

基于开放数据的 edX 课程学习者行为及相关因素分析

程志君, 罗鹏程, 罗玉文

(国防科技大学 系统工程学院, 湖南 长沙 410073)

摘要: 以 edX 平台的开放数据为基础, 研究 edX 课程学习者的行为和相关影响因素。首先根据开发数据集的特点选择相关数据项, 然后对描述学习行为的数据项, 如登录次数、互动次数、观看视频的次数、发帖数等进行聚类分析, 确定学习者的不同类型, 并根据其对应的数据均值进行比较, 讨论不同类型学习者的行为特点。同时联系其对应的年龄、国别和学历等背景数据开展关联分析, 探索学习行为以及学习背景对与学习者最终学习效果的综合影响, 为 MOOC 建设者改进课程设计, 优化在线交流方式等提供有益的参考。

关键词: 开放数据; MOOC; 聚类方法; 学习行为; 关联分析

中图分类号: G642 **文献标识码:** A **文章编号:** 1672-8874 (2017) 04-0046-04

An Analysis of Learning Behavior and Influencing Factors of edX Courses Based on Open Data

CHENG Zhi-jun, LUO Peng-cheng, LUO Yu-wen

(College of Systems Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract: The learning behavior of participants from edX courses and the influencing factors are analyzed in this paper based on the open dataset provided by edX platform. Firstly the data items related to the learning behavior have been chosen from the dataset, such as login, interaction, viewing and posting times, which are investigated by the clustering method to determine the types of learners. The characteristics of different types of learners are discussed through the comparison on the mean value of behavior factors. And their age, nationality and education are analyzed to define the relationship between the learning effect and these influencing factors, which will help the improvement of the design of courses contents and online communication style.

Key words: open data; MOOC; clustering method; learning behavior; relational analysis

一、前言

MOOC (大型开放式网络课程, massive open online courses) 作为一种新兴的教学模式, 凭借其优质的课程资源、开放的教学方式、自由分配的学习时间和在线自主的交流互动等特点, 自 2011

年诞生以来在全球范围内迅速发展。MOOC 突破原有传统教学模式在时间和空间上的限制, 极大地促进了教育领域的变革, 为现代教育提供了广阔的发展前景^[1]。

MOOC 学习者的学习兴趣、方式和能力各不相同, 为了让他们能尽快适应在线学习环境, 提高课程完成率和学习效果, 需要 MOOC 平台在提供

收稿日期: 2017-11-02

基金项目: 国防科技大学教育教学改革项目课题

作者简介: 程志君 (1978-), 女, 湖北武汉人。国防科技大学系统工程学院副教授, 博士, 主要从事装备系统工程, 数据驱动的可靠性分析与维修决策等。

学习支持服务的同时, 主动分析、了解学习者的学习特征和行为模式, 有针对性地设计课程内容, 优化教学模式, 改进互动交流及评价方式, 促进学习者个性化的教育。因此基于在线数据的学习者行为分析成为当前研究的热点之一^[2]。各大主流 MOOC 平台, 如美国斯坦福大学建立的 Coursera、哈佛大学与麻省理工学院建立的 edX、清华大学等高校建设的学堂在线、国防科技大学建立的首个军网 MOOC 平台 - 梦课等, 都开发了学习平台的数据收集与分析系统, 为全球研究者深入分析 MOOC 学习者行为提供了数据支持。蒋卓轩等以北大在 Coursera 平台上开设的 6 门课程学习者为对象, 将学习者分类, 考察学习行为与学习效果之间的关系, 并对其学习成果进行预测^[3]。王萍等利用 edX 提供的开放数据, 分别从性别、教育背景、年龄等方面探讨中国学习者的学习行为^[4]。李骁等以国防科技大学开设的军网梦课平台为对象, 从选课学习、成绩考核和互动交流 3 个行为方面分析, 总结了部队 MOOC 学习者的行为特征^[5]。

由上述分析可知, 目前的实证研究大多针对学习者学习过程中的单个变量进行统计分析, 如选课人数、退课人数、学习时间、学习完成率等, 对于将学习时间、互动次数、发帖数、学历背景等多因素综合的学习者分类方法及其对应特征的研究较少, 亟须结合大数据相关的分析方法, 如聚类分析等进行深入探讨。聚类分析是利用对象特征数据间的相似度来完成分类的一种多变量统计技术^[6], 能够综合利用多个变量的信息来完成样本的分类, 在数据挖掘、模式识别、决策等领域都有着广泛的应用。本文基于 edX 平台 2014 年发布的 2012 学年—2013 学年 16 门课程开放数据^[7], 利用聚类分析方法, 研究 edX 平台学习者的学习行为, 探究学习行为与学习者背景间的相互关联, 确定不同类别学习者的行为特征, 为促进学习者个性化学习提供一定的有益参考。

二、研究设计

(一) 研究框架

本文主要从学习者类型及对应特征的视角讨论以下几个问题:

(1) 学习者类型存在哪些类型?

(2) 不同类型学习者在学习行为上是否存在差异? 哪些行为上的差异较为显著?

(3) 不同年龄、国别、学历的学习者在 MOOC 中的学习行为存在哪些差异?

(二) 数据来源及说明

数据来源于 edX 2014 年发布的数据集——“HMXPC13_ DI_ v2_ 5 - 14 - 14. csv”, 其中包括 2012 学年—2013 学年秋季、冬季、春季三个学期哈佛大学和麻省理工学院在 edX 平台开设的 16 门课程数据。本文选取其中麻省理工开设的《The Challenges of Global Poverty》开展实证分析, 相关数据 7416 条, 涵盖用户 ID, 国家, 学历, 年龄等 20 个列项, 我们从中筛选出与学习者行为与学习成效相关的 13 项, 可以分为以下 4 类:

(1) 学习行为有关的数据项: 访问一半以上的章节数、登录学习次数、互动次数、观看视频的次数、学习的章节数、发帖数。

(2) 学习成效有关的数据项: 结业考试分数, 是否拿到合格证书。通过对拿到证书的学生进行筛选发现, 这些人的结业考试分数都在 0.5 以上, 故仅以结业考试分数作为最终衡量学习绩效的变量。

(3) 学习者背景有关的数据项: 国别、学历、年龄、性别。

(4) 选课行为有关的数据项: 观看过课程简介。

考虑到样本存在缺失的问题, 我们对样本数据进行整理, 剔除有缺失值的样本, 得到 3361 组有效数据。首先利用层次聚类方法对学习行为数据进行分析, 确定不同类型学习者的行为特点和规律, 然后结合年龄、国别、受教育程度等背景数据, 分析讨论不同学习行为、学习背景以及最终学习成绩之间的关联关系, 为下一步有针对性的改进措施和建议提供支持。

三、结果分析

本文利用 SPSS 软件进行相关数据的 K 均值聚类。以用户名为标识变量, 以学习行为相关的数据项为聚类变量, 计算得到 3—6 层聚类的结果, 考虑到样本容量和类与类之间的差异性, 更好地区别不同学习者的行为模式对学习效果的影响, 本文选取聚类数为 5 时的情况进行具体分析。

表 1 各类学习者行为数据的均值

学习类别	学习一半章节	登录学习次数	总的互动次数	视频观看次数	学习章节数	发帖数	总人数	得分
1	0.097	367.138	5.363	38.281	2.450	0.007	2981	0.01
2	0.997	4683.114	37.229	417.182	11.381	0.041	341	0.77
3	1.00	8779.556	51.639	984.139	11.889	0.028	36	0.82
4	1.00	14563	70.500	1566.500	12.000	0.000	2	1.00
5	1.00	11218	21.000	5865.00	11.000	0.000	1	1.00

（一）不同类型学习者的行为特征

在聚类数为 5 时，层次聚类得到每一类学习者的人数，以及完成一半章节数、登录次数、互动次数、视频观看次数、学习章节数、发帖平均数以及最终分数的均值，如表 1 所示，以最终得分的均值作为标准，将其分别称为体验学习者、一般学习者、积极学习者和学霸型学习者。下面分别对这些类别的行为进行比较分析：

（1）第一类：偶尔登录，完成一半章节学习的人不到 1%，课程结束时基本只完成前两章的学习，互动和视频观看数量相比其他类型也极少，这类学习者占总人数的 88.7%，但有 63.8% 的人结业考试 0 分，大部分没有参加考试，0.5 以上（可以拿到证书）的仅占 4.5%。这类学习者好奇心强，但明显缺乏持续学习的热情，学习效果自然很差，可以认为他们参与学习的主要目的是体验课程，因而将其归为体验学习者。

（2）第二类：课程结束时基本完成了所有章节的学习，与体验学习者相比，登录次数在其 10 倍以上，观看视频、互动和发帖数也有明显提升。这类学习者占总人数的 10.1%，大部分参加了结业考试，分数中等。但与后面三类学习者相比，各项学习数据较为一般。很明显，这类学习者对课程学习有一定的热情，但能够投入的精力有限，可以判断他们的目的性较强，期望能够获得认证，但效果一般，因而将其归为一般学习者。

（3）第三类：课程结束时所有人都完成全部章节的学习，登录次数、互动次数和视频观看数是一般学习者的 2 倍，唯一减少的是发帖数。这类学习者占总人数的 0.11%，都参加了最终的结业考试，分数良好。这类学习者对学习充满热情，投入较多精力进行课程学习，学习效果较好，将其归为积极学习者。

（4）第四类和第五类学习者的行为类似，登录次数和视频观看次数都远超前三类，考试成绩

满分，各项学习指标表现优秀，属于学霸型学习者。但在视频观看和互动上却出现明显差别，第五类学习者互动次数很少，甚至远低于一般学习者的平均水平，但视频观看数却达到惊人的 5865 次，几乎是第四类学习者的 4 倍，这说明第五类学习者属于不喜交流，属于性格内向型的学霸。考虑到学霸型的人数过少，我们将其归到积极学习者这一类，属于这一类中的特例。

（二）学习者行为的关联分析

为了进一步深入研究不同类学习者的行为特征，探索造成这些外在学习表现的深层原因，需要结合年龄、国别、学历等背景数据进行统计分析，确定它们与不同类型学习者学习行为之间的关联关系。

根据本研究统计，从年龄来看，体验学习者中 21—35 年龄段是主力，占比达到 73%，一般学习者中仍是这部分年龄的人数最多，但占比已经下降为 65%，到积极学习者年龄 21 岁—35 岁的比例只有 30%，其他年龄段占比较为平均。根据前面的分析，体验学习者中大多为好奇心强，但学习热情的持久性缺乏，这与青年求知欲强，但学习目的性大部分不太明确且耐心有限的特征较为符合。随着年龄的增长，学习的目的性更为明确，因此学习者的学习态度以及效果都明显好转，这与体验学习者、一般学习者到积极学习者的特征变化一致。

从国别来看，edX 将国家按教育普及程度分为 5 级（如表 2 所示），其中以美国为首的国家教育普及程度最高，印度为代表的国家则最低。根据四类学习者国别统计结果可以发现，不同国家中占绝大部分的都是体验学习者，比例都超过 70%，这说明了体验学习者的存在具有普遍性，与国家教育水平关系不大。但从后两类学习者占比来看，教育普及程度高的国家明显高出低级别国家，如第三类学习者中超过 92% 来自高级别国家，而第四类学习

者则全部来自教育普及程度高的国家，这从侧面反映了教育水平发达地区学习者的学习优势。

表 2 不同国家的教育普及程度^[8]

国 家	教育普及程度
巴基斯坦, 尼日利亚, 孟加拉, 摩洛哥, 印度, 其他	1
埃及	2
哥伦比亚, 印度尼西亚, 中国, 菲律宾, 墨西哥, 巴西	3
葡萄牙, 俄罗斯, 西班牙, 乌克兰, 希腊, 日本, 波兰	4
法国, 加拿大, 英国, 德国, 美国, 澳大利亚	5

从学历来看（如表 3 所示），以高中及本科以上为主，不同类型的学习者之间的差别不大。但随着学习热情和效率的提高，本科及以上学历人员的比例逐渐增大，这说明学历高的学习者目的更为明确，学习和理解能力相对较强，但不可忽视的是，体验学习者中本科及以上学历的人也近 40%，这说明高学历与其学习积极性、主动性和自觉性有时不一定成正比。

表 3 不同类学习者学历情况统计表^[9]

	1	2	3	4
高中	42.8%	44%	9.5%	0
本科及以上	39.2%	43.1%	87.5%	100%

注：1、2、3、4 分别表示体验学习者、一般学习者、积极学习者、学霸型学习者。

四、结论

本文利用主流 MOOC 平台 edX 提供的开放数据对相关课程学习者行为进行分析，确定不同类型学习者的行为特征，并结合其学历、国别和年龄进行关联分析。从中不难发现，学习者的学习行为，如登录次数、视频观看数、互动次数、发

帖数等从不同侧面反映其学习态度、能力和方法，对最终的学习效果影响显著，这与学习者的受教育程度、年龄等背景因素具有一定的关联关系，MOOC 建设者可以根据这些特点有针对性地设置课程内容和互动环节，从而达到提高学习者的学习热情和效率的目的。但同时由于聚类分析方法受初始类别设定和相似度指标的影响较大，因而对于前期数据筛选和分析者的要求较高，后续可结合相关方法进行进一步的探讨。

参考文献：

[1] 赵荣,马亮,张玉龙. MOOC 的理性思考:兴起、发展与未来[J]. 高等教育研究学报, 2014(2):9-14.

[2] 方旭. MOOC 学习行为影响因素研究[J]. 开放教育研究,2015(3):46-55.

[3] 蒋卓轩,张岩,李晓明. 基于 MOOC 数据的学习行为分析与预测[J]. 计算机研究与发展, 2015(3):614-628.

[4] 王萍. 基于 edX 开放数据的学习者学习分析[J]. 现代教育技术,2015(4):86-93.

[5] 李骁,柯水洲,敖永红,等. 梦课学习平台官兵学习行为分析与研究[J]. 网络化与数字化, 2015(6):89-96.

[6] 许丽利. 聚类分析的算法及应用[D]. 长春:吉林大学, 2010;20-28.

[7] MITx and Harvardx. Harvard X-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0 [EB/OL]. [2017-08-30]. <http://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147>.

[8] MITx and Harvardx. Harvard X-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0 [EB/OL]. [2017-08-30]. <http://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147>.

[9] MITx and Harvardx. Harvard X-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0 [EB/OL]. [2017-08-30]. <http://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147>.

（责任编辑：胡志刚）