 的大数据解决方案，使用其 Hadoop 分布式文件系统（HDFS）实现。用于从 Sakai 数据库传输数据的工具是 Sqoop。3 这些数据由于其分析功能而存储在 Hive [22]数据仓库中，如第 3.2 节所述 。图。1 说明了数据采集和存储步骤的体系结构模式以及所涉及的技术。

部署用于分析 Sakai 数据的 Azure 集群使用在 Ubuntu 16.04 内核上运行的 Hadoop v2.7.3。该集群由两个头节点（每个四个核心和 28 GB RAM）和四个工作节点（每个八个核心和 28 GB RAM）组成。

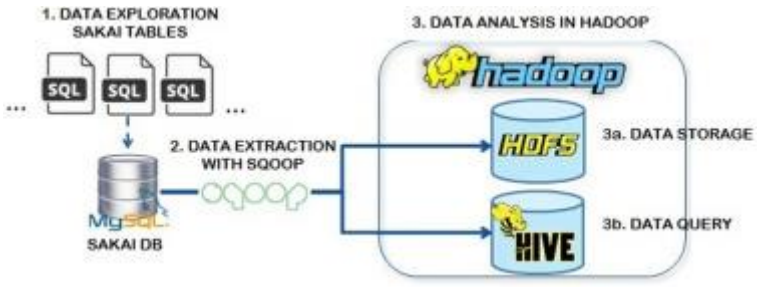

图 1. 用于获取和存储 Sakai 数据的大数据架构。

3.2 、数据分析

在存储数据之后，在开始从中获得结论之前，我们需要研究这些数据，以确定要考虑的最重要方面。应该注意，数据可能包含可能影响结果的噪声或不相关的特征。在这种情况下，我们有两个选择：我们可以执行数据预处理以删除所有不必要的功能，或者我们可以采用可以正常使用噪声和不相关功能的技术。在这项工作中，我们基于大数据特征设计和实现了几种技术，这些技术可以处理噪声和无关数据，而不会影响结果的质量。

如上一节所述，Hive 被选中用于存储学生数据，因为在 Hadoop 中使用此数据仓库系统使我们能够应用不同的技术，例如 HiveQL 查询或统计分析。

我们首先使用 HiveQL 执行定量分析，HiveQL 是 Hive 的临时查询系统。通过此分析，我们可以计算以下信息项。

- 工具排名：此分析研究每个学生在每个课程中使用的工具。会话定义为学生连接到电子学习平台的时间间隔。对于每个工具，我们分析每个工具最常用的事件。此外，在此过程中，我们分析会话中事件之间的相关性。将与 Apriori 算法一起使用和分析该相关性。该关联过程的目的是确定在同一会话期间哪些事件是相关的，以便确定学生的行为模式。因此，Pearson 相关系数 [23]通过 Hive 函数计算。
- 事件排名：此过程分析学生在每门课程中执行的事件，以便不仅识别每门课程中最常见的事件，而且还识别某些事件的缺席。它旨在检测活动高或低的课程，以及每种训练模式中发生的事件的全球排名。因此，该查询可以提供关于每个学生在特定课程/模态中执行的动作的某些见解以及可能缺乏可能与学生训练模态相关的动作（例如，反复不参加在线模态中的视频会议）。
- 事件趋势：此查询的目的是分析与电子学习平台中感兴趣的事件相关的时间线（例如，连接到 Sakai 的事件），以便识别某些重要的周期性模式。通过使用 Hive 提供的时间序列分析技术，可以在电子学习平台中识别具有高活动或低活动的时段。
- 连接趋势：此过程执行统计研究，以分析与 LMS 的连接中的每月和每周趋势，以及根据学术课程和培训模式分组的学生对 LMS 的平均访问次数。该信息可以帮助检测针对一周中的特定日期，月份，年份和模态的 LMS 访问次数的差异。

通过使用前面步骤中获得的信息，我们需要定义一种分析事件之间关联和序列的技术。该技术的理想特征是可解释性，鲁棒性和速度。因此，在研究了不同的可能性并考虑到数据量之后，我们选择了 Apriori 算法 [24] 。这是数据挖掘和教育数据中最流行和最广泛使用的算法之一[13]，[25]，[26]，[27]。这些参考文献在教育数据领域使用了 Apriori 算法; 在这种情况下，我们寻找一种模式寻求方法来分析学生的行为。

Apriori 算法是一种关联规则 数据挖掘技术，可以以分布式和并行方式实现 [28] 。其坚固性和可解释性使得可以获得可由非技术人员解释的可靠结果。

这是其选择的主要原因之一，增加了在 MapReduce 中实现它以便操纵大量数据的可能性。

该关联规则技术试图确定数据集中的项目或频繁模式之间的关联。为了并行化算法并能够处理大量数据，我们通过遵循 Hadoop MapReduce 框架 实现了这种技术 [29] 。该框架避免了网格计算的问题，其中节点总是存在失败的潜在机会，因此必须再次执行该任务。特别是，我们已经实现了 Apriori 算法的一个版本，同时考虑到教育背景的特征。此技术接收一组项目（在属性值中）format）作为输入并返回一组关联规则（项目规则）。Apriori 技术由两个阶段组成，如下所示。

- 在第一阶段，该技术计算每个项目的频率，然后计算不同项目组合的频率。对于最终项目规则集，将考虑超过特定阈值的项目（项目规则）的组合。此阈值称为支持，计算为项目重复次数除以事务数量，事务是包含不同项目的数据集中的条目。有必要建立最低限度的支持，以消除不太频繁的项目规则。
- 在第二阶段，从更频繁发生且超过置信度阈值的项目集生成一组项目规则。置信度是包含项目的事务的条件概率 X 还包含项目 ž。

Algorithm 1: General procedure of Apriori technique

```
Input Support, D
L1 ={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 ⋃k Lk
```

这两个阶段在算法 1 中得到证明。算法接收最小支持值和 d 要作为输入使用的数据集。首先，计算具有一个项目的所有项目规则的频率。接下来，算法使用该集合生成两个元素的子集，然后生成三个元素的子集，等等，直到不能创建其他子集组合。所有可能的对都遵守最低支持措施。最后，Lķ 生成满足置信度阈值的规则。最后一步被视为修剪，并反映在代码行中的算法 1 中 Lķ = { c ∈ Ci | c.support ≥ Support }。

算法 1 适用于 MapReduce 框架，以便在我们的 Azure HDInsight 配置中执行，如第 3.1 节所述 。图 2 描绘了算法的第一阶段; 具体来说，函数 Get-frequent-1-item-rules（d），计算最频繁的项目。在这里，'K'代表关键和我 X 表示不同的项目集（键，值）。这些值是使用 MapReduce API 提供的方法以及 HDFS 和 Hive 以分布式方式处理此算法的能力获得的。

关于本文所述的问题，项目是 LMS 事件，项目规则是一组事件，每个事件以特定频率发生，并且事务对应于学生 LMS 会话。输入数据集由所有会话和所有

学生的所有事件组成。根据这种技术，我们可以确定学生在 Sakai 平台上进行的不同事件的关联或反复行为。在本文中，Apriori 技术作为 MapReduce 过程实现了所提出的框架，以便从众多课程中获取大量事件作为输入。

关于最小支持度和置信度阈值，为支持值和置信度值建立的值分别在[20％至30％]和[80％至 70％]之间。经过几次评估测试，我们验证了这些时间间隔内的支持度和置信度值得到了类似的规则和结论。因此，对于下一节中介绍的研究案例，我们将支持率和置信度阈值分别固定在 20％和 80％。

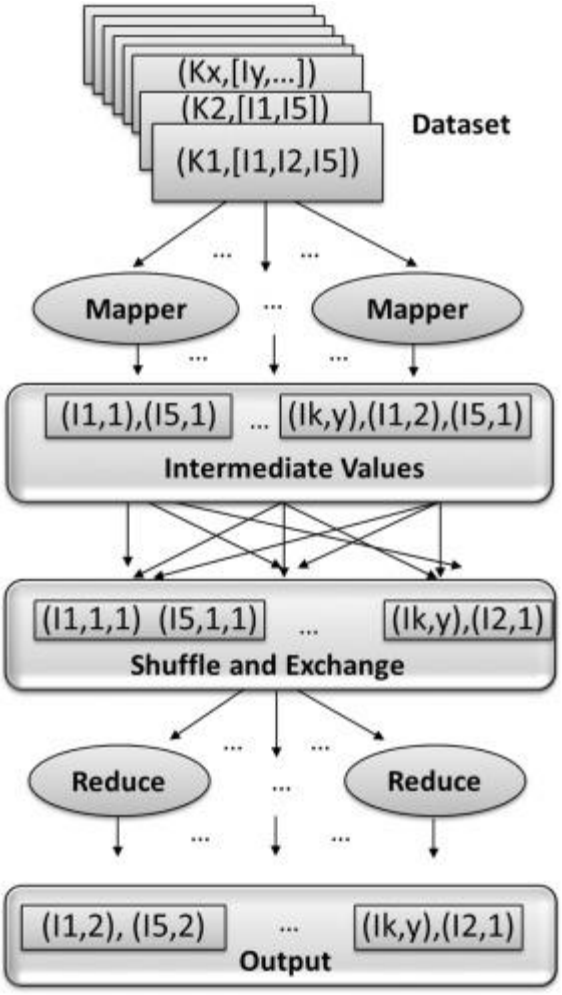

图 2. 函数 Get-frequent-1-item-rules 的一般方案。

3.3 、数据可视化

数据可视化，或说明在数据分析阶段后适当获得的结果的方式，是任何大数据相关项目的重要组成部分。结果必须以直观且易于理解的方式呈现，因为它们通常由非数据科学专家讨论。

为此，在提议的结果可视化框架中选择的工具是 Tableau [30]和 QlikView [31] 。Tableau 是图表，图形，地图和其他可视化类型的领先数据可视化工具之一。此工具允许以多种格式导出图形并将结果嵌入任何网页。此外，Tableau 还通过 Hive 管理存储在 Azure，Hortonworks，MapR 和 Amazon EMR 发行版中的大型复杂数据。我们使用 Tableau 的免费版本（称为

Tableau Public）作为我们的框架。它用于绘制排名事件图形和事件相关图形。

QlikView 是一种商业智能工具，可以处理来自多个来源的大量数据，以非常简单直观的方式处理和呈现这些数据。它的主要优点之一是它的仪表板可以在内存中实现数据集成 ; 因此，它可以在与数据源断开连接的情况下运行，并提供非常高的性能。本研究中使用此工具绘制与事件时间线相关的图形。

4 、案例分析

在本节中，我们将介绍我们的案例研究，其中通过我们提出的大数据框架分析从 Sakai 收集的 70 GB 事件数据。要分析的数据对应于我们大学所有学生在四个学年内为三种学习方式（在线，校内和混合）产生的 Sakai 事件; 也就是说，从 2012/2013 到 2015/2016。考虑到总共 41 个学士学位和 93 个硕士学位，如表 1 所示，根据训练方式分组。如表 2 所示，在此期间注册的学生总数为 76,268 。学生在此期间产生的事件达 79,432,423，按模式和学年分配，如图所示表 3。

以下部分讨论了我们案例研究中最相关的结果和要点，考虑了最常用的工具和事件，工具使用的趋势，包括检测到的工具之间的使用关联，最后，分析了登录记录以确定潜在的连接模式。在所有这些部分中，我们根据他们的训练方式分析了学生群体的行为。

表 1。按模态和学习领域分列的硕士和学士数量。

研究领域		**线上**	**在校园**	**混合**
社会	**学士**	**1**	**6**	**3**
科学	硕士	6	3	2
健康	学士	1	9	1
	硕士	3	4	22
运动	学士		3	
	硕士	2	2	12
工程	学士	1	五	
	硕士	3	3	4
商业	学士	2	五	1
	硕士	10	3	3
法律	学士	1	1	1
法	硕士	4	4	3

表 2。按学年和学年分组的学生人数。

形态/年	2012/2013	2013/2014	2014/2015	二千零十六分之二千零十五
线上	**628**	**863**	**1526**	**2849**
在校园	12114	13483	15333	16960
混合	2425	1885 年	3457	4745

表 3。按学年和模态分组的事件数。

Modal./year	2012/2013	2013/2014	2014/2015	二千零十六分之二千零十五
线上	**335557**	**593423**	**2169552**	**4205723**
在校园	12410137	14689112	16420582	9745588
混合	1467401	3637706	6373376	7384266

4.1 、工具排名

根据图 3，图 4，图 5 ，通过使用关于学生数量的标准化数据，我们在使用 Sakai 工具方面突出了关于四个学年的演变的以下发现。

- 对于在线模态，如图 3 所示几乎所有工具中，14/15 学年的学生活动都有显着增加。这是对讲师必须采用的教材和方法规则的直接结果。这种增加在使用课程生成器和资源工具方面非常重要，在较小的范围内，但在分配和公告中也非常重要。这种增加是由于虚拟校园中模板和内容容器的应用，这使得内容和活动的顺序更具吸引力和直观性，特别是通过 Lesson Builder 自己的工具，它定义了单元模板。在 13/14 和 14/15 的学年中，唯一能够减少在线形式的工具是论坛。其理由是因为根据学术规定，主题内容，而是组织问题，随着前一段讨论的措施的应用，现已得到解决。为了增加论坛的参与度，在 15/16 学年期间，建立了一项新措施，即使用讲座向学生提出挑战的讨论论坛，这导致该工具的使用大幅增加。
- 混合模态在使用 Sakai 工具时表现出与在线模态类似的行为，但论坛工具除外，如图 4 所示。尽管有关材料和模板的相同规定适用于两种模式，但该工具在混合模态中得到更广泛的应用。这可能是因为即使在课堂以外的活动中，这种模式中发生的这么小比例的校园课程也会增加参与。
- 对于校园模态，大多数工具表现出逐渐降低的使用趋势，但有一些例外情况，如图 5 所示。公告和分配工具是唯一具有增加趋势的工具，这可能是由于在发布某些信息时引入发送给学生邮件的自动通知。

根据每种模式中学生的数量比较工具的规范化使用，我们发现在大多数情况下，工具的最大用途是在在线模态中找到，其次是具有相似值的混合模态。在

最后一学年，校园模式的使用率大幅下降。在三种模式中具有相似行为的唯一工具是公告。最后，所有学年和模式最常用的工具是资源。

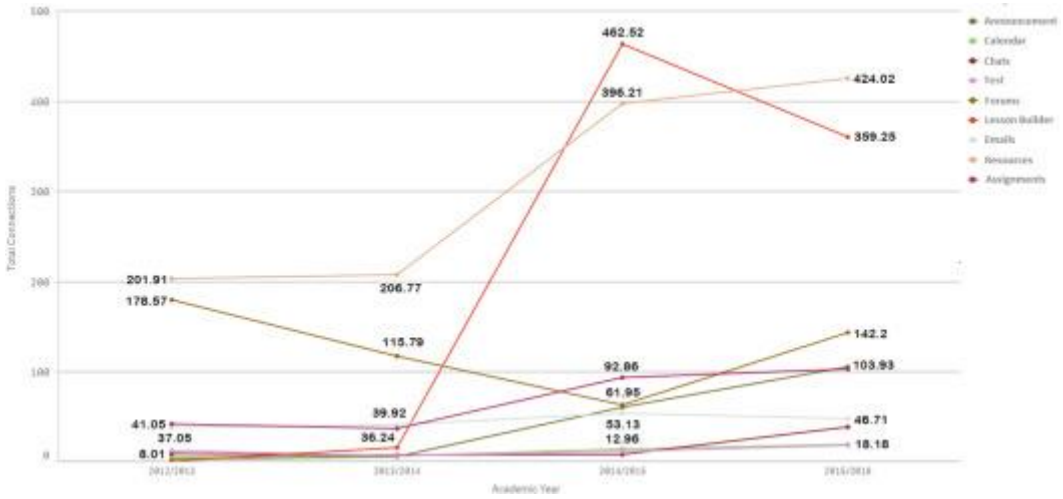

图 3. 关于在线模态中使用 Sakai 工具的学年的演变。

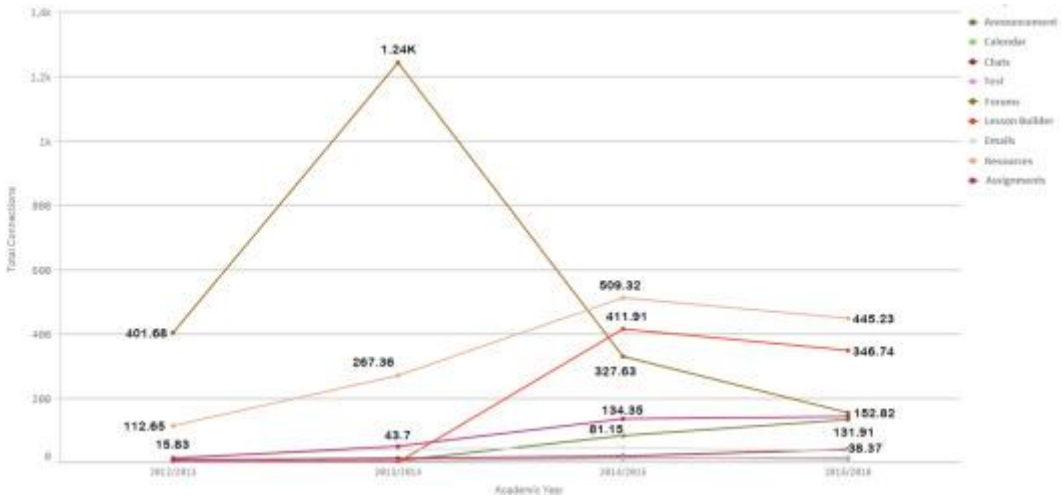

图 4. 关于在混合模态中使用 Sakai 工具的学年的演变。

我们根据每个 Sakai 工具生成的特定事件进一步分析工具的使用。在研究了表 A.6 之后，如附录所示，我们可以添加以下与使用 Sakai 工具相关的要点。

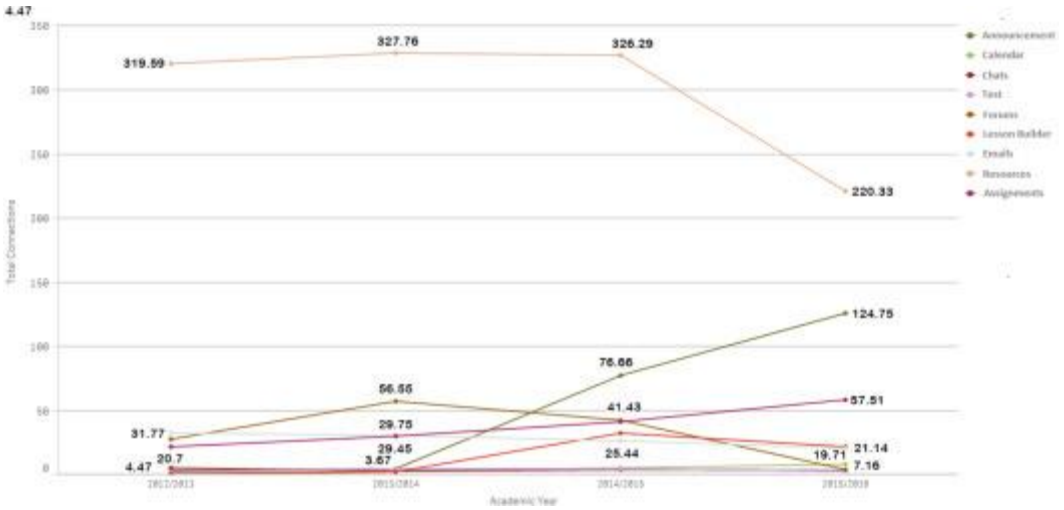

图 5. 关于在校园模式中使用 Sakai 工具的学年的演变。

课程生成器工具未在 12/13 和 13/14 学年实施。在 13/14 年，它被实施为某些等级的试点经验。然而，直到 14/15 学年才开始在大学开设所有学位课程。可以看出，在资源工具之后，课程构建器工具及其事件“访问单元”是最常用的。这个工具得到了学生们的好评，并且课程的教学大纲和内容更清晰，更轻松。对于聊天工具，在 12/13 和 13/14 课程中，学生创建的聊天消息比他们阅读的要多得多。相比之下，在 14/15 和 15/16 期间，他们阅读的聊天消息比他们创建的更多。最后，论坛工具还展示了第一年和第二年之间的差异。在头两年，学生们在论坛上写的比他们阅读的帖子更多。然而，在最后两年，学生们从同龄人那里读到了比他们写的更多答案。

总之，我们必须强调学生的行为变化。在 12/13 和 13/14 的学术课程中，学生们更加活跃，创造了更多的论坛和聊天消息。然而，在 14/15 和 15/16 的学术课程中，学生变得更加被动，阅读更多论坛和聊天消息而不是创建它们。学生的这种被动性与课程生成器工具的实现相吻合。

图 6. Sakai 的 Lesson Builder 工具的学习单元的典型组织示例。

如结果所示，课程生成器工具是最杰出的工具之一，是学生最常用的工具之一。图 6 示出了网络编程主题的学习单元的屏幕截图。为了隐私，某些数据被故意模糊。学生可以查看单元的开始和结束时间表，以及估计的学习时间。此外，这些材料被组织成主要和其他材料，并有视频，论坛和视频会议日期和时间，以及这些材料的链接。

4.2 、活动排名

在本节中，我们将独立于相关工具分析事件排名。图 7，图 8，图 9 示出了根据模态分组的前 10 个事件排名。该 ÿ-axis 表示事件标识符，而 X-axis 显示每个事件的记录数。这些排名是使用 3.2 节中解释的 ad-hoc HiveQL 查询获得的 。这些排名的评论如下。

- 图 7 显示了在线模态的事件排名。此模式中最常见的事件是“下载资源”和“访问单元（课程构建器）”事件; 中等事件是“创建资源”和“自定义网站”; 最后，活动最少的事件是“读取分配”和“读取消息”事件。
- 图 8 显示了校园模态的事件。最常见的事件是“下载资源”和“阅读公告”; 观察到“阅读帖子（论坛）”和“更新个人资料”事件的适度发生; 活动最少的事件是“阅读消息”和“保存草稿（分配）”。
- 图 9 涉及混合模态。最常执行的事件是“下载资源”和“访问单元（课程构建器）”，而“读取分配指令”和“读取通知”被认为是适度的; 活动最少的事件是“阅读聊天”和“自定义网站”。

在分析了这些排名之后，我们发现学生在三种模式中最重复的事件是“下载资源”事件。我们必须记住，Sakai 是一名内容管理员，其基本功能是为学生提供不同的资源类型，这可以通过收集的事件数据得到证实。

图 7. 在线模式的事件排名。

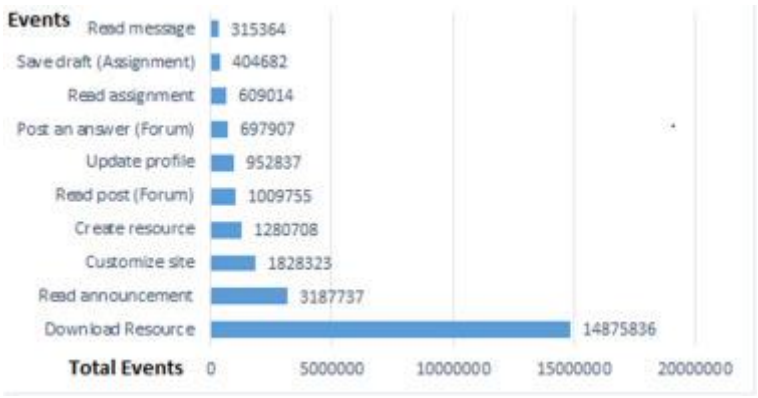

图 8. 校园模态的事件排名。

图 9. 混合模态的事件排名。

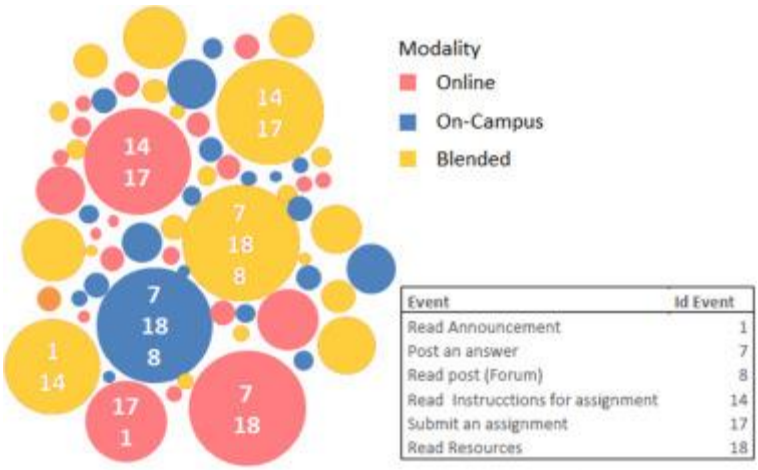

图 10. 学生在同一课程中按照模态分组的事件组合。

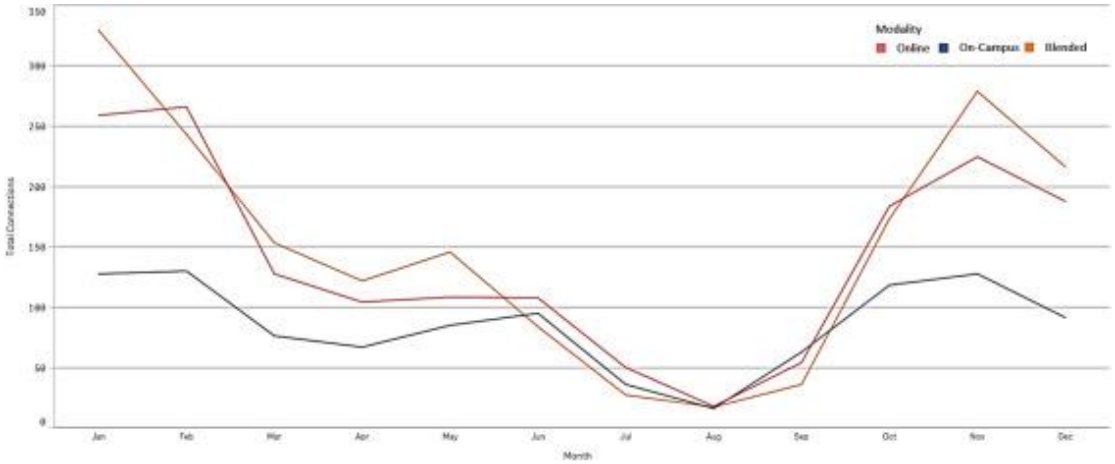

图 11. 每种模式的每月连接趋势。

可以观察到，在线和混合模态在最常用的 Sakai 事件中呈现相同的行为，这是由于使用 Lesson Builder 模板作为两种模态中剩余工具的访问点。此外，应该注意的是，在线模态中的学生不会从私人消息中读取消息在 Sakai 集成的工具。这个事实是合理的，因为这些消息与他们的学术电子邮件相关联，他们可以直接阅读这些电子邮件而无需访问平台。但是，在线模式中“读取分配”事件的低活动性是令人惊讶的。这是因为每个分配的指令都在附加文件中指定。此活动对应于事件“读取分配指令”，它在定义分配时提供了进一步的可能性，例如使用增加的写入空间，将文本与图像集成以及格式化选项等。在其余的方式中，这些说明在课堂上提供。

在下一节（第 4.3 节 ）中，对最常用事件之间的可能关系进行了深入分析。

4.3 、事件关系

通过应用上一节中介绍的 HiveQL 查询和 Apriori 算法，我们获得了几个关联规则，这些规则提供了有关学生行为的信息以及他们在 LMS 平台上执行的事件之间的关系。获得的事件关系考虑了不同的方式。因此，图 10 使用气泡图来描绘每种模态中事件之间关系的概率，如 Apriori 算法所获得的。每种模态分别按红色，黄色和蓝色分组，分别用于在线，混合和校园模式。

在同一会话中执行的事件之间最具代表性的关联的分析如下，可靠性指数高于 70％。

- 对于在线模式：
  - 学生在提交作业之前阅读指令（事件 14）（事件 17）。
  - 由于讲师通过公告提供最终指示，学生在发送任务（事件 17）之前阅读公告（事件 1）。
  - 在论坛（活动 7）中发布答案之前，学生会咨询必要的资源（事件 18）以提交正确的答案。这种行为允许教师在论坛中提出更多问题，以便学生在咨询资源学习时间接学习。
- 对于校园方式：

- 与在线学生一样，校园学生在论坛中发布答案（事件 7）之前访问资源（事件 18）。但是，学生在阅读论坛帖子（活动 8）时也会查看资源（事件 18）。这种行为允许学生使用论坛作为讨论工具间接地咨询和学习资源，尽管他们是在校学生和上课。

• 对于混合模态：

- 同样，混合组中的学生在阅读（事件 8）或在论坛中发布答案（事件 7）之前访问资源（事件 18）。

- 与在线学生一样，混合学生在提交作业之前阅读说明（事件 14）（事件 17）。这种行为是由于在线和混合学生更加细致并且没有在课堂上接受指示，因此他们必须阅读说明以避免在提交作业时出错。

- 学生阅读公告（事件 1），然后阅读任务说明（事件 14）。这是因为，在许多情况下，讲师通过这种方式的公告通知激活未决任务。

在同一会话中执行的事件之间的进一步关系，可靠性指数高于 50%，如下。在线模态学生阅读公告，然后阅读指令。有公告表明新的任务可用，这就是这种关系的原因。对于校园模式，学生阅读帖子，而不是发布答案，他们在论坛中创建一个新线程进行回答。这通常发生在他们回答讲师的问题时，他们在同一个帖子中回答同学。最后，对于混合模态，学生在提交作业之前阅读公告并在课程构建器工具中访问该单元。

除了事件之间的这些关系之外，重要的是要注意校园学生在访问平台时不遵循行为模式。然而，由关联规则确定的使用模式在电子学习平台上的在线和混合模态之间不同。当在线学生访问 LMS 平台时，他们首先使用课程生成器工具访问单元，然后阅读作业说明，完成阅读公告，最后提交作业。但是，一旦混合学生访问了该平台，他们就会开始阅读作业的说明，然后使用 Lesson Builder 工具访问一个单元，以便稍后提交作业。总而言之，论坛是学生一直使用的工具，确定的模式表明他们必须由讲师煽动，因为这为学生提供了学习，咨询资源和了解作业的手段。

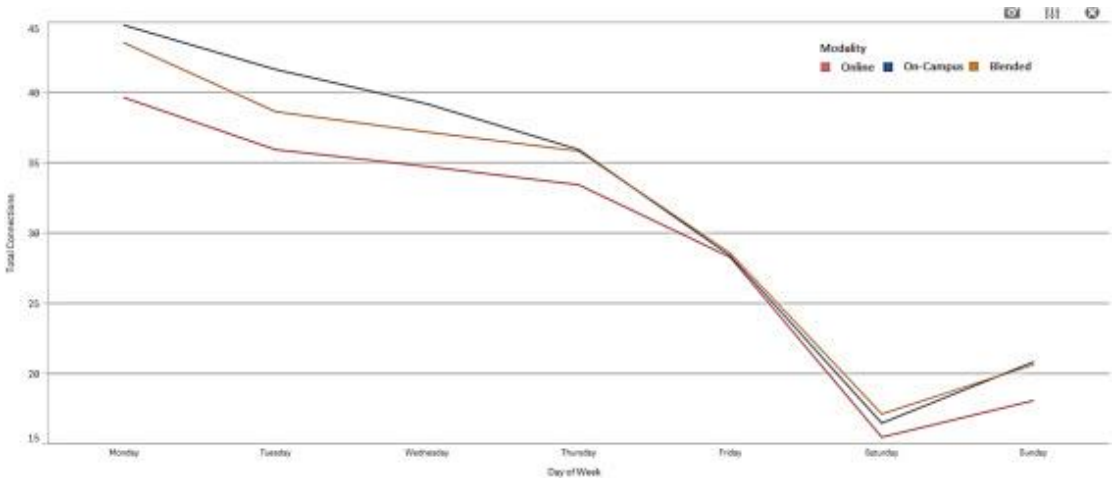

图 12. 每种模态的每周连接趋势。

4.4 、连接趋势

在本节中，我们将根据在所有分析年份中每种模态对 Sakai 的访问次数搜索连接趋势。为了实现这一目标，我们分析了每年和每周与 Sakai 学生人数的关系的趋势，按模态分组，以及每年和每年的平均访问次数。

首先，图 11 说明了学生在研究期间（即 2012/13 至 2015/16）的 Sakai 访问次数的趋势。可以观察到，活动较高的时期对应于 1 月和 2 月，与我们大学的中期考试时间相吻合。请注意，对于 5 月和 6 月的期末考试，与中期相比，活动较少。我们确定了造成这种差异的三个主要原因：首先，需要更多个人和自主工作量以及与讲师联系较少的科目（例如，学士或硕士学位的最终项目）计划在第二学期; 其次，具有外部实习（特别是在健康和教育学位）的科目大多在第二学期; 最后，在较小程度上，辍学学生的数量也会对访问次数产生影响。而且，在图 11，9 月至 11 月的开放月份以及 7 月至 8 月的假期期间可以识别。

其次，图 12 显示了三种方式中一周中每一天的连接数的趋势。从周一开始，整个星期的访问次数持续下降，周六下降幅度较大，周日则略有增加。不同寻常的是，这种模式由三种模式共享，因为在线学生的周末预计会有更高的活动。然而，事实证明，学生更喜欢遵循传统的组织方法，在一周内一点一点地工作，并在周末休息。

最后，在表 4 中，分析了每种模式和学年的平均访问次数。由于此类手段的标准偏差非常高，我们对每位分析的访问总次数的 25%（Q1）和 75%（Q3）之间的访问次数进行了筛选，以便排除学生访问量非常低或很高。然而，2013/14 年度混合模态仍有一个标准偏差非常高的项目。

表 4。按模态和学年分组的 Sakai 的平均访问次数。在每个分析中访问次数的 Q1 和 Q3 之间选择学生以减少标准偏差。

情态		ñ	Q1	Q3	意思	SD
线上	**2012/13**	**236**	**44**	**731**	**215,00**	**172093**
	2013/14	363	45,00	868,00	271,50	215040
	2014/15	860	49,00	1563,00	517,18	441097
	2015/16	1542	33,00	1298,00	420,94	359464
在校园	2012/13	4754	303,00	1009,25	590,57	193253
	2013/14	5138	278,00	854,25	524,05	160786
	2014/15	5899	210.00	904,00	526,88	193451
	2015/16	8106	24,00	438,00	164,48	124033
混合	2012/13	270	326,00	3428,25	1265,63	843372
	2013/14	590	259,50	4649,00	1647,22	1176,263

情态		ñ	Q1	Q3	意思	SD
	2014/15	1777	111.00	2630,50	988,65	700200
	2015/16	2244	16,50	1123,75	364,19	338107

该表中最相关的发现是混合学生的访问次数最多，而在线学生的访问次数相对较少。同样，这表明校园会议的比例较低会导致学生参与度提高。在线学生关注每种模式，从 2013/14 年度到下一年度的访问量增加了一倍。这是因为课程构建器包含在 14/15 课程中，这促使在线学生更频繁地访问虚拟校园。然而，在 15/16 岁的时候，校园和混合学生的就诊次数大幅下降。这是因为改变了该年度不活动时间的流逝，减少了过期会话的次数以及重新连接的必要性。

表 A.5。事件首字母缩略词。

缩写	活动名称
E1	**阅读公告**
E2	新消息聊天
E3	阅读聊天
E4	考试开始了
E5	考试修改
E6	发送考试
E7	发表回答
E8	阅读帖子
E9	更新元素
E10	参观单位
E11	阅读邮件
E12	新文件夹邮件
E13	阅读作业
E14	读取分配指令
E15	保存草稿
E16	下载资源

表 A.6。Sakai LMS 的每个工具生成最多的事件。对于每个学年（行）和每个工具（列），每个模态显示学生最常用的事件，其中'Ex'表示事件，数字表示执行这些事件的时间。对于每个学术课程，模式分别由在线，校内和混合模式的首字母缩写“O”，“C”和“B”表示。名为“Announ”的专栏。指公告工具。带有首字母缩写词的事件名称如表 A.5 所示。

Acd.Year	工具	公告	聊	考试	论坛	课	内部	分配	资源
						生成器	邮件		
2012/2013	**Ø**		**E2**	**E4**	**E7**		**E11**	**E13**	**E16**
			5002	2693	54487		10793	10316	95342
	C		E2	E4	E7		E12	E13	E16
			53977	6887	159317		282814	86550	3467309
	乙		E2	E4	E7		E12	E13	E16
			15065	2075	484938		16091	10088	234625
2013/2014	Ø		E2	E4	E7	E9	E11	E13	E16
			5812	2245	47552	4860	15170	13666	136146
	C		E2	E4	E7	E9	E12	E13	E16
			25576	12292	374865	6237	269721	136365	3886253
	乙		E2	E4	E7		E11	E13	E16
			19737	1192	1165302		34958	34103	430336
2014/2015	Ø	E1	E3	E5	E7	E10	E11	E14	E16
		84409	6276	5084	50167	874681	45564	43731	507730
	C	E1	E3	E5	E8	E10	E11	E14	E16
		1125143	45163	12097	450102	395650	273242	171010	4391931
	乙	E1	E3	E6	E8	E10	E11	E14	E16
		267598	42074	14300	640636	1201058	67086	166321	1526640
二千零十六分之二千零十五	Ø	E1	E3	E5	E8	E10	E11	E14	E16
		8285712	93303	12978	355962	495685	69752	99121	1072763
	C	E1	E3	E5	E8	E10	E12	E14	E16
		2062594	44953	9974	42519	267586	257822	281487	3130343
	乙	E1	E3	E6	E8	E10	E11	E14	E16
		609319	148608	9314	642336	1486544	60919	215965	1826641

5 、结论和未来的工作

本文的主要目标在于从存储在电子学习平台（如 Sakai LMS）中的数据中获取知识。我们已经提议使用大数据技术和框架来尝试获取学生的行为模式，并能够通过改进学习过程来提供结论以提高学生的表现。我们使用其 HDFS 实现

选择了基于 Azure HDInsight 的大数据解决方案。用于从 Sakai 数据库传输数据的工具是 Sqoop，这些数据存储在 Hive [22]数据仓库中。此外，我们按照 Hadoop MapReduce 框架实现了 Apriori 算法以获得 学生在 Sakai LMS 中进行的活动的关联规则。使用这些技术和大数据框架，我们研究和分析了一个包含 70 GB 有关 UCAM 学生行为的信息的数据库，包括有关学位和硕士学位的所有可用数据。已经对所获得的结果进行了讨论，翻译和视觉描绘，以便由与大数据领域无关的人员（例如学位协调员，讲师或学生）容易地解释。

结果表明，所有模式的学生在发布和阅读资源和学术资料之前/之后使用论坛工具。因此，学生通过论坛间接地加强他们的学习过程。这种安排令人惊讶，因为论坛甚至在校园模式下使用，由于课堂上的面对面互动，预计某些工具的使用会更低。然而，即使对于这种形式，这也是对学生学习过程的强化。因此，讲师应该鼓励使用这些并提出额外的挑战来增加和鼓励使用它们。此外，课程生成器工具，特别是其事件“访问单元”，是“资源下载”事件之后执行最多的工具。因此，讲师应继续使用此工具提供的模板，因为学生会发现内容更清晰，更易于使用。这一事实表明，LMS 中的内容组织，无论是使用 Sakai 课程构建器还是其他平台中的类似工具，都可能成为促进学生参与的关键因素。最后，混合的学生更频繁地使用 LMS Sakai，展示更多的访问，从而进行更激烈的学习过程。这个结果令人惊讶，因为 展示更多的访问，从而进行更激烈的学习过程。这个结果令人惊讶，因为 展示更多的访问，从而进行更激烈的学习过程。这个结果令人惊讶，因为在线学生应该最频繁地使用该工具，因为他们没有机会参加面对面的课程。

本文提出的框架可用于该领域的进一步研究; 例如，也要研究讲师的行为模式。它还可以用于其他领域，例如智能汽车，以识别良好（或差）驾驶员行为，或智能家庭来研究居民的能量使用。

这项工作的一个直接未来路线是确定学生行为模式与其成绩之间可能存在的相关性，以便识别和促进有助于提高学生资格的行为。此外，我们正在研究为什么某些工具在培训模式中比其他工具更容易接受的原因，以便以较低的接受程度升级这些工具。

致谢

这项工作得到西班牙 MINECO，西班牙的支持 ，授予 TIN2016-78799-P（AEI / FEDER，UE）。作者要感谢本大学在线部门的成员参与本文。他们还要感谢参与该研究的学位协调员，讲师和学生。

附录。

请参阅表 A.5，表 A.6。

参考文献

[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.2009.06.011.

[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, EDUCAUSE Rev. 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 Gener. Comput. Syst. 68 (2017) 31–37, 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.campuscomputing.net/. (Accessed 15 January 2017).

[7] C. Romero, S. Ventura, Educational data mining: a review of the state of the art, IEEE Trans. Syst. Man Cybern. C 40 (6) (2010) 601–618.

[8] C. Romero, S. Ventura, E. García, Data mining in course management systems: Moodle case study and tutorial, Comput. Educ. 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: 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, Oper. Res. 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, 2012, National Training System Association Orlando.

[12] J.-L. Hung, Y.-C. Hsu, K. Rice, Integrating data mining in program evaluation of K-12 online education., Educ. Technol. Soc. 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 J. Soft Comput. 5 (4) (2015) http://dx.doi.org/10.21917/ijsc.2015.0145.

[14] B. Tulasi, Significance of big data and analytics in higher education, Int. J.

Comput. Appl. 68 (14) (2013).
[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/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.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, 2012,
[18] A.G. Picciano, The evolution of big data and learning analytics in american higher education., J. Asynchronous Learning Netw. 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 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 case study, Cutter Consortium White Paper 13 (8) (2013).
[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.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, Proc. VLDB Endow. 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 J. Sci. Technol. 9 (3) (2016).
[27] S.K. Verma, R. Thakur, S. Jaloree, Pattern mining approach to categorization of students' performance using apriori algorithm, Int. J. Comput. Appl. 121 (5) (2015).

[28] S. Singh, R. Garg, P. Mishra, Review of apriori based algorithms on mapreduce framework, arXiv preprint arXiv:1702.06284, 2017.

[29] J. Dean, S. Ghemawat, MapReduce: simplified data processing on large clusters, Commun. 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

Magdalena Cantabella 于 2008 年获得穆尔西亚天主教大学计算机科学学士学位，2012 年获得穆尔西亚大学生物医学专业计算机科学硕士学位。2010 年以来，她是该系理工学院的副教授。穆尔西亚天主教大学计算机工程学位。她的研究领域包括大量的数据统计分析，电子学习和用户档案的定义。

RaquelMartínez-España 是西班牙穆尔西亚天主教大学（UCAM）技术学院的副教授。她于 2009 年获得计算机科学硕士学位，并获得博士学位。2014 年在穆尔西亚大学获得计算机科学博士学位。她曾参与过人工智能和教育方面的多个研究项目。拉奎尔参与了各种学术和工业项目。他的研究兴趣包括数据挖掘，大数据，软计算，人工智能和智能数据分析。

BelénLópezAyuso 在穆尔西亚大学获得计算机科学硕士学位，并获得博士学位。在同一所大学的计算机科学专业。她拥有 18 年的教学经验，包括大学水平的学位和硕士课程，包括电子学习方法。她参与了几个教育创新项目，从中获得了教育创新领域的出版物。目前，她是穆尔西亚天主教大学计算机工程学院院长和该大学网络系热学院院长。她的研究领域包括教学评估和电子学习方法评估。

JuanAntonioYáñez 于 2015 年在穆尔西亚天主教大学获得计算机科学学士学位，目前在一家技术公司担任计算机顾问，并开始在研究领域进行博士研究，包括大量的数据统计分析。

AndrésMuñoz 是西班牙穆尔西亚天主教大学（UCAM）技术学院的高级讲师。他获得了博士学位。2011 年在穆尔西亚大学获得计算机科学学士学位。他曾参与过多项人工智能和教育研究项目。他的主要研究兴趣包括智能系统中的论证，语义 Web 技术以及应用于教育的环境智能和智能环境。