基于 moodle 日志的教育数据可视化技术*

黄 镭[1] 吴宙华

（1．广西广播电视大学 教学资源中心，南宁 530021）

摘要：随着国家开放大学逐渐将网络课程迁移到基于 moodle 的学习平台，越来越多的学习者在平台上浏览资源、下载课件、完成在线测试、编辑 wiki 和参与互动，并由此生成了海量的日志数据。该研究以国家开放大学广西分平台的几门统设课程为例，主要利用数据可视化和 web 日志挖掘的方法，借助 moodle 的日志数据，对网络课程的使用情况进行画像，由此来掌握一门上线课程的运行状况，以及学生的参与度，并促进教学管理决策和课程资源的持续改进。

关键词：国家开放大学；moodle；web 日志挖掘；数据可视化

【中图分类号】G40-057【文献标识码】A【论文编号】【DOI】

一 研究背景

大数据为在线教育实现个性化教学、重新构建教学评价方式和提升教学管理水平提供了可能。为了分析学习者在学习平台上的学习活动所产生的海量的记录，挖掘出有用的信息，许多学者进行了大量的研究，诞生了教育数据挖掘（Educational Data Mining）和学习分析（Learning Analysis）等研究领域。其中，教育数据可视化（Educational Data Visualization）作为教育数据挖掘的一种应用类型，在协助学习者自我评估，指导优化教学和为教学管理提供决策等方面提供了非常直观的依据。因此，在许多主流的商业学习平台如（Coursera、Blackboard、Sakai、清华学堂在线……）以及开源学习平台如 moodle 中均提供了所谓“学习仪表盘”或者学习分析报表这样的工具，用于将学习平台上的统计与分析可视化。然而，由于教育问题的复杂性和具体性，对数据报表的要求呈现出多样性和动态性等特点，而平台原有的数据报表系统往往并不能完全满足各高校开展教学检查和动态监控教学质量的要求；同时，随着国家开放大学将网络课程逐渐迁移到 moodle 平台上来，截至目前为止，已经有大量的开放教育学习者得以在平台上开展学习，如何将海量的访问日志转化为“会说话”的图表，改进 moodle 的报表功能，是本文的研究内容。

二 教育数据可视化及其研究现状

数据可视化技术是指以人作为分析主体和需求主体，强调基于人机交互的、符合人的认知规律的分析方法，意图将人所具备的、机器并不擅长的认知能力融入分析过程中，以获得对于大规模复杂数据集的洞察力(insight)[1]。

Ben Fry[2]根据数据可视化的最新发展与需求，提出可视化流程由 (1)获取、(2)分析、(3)过滤、(4) 挖掘、(5)表述、(6)修饰、(7)交互共七个阶段组成。其中(1)～(3)点描述的比较宽泛，针对诸如 web 日志挖掘等特定研究领域，可以参考数据挖掘中数据预处理的手段和方法来进行。

教育数据可视化，研究的对象主要包括：教学资源的内容情况，以及学习者学习过程中产生的行为数据。教育数据可视化的作用，从教师角度来说，可以直观获得有关学生的学习绩效、学习过程以及学习环境的信息，这些信息可以为教师设计和改进课程、为学生提供教学支持提供依据；从学生角度，可以促进其知识的内化与自我评估；从教育管理者角度来说，对学校教学过程数据进行可视化，就有利于合理考核教师绩效，统筹教学资源规划。国家开放大学每年要组织对地方电大进行日常的网上教学检查工作，以此作为评价地方电大进行网上教学的重要指标，这时课程数据的统计与可视化处理就起到了重要的作用。

教育数据可视化是一个包含了设计与美学、教育学和教学法理论、计算机和 web 数据挖掘算法的交叉学科，往深度来递进，可以结合大数据挖掘的理论和算法，揭示更复杂层次的数据间的关系，往广度来说，可以为学习者的自我评价、教学者的教学优化和教学管理者的管理自动化提供帮助。

目前，国内外在教育数据可视化领域随着教育数据挖掘等领域的兴起，正在开始走向研究热点。在一些研究中，这种教育数据可视化方法也被称为信息仪表盘、学习仪表盘。比如，围绕 moodle 平台的数据可视化这方面的研究，较早的有 GISMO[3]；Calvani[4]等人则利用雷达图对 moodle 平台上课程论坛中交互式讨论的学习有效性进行建模和可视化；Podgorelec[5]等开发了针对 moodle 平台的“信息仪表盘”模块，可以看到课程的访问情况，各模块的点击量，以及教师和学生的各自登录情况等；Aguilar[6]等利用社交网络图、蛇形图和词云三种数据可视化方式对 moodle 平台上的用户关系进行刻画，突出显示围绕某个关键词进行讨论的社交网络图谱。Mazza[7]等在 GISMO 的基础上开发了 Moclog 集成模块，在教师和学生两个角度，提出了各自的教学／学习参与度的多个方面的指标，并根据这些指标形成了可视化报表。国内目前开展这方面的研究还较少，张振虹[8]等人对学习仪表盘的概念和应用进行了一些介绍；张金磊等人[9]对数据可视化的概念做了介绍，并从受众角度指出了数据可视化对远程教育的推动作用。一些其他的研究则着眼于利用 SPSS／WEKA／UCINET／EXCEL 等现成的统计／挖掘软件的功能对平台日志进行分析，并以软件的可视化功能作为结论形式，并未单独开发可视化工具，毫无疑问失却了一定的灵活性[13][14]。

目前对教育数据可视化的研究大多集中在特定平台（比如 moodle，或者商业化集成环境）之中，并借助一些开源的平台无关的可视化工具。比如商业上，Coursera（https://zh.coursera.org/）、清华学堂（http://www.xuetangx.com/）等学习平台均提供了类似于学习仪表盘这样的模块，在访问统计等方面均有相应的可视化应用；作为服务器日志分析和可视化的开源方案，AWStats(http://www.awstats.org/)更多的是着重于平台统计数据的可视化；另外一个针对 web 服务器日

志的开源分析工具 visitors(http://www.hping.org/visitors/)则主要是针对平台超文本链接的访问分析，以及点击流的网络可视化，没有直接针对学习对象的统计分析可视化功能。

本研究也是立足于 moodle 平台，但是相对于前述大多数研究，本方案独立于 moodle 平台，开发了一个基于 Python 的，从日志预处理、指标客制化，到统计可视化的开源 moodle 日志分析工具，并利用该工具完成对数门广西广播电视大学的 moodle 课程的画像和分析工作。

三 研究案例与方法改进

本研究选择以国家开放大学 moodle 学习平台广西分平台上的统设课程为研究对象，通过 moodle 的日志报表数据，经过数据预处理和数据可视化，将平台的教学情况以更为直观的、个性化的呈现出来。可视化过程按照流程顺序分为如下几个步骤：

1 数据预处理

一条典型的 moodle 日志格式（原国开平台日志格式，现已更新）如下所示：

16002315 "2016 年 06 月 30 日 18:12" 10.111.2.40 吴宙华老师 "course report log (http://guangxi.ouchn.cn/report/log/index.php?id=5106)" 16 春审计学

可以看到，moodle 的访问日志包含了访问时间、访问 IP 地址、用户名等，其中双引号包围的内容，在 moodle 日志中称为事件（event），包括事件名和括号中的 URL，最后是描述（description）。比如在上例中，日志时间就是“2016 年 06 月 30 日 18:12”，IP 地址就是“10.111.2.40”，用户名就是“吴宙华老师”，事件名就是“course report log”，URL 就是“http://guangxi.ouchn.cn/report/log/index.php?id=5106”，描述就是“16 春审计学”。

一般的 web 日志挖掘均是针对标准的 NCSA 扩展格式（ECLF 格式），这种日志一般来自比如 apache 服务器，moodle 日志与其在格式上稍微有一些区别，但是在内容来源上则有较大差别，因为 web 服务器上的日志不但包含了 web 的页面访问，连同页面内嵌的 javascript 代码请求、图片、字体下载等请求都包含在内，而这些则是进行 web 数据挖掘的时候需要过滤的数据，因此，针对 moodle 日志进行的数据预处理过程，要比直接针对 web 服务器日志的数据预处理过程简单。实际上，moodle 日志就是来源于服务器访问日志，moodle 内部进行了过滤和包装重组。在实际实践中，针对 moodle 日志进行的数据预处理，既有方便之处，又有不便之处。

方便之处在于，网站的访问行为的记录粒度得到了提升，略过了无关信息的访问记录，省去了数据筛选的过程，并且，由于 moodle 日志包含了用户名，避免了预处理过程中的用户识别阶段的工作；不便之处则在于，moodle 日志报表的时间格式只精确到分钟，因此，在需要挖掘的指标中包含用户在线时间和会话信息等对象的时候，需要额外的处理。下面详细说明整个 moodle 日志的数据预处理过程。

通常，Web 日志数据预处理过程包括数据清理、用户识别、会话识别、事务识别、路径补充等几个步骤。

数据清理（Data Cleansing）

是数据预处理工作的基础,在数据挖掘过程中也起着至关重要的作用。数据清理的任务是要把 Web 日志中和挖掘目的无关的数据项清除,并把对挖掘目的有用的数据转换成数据挖掘需要的格式。由于直接使用 moodle 日志，避免了无关数据筛选，但是还需要转换格式。比如，web 日志中的访问时间只能精确到分钟，在没有服务器日志读取权限的情况下，一分钟内的访问，其日志中的时间戳都是一样的。这样在统计访问次数的时候不会有什么问题，但是在统计会话内在线时长的时候，就会出问题。一种解决办法是记录次数，但是时间长度近似为 0，这样会丢失一些精度，若日志中存在大量不足 1 分钟但是接近 1 分钟的会话内的连续访问（实践表明这种情况是很常见的），则误差会进一步加大；第二种近似办法是找到会话内连续相同时间后面紧跟着的第一个不同时间，然后从连续相同时间的第一次访问开始，把访问时间平均分配到这段时间内，意即修改这一段相同时间的时间戳，使其变成精确到秒的近似等间距时间（因为不一定能够整除，完全整除也没有意义）；退一步讲，若后续数据分析工作不关注会话内的细节，则不必处理。

用户识别（User Identification）

moodle 日志里面已经带有用户信息，不必进行用户识别。

会话识别 （Session Identification）

会话是指同一用户连续请求页面的集合,不同用户访问的页面属于不同的会话。经过上一步用户识别出来之后,就要把每个用户在一段时间内的所有的请求页面分解为单个的会话。最简单的方法莫过于,也就是一个时间间隔,如果用户访问的时间差超过了一个 Timeout ,则认为用户开始了一个新的会话,通常 Timeout 默认的时间阈值为 30 分钟,很多学者用各种不同的算法去计算这个阈值,以达到与现实会话更加接近，一些研究认为比较合理的时间阈值是 25. 5 分钟[10]。例如用户 A 向服务器发出 5 个请求，如果前面 4 个请求时间相近，后面一个比第 4 个晚了一个时间戳，则将访问记录分为 2 个会话。除了时间戳(Timeout)的方法之外，还有其他方法。本实验采用 30 分钟作为会话的 Timeout 值。

事务识别(Transaction Identification)

事务识别是建立在用户会话识别的基础上的，把会话进一步分成具有一定语义的事务，其目的是依据数据挖掘任务的需求将事务做分割或合并处理，使预处理结果更加精确，适合于数据挖掘需求。本实验处理的可视化对象只需要会话级别的访问粒度，因此无事务识别的需要。

路径补充(Path Supplement/Completion)

在会话识别过程中的另一个问题是确定访问日志中是否有重要的请求没有被记录。由于服务器在访问过程中采用缓存机制，这就使得在访问特定资源的请求路径上会丢失一些引用信息，导致用户的访问过程没有被 Web 日志完整地记录，比如大多数访问路径的浏览器的“后退”操作没有被记录在日志中。路径补充的任务就是将这些遗漏的请求补充到。由于本实验不涉及到点击流分析，因此不需要进行路径补充。

2 指标选取的经验以及改进

在基于数据可视化技术的学习分析研究里面，要始终明确可视化形式只是手段，核心还是指标的选择和开发，因此，在本研究中，既借鉴性的采纳了一些目前常规的指标，也创造性的开发出了一些更能表征学习状态的指标族。

本可视化课程报告具体分为 4 组模块，共 9 个数据可视化对象，图像全部是基于 python 开源图形库 matplotlib 编程绘制。

模块 1 平台访问量可视化

从日志中最容易得出的统计就是按照某个时间单位汇总得到的用户访问情况。以下图 1 就是《国家开放大学学习指南》课程广西 moodle 平台上的各月访问量统计柱状图，用以表征课程总体访问情况。

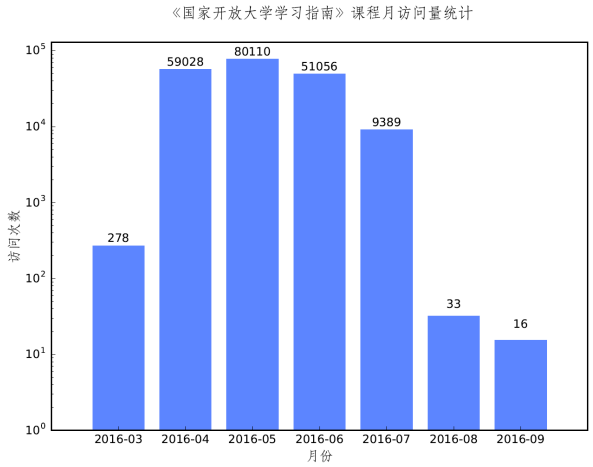

图 1《国家开放大学学习指南》课程各月访问量统计柱状图

统计时段选取 2016 春季学期，跨度为 7 个月。由于 2016 年春季学期该课程的访问量各月访问量差别过大，只有对数阶才能看清楚其他月份的访问情况。我们可以看到平台上这门课程的月访问量主要集中在 2016 年 4 月～6 月三个月，其后的访问量下降的很快。这里可以看出来，课程在开学的前三个月，由于班主任的推动，以及开学典礼上的动员和宣讲工作，平台人气最高，访问量是很可观的。8~9 月份属于暑假期间，访问量可以忽略。月度访问量只是一个比较整体的指标，一些更细致的统计数据还需要进一步获取和分析。

比如在时间上，为了获得更有趣的观察角度，需要统计一天 24 小时各个时段里面的课程访问情况。如果以月份和时间间隔作为 x 轴和 y 轴，要想统计用户的访问量就至少需要三个坐标。有一种称为热力图的可视化形式，利用人眼对颜色的感知能力，将第三维数据以颜色深浅或者色温来表示，是一种将三维数据在二维尺度上直观表达的重要手段。以下就是各月份/各时段访问量统计热力图 ，用以表征课程用户的学习时间模式，如图 2 所示。

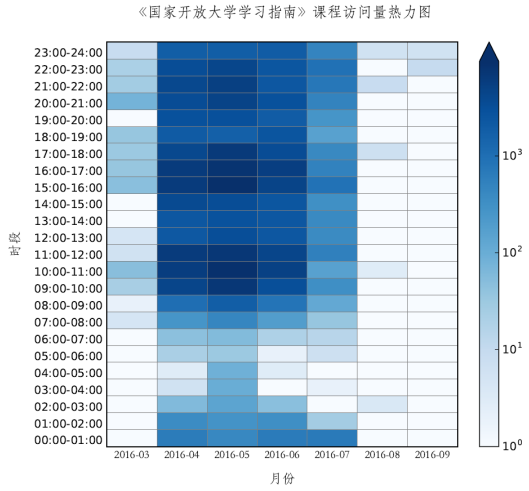

图 2《国家开放大学学习指南》各月份/各时段访问量统计热力图

通过图 2，我们可以很直观的看到前面月度访问量柱状图所不能表达出来的信息。比如，我们可以看到，不同的月份，一天之中大致存在三个访问高峰期，这三个访问峰值主要集中在早上的 9 点到 12 点，下午 3 点到 6 点，以及晚上的 8 点到 11 点之间。进一步的，为了掌握星期各天的登录是否存在某种模式，可以绘制各星期/各时段访问量统计热力图，用以表征课程用户的学习时间模式，如图 3 所示。

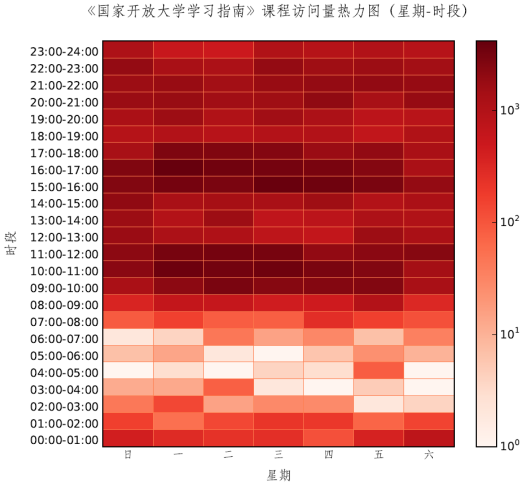

图 3《国家开放大学学习指南》各星期/各时段访问量统计热力图

通过图 3，我们可以看到从星期一到星期日，用户的白天学习时间并没有太大的区别，基本上还是集中在图 2 所显示的三个峰值区间附近，但是第三个峰值的结束时间随着周末的临近逐渐向更早的时间推移。图 3 相对于图 2，补充揭示了学习者夜间自学时间的一种规律。

魏顺平[11]曾经提到，在线学习平台中普遍存在学习者访问情况的无尺度现象，或者说统计量具有幂律（power law）分布的特点，简单来说，就是存在极少量高访问数的访问者和大量低访问数学习者的情况。本文用另外一种数据可视化方式对该研究结论进行了验证。通过将访问数以 50 为一个等级，对课程访问用户进行了划分，绘制了如下饼图，如图 4 所示。

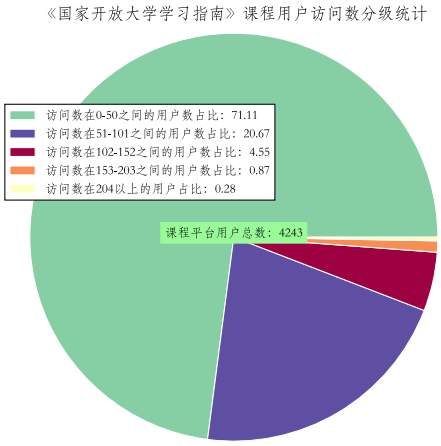

图 4《国家开放大学学习指南》课程用户访问数分级统计

从图 4 可以看到，访问量低于 50 的用户占了超过 2/3 的全部平台访问用户数。访问量在 200 以上的用户占平台总访问用户数量的 1%不到。考虑迄今为止我们的数据统计均是跨度为 7 个月的时间周期，这意味着大部分用户月平均访问量只有 7 次左右。以上均是对访问量进行统计得到的结果，接下来研究用户的在线模式。

模块 2 平台用户情况

考察平台用户的在线情况，除了统计数量以外，还需要进一步跟踪用户访问情况在时间上的动态变化。要想得出这种动态情况，不妨设定三个集合：沉睡用户集 S1、活跃用户集 S2、新增用户集 S3，分别代表当日往前一周以内无登录的历史用户集合、当日往前一周内访问过课程的用户集合以及当日往前一周内的新增用户集合。设当前日期往前 1 周统计不同在线用户集合为 W，迄今为止平台历史用户集合为 D，则集合序列 S1、S2、S3 的计算满足下列关系：

初始化：

第一个时间点：

S1=0; D=S3=S2=W

此后每日依次计算：

S2=W;

S1=U-S2

S3=S2-U

D=D∪S2

上述“-”、“∪”运算为集合运算符，其中等号表示赋值计算。通过获取每一天序列 S1/S2/S3 的集合大小，可以得到沉睡用户、活跃用户和新增用户数量上的动态变化情况，以上述关系绘制 S1、S2 和 S3 的点线图，其中 x 轴为日期，y 轴为周用户数，同样以《国家开放大学学习指南》课程为例，如图 5 所示。

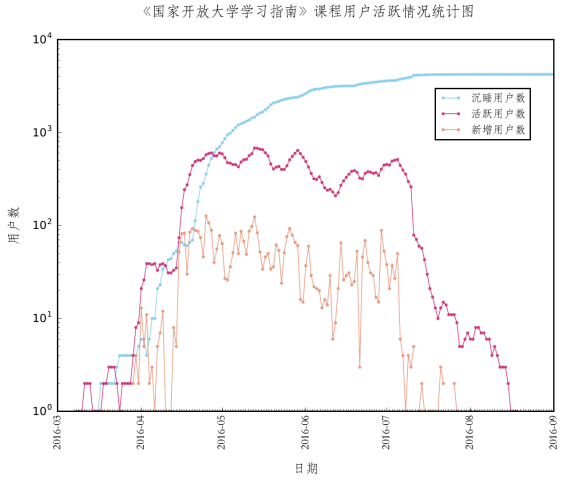

图 5《国家开放大学学习指南》课程用户活跃情况统计图

目前为止的所有统计图形均只考虑平台访问用户，不考虑选修了该课程但是从未登录过的用户。我们可以很容易证明，未登录过的用户对上述图线形态均无任何影响，比如对图 5 来说，若设从未登陆过用户为 S0，因为逻辑上来说，无法从过往历史来判断某用户是否永不登陆，但是在期末的时候，可以把该 S0 加入计算。则可以设初始化的时候 S1=S0。由前述关系可知，S0 的加入不影响 S3 和 S2 曲线，对 S1 的曲线形态也并无影响，因此 S0 仅作为图线在 y 轴上的偏置而已。

考察图 5，可以发现 2016 年 4 月份期间，平台新增用户数量、活跃用户数和沉睡用户数都在同步上升。其中 2016 年 4 月上旬的时候，沉睡用户数甚至有了一次下降，伴随着沉睡用户数的这次下降，几乎同时期新增用户降至零，说明沉睡用户中的一部分从沉睡中“复苏”过来，重新访问了网络课程；从 2016 年 4 月份下旬以后开始到 2016 年 5 月，活跃用户和沉睡用户的数量同时急剧上升，而新增用户也同时达到其最高点，标志了平台的第一次访问高峰，其后活跃用户和新增用户均震荡中缓慢下降，同时沉睡用户急剧上升。这个阶段的数据因为对数尺度得以放大细节，实际上用户在线情况的震动是非常大的：可以看到周新增用户数主要是在 100 左右震荡，而周活跃用户数在 500 人左右震荡，两者波动情况大致相符，说明这一时间段内，用户的重复访问情况是十分普遍的；其后 2016 年 7 月份以后，沉睡用户数量基本上维持不变，活跃用户数和沉睡用户数曲线大幅下降。从上述分析中，我们可以得出这样的印象，就是开学第一个月是吸收用户最为重要的时间点，这个时间段内无论是活跃用户还是新增用户的数量增长是十分迅速的，后面还会有关于在线时间的分析。

Moodle 的日志里面还包含了用户的 IP 地址信息，为了对课程的用户来源有一个直观的现示，对用户的 IP 地址所属的地理位置进行了统计，并选择词云作为这个统计的可视化表现形式。由于国开的 moodle 平台基于云计算平台，为了优化各个省级分平台的访问，存在众多 CDN 加速节点，加上互联网上代理的广泛使用，仅仅通过平台访问的 IP 数据，是无法得到非常精确的数据的，但是根据不同地域的访问频度，仍然可以获得课程在不同地理位置的访问情况，以《中国特色社会主义体系概论》这门课程为例，绘制的不同地域用户来源的词云，可以以访问频度为权值，给出不同地理位置的缩放显示，如图 6 所示。我们可以一目了然的看出，这门课程的学习者主要来自于省内的南宁市、防城港市、桂林市、钦州市和贵港市。通过这种图示，我们可以比较明了的知道各个地市级电大在课程的推广上的力度，更加有的放矢的督促分校在课程宣传推广上投入更多的精力。由于课程的内网访问量较大，而内网用户主要来自于省级电大教师登录，为了滤除这部分影响，去掉了内网访问的统计。

图 6《中国特色社会主义体系概论》不同地域用户来源词云

模块 3 主题化的课程情况描述

所谓主题化的课程情况描述，就是不再以直接的统计量作为观测对象，而是设计出一种复合式的指标，这一套体系就称为主题。

在数据可视化里面，有一种常用的表现形式，叫做雷达图，或称为蜘蛛图。雷达图本质上就是以每个标量一个轴，将多个标量共同绘制在共一个原点的图表之中，然后不同的统计样本的各维分量的比较可以以覆盖范围来看出指标的优劣。雷达图可以分为等量纲的和非等量纲的两种。等量纲的雷达图，各个标量分量的单位和刻度都是一致的；而非等量纲的雷达图，各个轴都有各自的标度，由于雷达图各个轴讲究角度等分，长度相等，因此非等量纲的雷达图，还需要将不同的标量的范围进行映射，映射到同一个均匀的范围内。

参考国内著名慕课平台清华学堂在线的指标，以及 Lehmann 等[12]所提到的互联网用户的参与度模型，本文提出以课程学习热度和课程交互度两个方面，两个方面各 3 个指标，绘制雷达图，作为课程健康度主题。接下来对该主题的这几个指标进行说明。

课程学习热度：

（1）课程用户活跃度=访问次数在 30 以上用户数/历史在线用户数 — 这一比例表征活跃用户数的度量；

（2） 人均课程使用时间=历史用户总在线时长/历史在线用户数 — 表征课程的使用情况；

（3）人均论坛访问量=论坛总访问量/历史在线用户数 — 表征论坛的总体使用情况；

课程交互度：

（4）（发帖+回帖）总数 — 统计课程论坛中用户生成内容（User Generated Content, UGC）情况；

（5）论坛发帖回复率=论坛回帖数/论坛主题数；

（6）论坛生命力=贡献内容（发帖回帖）的用户数/曾访问论坛的全部用户数 — 这个指标主要是表征论坛有意愿产生交流的用户的覆盖情况，因为论坛的生命力主要是来自用户对论坛内容的持续贡献。

选择《中国特色社会主义体系概论》和《国家开放大学学习指南》这两门课程作为对比，以上述指标绘制雷达图，如图 7 所示。

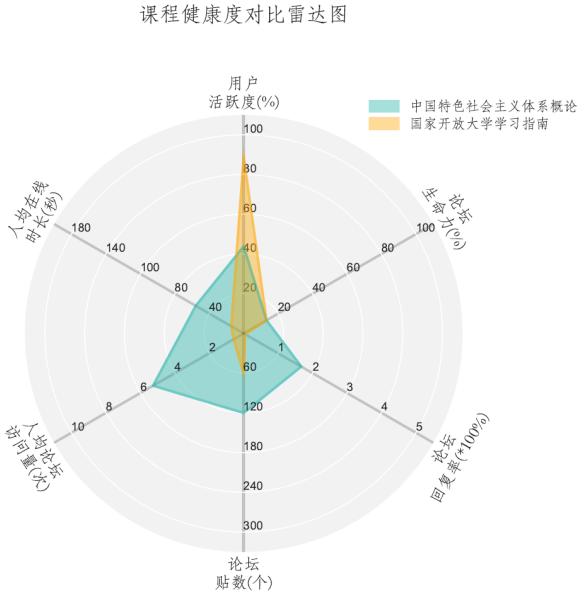

图 7 课程健康度对比雷达图

对于指标的选择，整体上来说，由于仅仅通过日志无法得到课程的选修学生数量，只能得到访问平台的情况，因此去除掉比如课程总人数等统计量。由于课程总选修人数有多有少，选修课程人数多的课程的访问人数也会较大，因此指标选取的时候要尽可能的去除这类因素，尽可能选择与总量相关度较小的特征统计量。当然，若在生成图表之前，可以获得几个对比课程的全部数据，也可以做一个正规化处理消除这类影响因素。

而在个别指标的挑选上，本文参考了清华学堂在线的一部分指标，以及 Lehmann 等人所提出的指标，最终选择了这六个指标，主要也是考虑了这六个指标相互之间的互补性，并且去掉一些与课程规模相关的指标，比如清华学堂在线里面使用的课程选课人数，以及 Lehmann 论文中使用的总访问量等指标；借鉴了 Lehmann 论文中关于用户停驻时间的指标，使用人均在线时间来刻画课程的在线情况；创造性的提出了一些表征课程和论坛使用情况的指标（用户“活跃度”、论坛“生命力”、回复率等）。这些指标的选择最终证明对于分析和对比课程的在线教学情况是非常有利的。

比如，从图 7 我们可以看出来，《国家开放大学学习指南》课程在用户活跃度上面比《中国特色社会主义理论体系概论》要高，说明访问量高于 30 的用户数占比前者要高于后者，而人均在线时长上，前者却远低于后者。从这两个指标，可以很容易看出来，《国家开放大学学习指南》这门课程的人均课程使用程度并不比《中国特色社会主义理论体系概论》要好。我们进一步可以从人均论坛访问量的差异得出这样的结论，《中国特色社会主义理论体系概论》的人均在线时长主要得益于人均论坛访问次数，说明这门课程的学生在课程论坛上花了更多的学习时间，同时，从论坛帖数和回复率我们可以看出，《中国特色社会主义理论体系概论》比较于《国家开放大学学习指南》在帖数和回复率上均高于后者。最后，我们在论坛生命力这个指标中可以看到，两门课程的论坛核心用户占比是比较接近的。通过对比，很自然得出结论，就是《国家开放大学学习指南》这门课程若是任课教师能在课程论坛上多为学习者提供更多的帮助，对于提升课程的健康度，会比较有帮助。

第二个主题主要是课程各模块使用情况监控。

这个主题是监控同一门课程的不同模块访问情况。moodle 平台上课程的模块主要是按照 moodle 的资源类型进行区分，moodle 的内建类型大致可以分为论坛（forum）、课程（course）、作业（assignment）、资源浏览（view）、测验（quiz）以及其他模块，不同的课程，在资源的比重和资源的信息量上面是不一样的，因此，可以将访问量前几的模块单列绘制曲线，剩余模块合计绘制曲线。值得注意的是，如果在创建课程的时候，对课程的描述更为准确的话，是可以按照课程的设计逻辑，而不是按照课程的模块类型来统计的。按照课程的设计逻辑来统计的话，采用的日志字段就是

“描述信息”字段，而不是“事件”字段了。关于 moodle 的各模块含义和描述，见 moodle 官网关于新的事件系统的文档（https://docs.moodle.org/dev/Event_2），这里不再赘述。以《国家开放大学学习指南》课程为例，根据各模块的在线时长绘制的课程各模块/整体用户日平均在线时长曲线如图 8 所示。

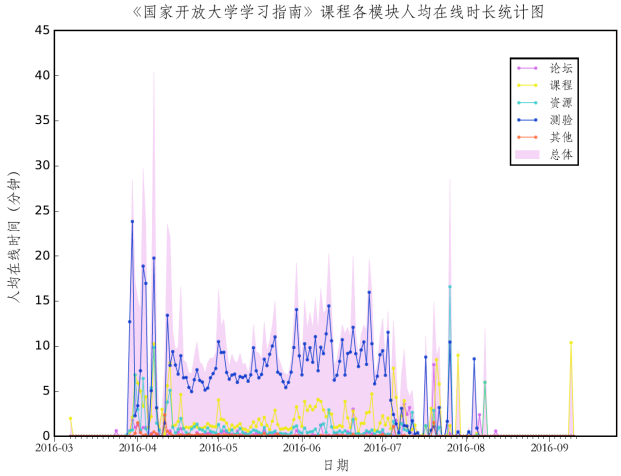

图 8 课程各模块/整体用户日平均在线时长曲线

从图 8 可以看出，课程的测验模块访问情况在大部分时间里面都是占据最大的时间比例，接下来是课程模块。说明学习者在该课程中主要花时间在完成在线测试之中。通过对课程设计的实际考察，我们可以认为学生的学习行为与课程的设计与考核要求基本契合。

模块 4 不同课程间的横向对比

为了让不同课程得到更好的对比，将不同的课程放到一起进行对比是非常有必要的，图 9 是广西分平台上访问量排在前五位的课程在 2016 年春季学期的学生在线情况对比统计图，其中上方的图是人均在线时长的曲线，下方的图是课程在线用户数的统计。从整体上，可以看出《国家开放大学学习指南》课程的日在线人数与其他课程相比高很多，但是其日人均在线时长与其他课程相差无几，大约为 20 分钟；《审计学》课程在 2016 年 6 月下旬人均在线时长较为突出，但是在线人数不到 10 人，因此只有少数用户对在线时间做出了贡献；《中国特色社会主义体系概论》课程的整体情况较好，而《儿童心理学》课程在 2016 年 4 月份的人均在线时长尖峰这一异常情况经考察日志可以知道主要是由于教师用户在线修订课程造成的。

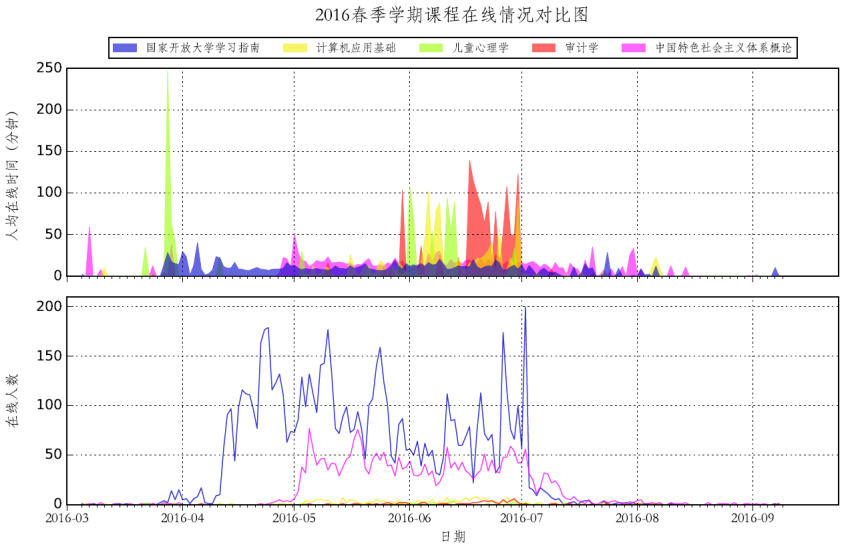

图 9 各月/各星期/各时段课程各模块访问量统计柱状图

通过对不同的课程进行横向比较，可以对整个平台的课程使用情况有一个全局的掌握，方便教学管理人员有针对性的开展教学督导工作。

结合上述模块和不同的图表，还让我们有能力给出更进一步的分析，以下讨论均以《国家开放大学学习指南》为例。首先，从图 9 下半部分，我们可以看到日在线人数一般在 50～150 的区间，图 5 里面看到，4 到 6 月份三个月的活跃用户数基本维持在 400 以上，考虑到图 5 中的活跃用户定义为往前一周里面累计的用户数，因此基本可以推算出日活跃用户数为 400/7，约等于 60 左右，若考虑到周内重复登陆情况（同一个用户周内重复登陆仅统计为一个）对推断日活跃用户的影响，日活跃用户数应该是高于 60 的一个值。无论如何，图 9 和图 5 在不同的统计尺度下相互印证了用户的登陆情况；其次，观察图 7 雷达图，我们可以看到在用户活跃度这里，该课程有大约超过 80%的用户登录次数超过 30，此统计结果要超过另一门对比课程（《中国特色社会主义理论体系概论》），然而，与之对应的是超低的人均在线时长（不足 20 秒）。这个数据与图 8 中，4～6 月之间人均在线时长约 10 分钟的事实表面上是冲突了，其实不然。因为图 8 中统计的值是当日登陆用户的人均在线时长。而图 7 中的人均在线时长是整个学期的人均在线时长。10 分钟与 20 秒，差了 30 倍，这并不仅仅是由于整个学期剩余时间（4 月初、7 月初）的用户登录造成的统计差别，因为我们可以看到，4～6 月份是整个学期课程最集中的使用时段，其他时段的数据较少，对统计的影响微乎其微。造成这个统计差别的原因在于，在时间方向上，用户重复登陆的情况较为稀少（除了图 5 标志的 4 月份的一波重复访问情况以外，见前述分析），大量的用户在很少的登陆次数上就完成了课程学习并不再登陆。这样的情况可以借助图 5 看出来。图 5 此三个月之间的新增用户数在震荡中维持了一个恒定的水平，而活跃用户数则相应的也在一个平均水平线上下波动，同时沉睡用户持续上升。重复登陆情况越低，统计时间段内的用户数就越大，相应的人均在线时长就下降了；最后，我们可以在图 7 中看到，两门课程的“论坛生命力指数”基本相同，然而论坛的使用情况却有较大差别。根据“论坛生命力指数”在本研究中定义，我们可以得出这样的印象，就是无论课程的使用情况如何，访问论坛的用户里面，有意愿参与讨论和发言的用户占比是比较恒定的。

四 总结与思考

Moodle 与慕课，对在线教学活动的开展，以及课程论坛的组织和引导向来有不同的观点，一些传统高校开始利用慕课这种形式进行教学，基本上主要是把基于授课视频的在线自学作为开展线下授课的一种补充形式，哪怕这部分教学时间占比再大，也无法替代日渐稀少的面授课时在翻转课堂模式教学中所起到的核心作用；而对于以全面远程授课为主的开放教育来说，由于时间空间的约束，以及学习者多样化所带来的挑战，面授课只是全面铺开远程教育之前的一种补充而已。因此，如何保障网络教学的质量，提高课程的在线使用率，只有在课程设计上和在线教学支持上下文章。而通过基于教育数据挖掘的数据可视化方法，可以对在线学习情况的趋势进行跟踪，对学习者的学习习惯、作业完成的方法进行分析，为开展在线教学检查和评估提供了更丰富、直观的依据，同时还可以为课程设计的持续改进，提供方向和意见。本研究的另一个贡献在于开发了一个支持扩展的 moodle 日志自动分析框架，为基于 moodle 的在线课程教学效果评估提供了一个自动化的方案。

参考文献

[1]任磊,杜一,马帅,等.大数据可视分析综述[J].软件学报,2014,25(9):1909-36.

[2](美)Ben Fry 著. 张羽译.可视化数据 [M].北京:电子工业出版社,2009:5-6.

[3]Mazza R, Milani C. GISMO: a Graphical Interactive Student Monitoring Tool for Course Management Systems[C]. International Conference on Technology Enhanced Learning, Milan. 2004:1-8.

[4]Calvani A, Fini A, Molino M, et al. Visualizing and monitoring effective interactions in online collaborative groups[J]. British Journal of Educational Technology, 2010, (41): 213–226.

[5]Podgorelec V, Kuhar S. Taking advantage of education data: Advanced data analysis and reporting in virtual learning environments[J]. Elektronika ir Elektrotechnika, 2011, 114(8):111-6.

[6]Aguilar DA, Therón R, Peñalvo FJ. Understanding Educational Relationships in Moodle with ViMoodle[C]. ICALT 2008 Jul 1: 954-956.

[7]Mazza R, Bettoni M, Faré M, et al. MOCLog -Monitoring Online Courses with log data[C]. Proceedings of the 1st Moodle Research Conference, Heraklion, Greece. 2012:132-139.

[8]张振虹,刘文,韩智.学习仪表盘: 大数据时代的新型学习支持工具[J].现代远程教育研究. 2014,(3):100-7.

[9]张金磊,张宝辉,刘永贵.数据可视化技术在教学中的应用探究[J].现代远程教育研究. 2013,(6): 98-104.

[10]殷贤亮,张为.Web 使用挖掘中的一种改进的会话识别方法[J].华中科技大学学报: 自然科学版.2006,34(7):33-35.

[11]魏顺平.在线学习行为特点及其影响因素分析研究[J].开放教育研究.2012,18(4):81-90.

[12]Lehmann J, Lalmas M, Yom-Tov E, et al. Models of user engagement[C]. International Conference on User Modeling, Adaptation, and Personalization. 2012:164–175.

[13]王苗.学习分析技术在网络课程学习中的应用实践研究[D]. 长春:东北师范大学,2014.

[14]魏顺平. Moodle 平台数据挖掘研究——以一门在线培训课程学习过程分析为例[J]. 中国远程教育. 2011(01).

Profiling Online Courses Using Moodle Logs*

HUANG Lei[1] WU Zhou-hua

(1. Centre of Educational Resources, Guangxi Radio and Television University, Nanning, Guangxi, China 530021)

Abstract: Along with the progressively online courses migration to moodle-based learning platform of The Open University of China, more and more learners start viewing pages, downloading courseware, taking online quizzes, editing wikis and taking part in interactions, which generates huge volumes of log data within the platform. This research paper, taking courses deployed in Guangxi branch for example, aims at online courses usage profiling from moodle log, by approaches of data visualization and web log mining, to monitor the operating situation of the courses, the engagement of the learners, and to help improve the administration and design of the courses.
Keywords: The Open University of China; moodle; web log mining; data visualization

---

*基金项目：本文为 2014 年度国家开放大学科研项目（移动学习应用模式的创新与实践 - - 以广西北海电大为例，项目编号：G14A3103Y）的阶段性研究成果。

作者简介：黄镭，科长，助理研究员，硕士，研究方向为教育数据挖掘与人工智能。邮箱为 1462044063@qq.com