DOI:10.19605/j.cnki.kfxxyj.2017.01.006

【学习支持与学习分析】

基于Canvas Network开放数据集的MOOC学习分析

胡红梅[1]　宗　阳[2]

（1.中央民族大学附属中学丰台实验学校　科研信息中心，北京　100074；2.北京师范大学远程教育研究中心，北京 100875）

【摘　要】大规模学习者的参与使MOOC平台记录了海量学习者行为数据，这为基于大数据进行学习分析提供了基础。基于Canvas Network发布的开放数据集特点，对数据集字段进行统计分析，力图全面展现 MOOC 学习者实际学习状况。在统计分析的基础上，选取有学习者自我评价数据的样本，分析不同类型学习者在学习目标、每周期望学习时间、学习行为和学习成绩上的差异性。

【关键词】MOOC；学习分析；学习行为；开放数据集

【中图分类号】G51　　**【文献标识码】**A　　**【文章编号】**2096-1510（2017）01-0055-08

一、引言

大数据时代，数据中蕴藏着巨大的价值信息等待挖掘，数据的重要性不言而喻。MOOC自2012年兴起以来在不同层面都不断引发新的研究和思考，由于其大规模的特质所产生的海量数据为学习分析与教育数据挖掘研究提供了基础。（王萍，2015）在互联互通的时代，数据开放已经越来越成为人们的共识。目前，很多国家和地区，都在互联网上开放了本国的公共数据，以供全世界人民使用。（姜桂兴，2015）在教育科研领域，随着MOOC的兴起，越来越多的开放数据集被发布出来供全球研究者使用。Coursera、edX等在线学习平台在创建系统时就已经考虑到大规模数据的收集和分析的重要性，因此这些平台对在线学习者特别是MOOC学习者的学习数据记录较为完善。作为非盈利性在线学习平台，很多MOOC平台在数据开放和教学研究方面做着重要的推进工作。2014年5月，哈佛大学和麻省理工学院联合发布了第一个大规模MOOC开放数据集，该数据集包括2012-2013 学年edX平台上共16门课程开放数据，为研究者深入分析和研究 MOOC 学习者和课程提供了数据支持。（Dataverse，2014）

Canvas Network（2016）是美国教育技术公司Instructure旗下的一个开放在线课程平台，MOOC是它提供的开放在线课程的一种。为了让研究人员更好地理解和描述MOOC现象以及在线学习行为和课程设计特征，Canvas Network依据edX2014年发布的数据集的特征于2016年3月公布了一份2014年1月-2015年9月的公开数据集。（Canvas,2016）本研究基于Canvas Network平台开放数据集，对数据集字段进行统计分析，并且选取有学习者自我评价数据的样本，探寻不同类型学习者在学习目标、期望学习时间、学习行为和学习成绩上的差异性。

二、Canvas Network开放数据集

Canvas Network 发布的开放数据集包括来自Canvas Network 开放课程（2014年1月-2015年9月）中的数据以及相关说明文档。由于用户在线数据中的隐私信息保护问题，与edx2014年开放数据集类似，该开放数据集在开放之前进行变量的选择规划，变

量的确定是基于Canvas Network 平台实际记录数据情况以及与相关研究者探讨决定，在这个过程中，由于学习者参与的不持续性去掉了像小组活动、同伴互评和wiki活动等学习行为变量。确定提供的目标变量后，进行了去身份识别（De-identification）处理。在去身份识别过程中，首先去掉了那些直接标识符和学习者年龄18岁以下的学习者数据，经过处理后，数量上相对原始数据集有所减少，对部分数据项的统计略有影响，但不影响整体性数据分析。

该数据集共包括Canvas Network 平台上10个学科门类238门课程，其中共130门课程学科为职业与应用类和教育类，而医学、数学、物理、计算机等科学类课程较少。238门课程开设时间多为1-2个季度，跨度达3个季度及以上的课程仅26门，课程长度多集中在35至65天之间。这些数据包括325 199条聚合记录，每条记录代表一个学习者个体在一门课中的学习行为。该数据集变量来自三部分：课程管理类变量、学习者与课程交互过程中产生的变量和来自学习者调查问卷的变量。该开放数据集的结构是基于edx 2014年开放数据集，由于去身份识别处理相较edx 2014年开放数据集去掉了学习者性别和学习者国家两个有分析价值的隐私字段，但是增加了学习者问卷调查数据。

数据集共提供了26个列项，列项可分为包括课程和学习者四类数据，分别为课程信息、学习者基本信息、学习者学习意图信息和学习者学习行为信息（见表1）。

三、学习数据分析

在该部分，本研究基于Canvas Network开放数据集来分析MOOC学习者的实际学习情况，从学习者基本信息、学习者学习意图信息和学习者学习行为三个方面进行分析。

1. 学习者基本信息

该数据集共包括224 914名学习者，由于去身份

表1　Canvas Network开放数据集描述

类别	列名	列项描述	实际数据存储说明
课程信息	course_id_DI	课程ID	去身份识别后随机生成的课程编码
	course_start	课程开始时间	以一年四个季度形式存储
	course_end	课程结束时间	以一年四个季度形式存储
	discipline	学科类别	学科名称
	grade_reqs	成绩要求	0或1，如果是1说明课程中至少3个作业
	course_reqs	课程内容模块	0或1，如果是1说明课程中至少3个内容模块
	course_length	课程天数	课程开设的天数或者课程有活动的天数，比实际开设天数要大
学习者基本信息	userid_DI	用户ID	去身份识别后随机生成的用户编码
	registered	记录信息	0或1，如果是1说明所有记录都在该数据集中
	final_cc_cname_DI	学生国籍	当前版本去身份识别处理而去掉，后续版本可能会提供
	LoE_DI	学历水平	去身份识别的来自课程问卷选择数据
	age_DI	年龄	去身份识别的年龄段
	gender	性别	由于去身份识别当前版本未提供，后续版本可能会提供
学习者学习意图信息	learner_type	学习者类型	来自课程问卷选择数据
	primary_reason	学习目的	来自课程问卷选择数据
	expected_hours_week	每周期望学习时间	来自课程问卷选择数据
学习者学习行为信息	start_time_DI	开始交互时间	去身份识别的学习者第一次交互时间，以一年四个季度形式存储
	last_event_DI	最后交互时间	去身份识别的学习者最近一次交互时间，以一年四个季度形式存储
	nevents	学习事件数	课程汇总不同交互数，记录为页面浏览次数
	ndays_act	学习事件跨度	不同事件涉及的天数
	nforum_posts	发帖数	在课程中发帖总数
	viewed	交互数	0或1，如果是1说明在课程内交互数大于等于1
	explored	学习程度	0或1，如果是1表示学习者与课程全部模块的交互或浏览比例大于等于50%
	ncontent	模块浏览比例	不同课程模块浏览的百分比
	completed_%	模块完成比例	完成全部课程内容模块的百分比，针对内部要求的内容模块大于一个内部门槛
	grade	成绩比例	估计的或者实际成绩百分比，针对那些有足够给分活动来计算最终成绩的课程

识别的处理，学习者基本信息里面学习者国籍和性别字段被去除，年龄字段也不包括18岁以下年龄段数据。去身份识别后剩余31 890名学习者有年龄信息，学习者的年龄段信息统计如表2。

表2 学习者年龄分布表

年龄段	学习者人数	所占比例
{19−34}	13020	40.83%
{34−54}	14276	44.77%
{55 or older}	4594	14.41%

可以看出Canvas Network平台上学习者年龄集中在19-54岁之间，占85.6%，55岁以上学习者占14.41%的比例。

在学习者学历信息部分，剔除无记录或丢失信息数据后共剩32 783名学习者学历信息，统计如表3。

表3 学习者学历分布表

学　历	学习者人数	所占比例
完成2年大学学位	1842	5.62%
完成4年大学学位	7479	22.81%
高中或大学预科	1533	4.68%
硕士学位（或同等学历）	11969	36.51%
其它	442	1.35%
哲学博士，法学博士，或者医学博士学位（或同等学历）	2501	7.63%
大学在读,还未完成学位	3896	11.88%
上过一些研究院	3121	9.52%

可以看出Canvas Network平台上学习者人数最多的学历硕士学位（或同等学历），占36.51%，其次是完成4年大学学位的学习者，占22.81%。

2. 学习者学习意图信息

该部分信息数据来自学习者调查问卷，学习目的、学习者类型和每周期望学习时间三个变量共有34 896条记录有数据。在学习目的方面，不同学习目的人数分布如表4。

表4 学习者学习目的分布表

学习目的	记录数	所占比例
1.对MOOCs感到好奇	2170	6.22%
2.第一次准备考大学	283	0.81%
3.准备返回学校	766	2.20%
4.喜欢成为社区学习者中的一员	1219	3.49%
5.喜欢学习那些吸引我的主题	19566	56.07%
6.希望能获得新的职业技能	4262	12.21%
7.希望能获得促进工作提升的技能	2150	6.16%
8.希望能获得在工作中使用的技能	144	0.41%
9.喜欢在线学习的形式	2814	8.06%
10.想要尝试Canvas Network平台上的课程	1522	4.36%

可以看出Canvas Network平台上的学习者主要学习目的是“喜欢学习那些吸引我的主题”，占56.07%，其它学习目的均不超过13%。

学习者类型方面，不同类型学习者分布情况如图1。

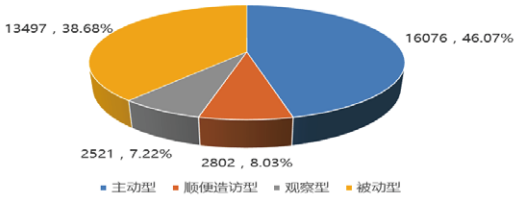

图1 学习者类型分布图

可以看出Canvas Network平台上主动型学习者占46.07%，被动型学习者占38.68%，其它两类学习者人数不足16%。

学习者每周期望学习时间方面，不同期望学习时间分布情况如图2。

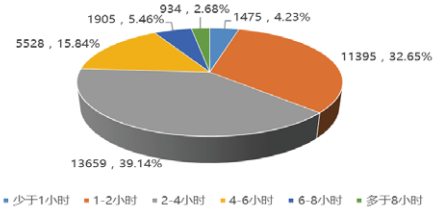

图2 学习者每周期望学习时间分布图

可以看出，大部分学习者每周期望学习时间为1-4小时，占71.79%，两个极值（少于1小时和多于8个小时）的学习者人数共占6.91%。

3. 学习者学习行为信息

学习者学习事件数共86 287条记录有数据，最小事件数为1，最大事件数为530 411。随着事件数增多，记录数整体呈下降趋势，事件数为3时急剧下降；最高记录数是事件数为1的记录数（6812），事件数为2的记录数为6760，是事件数为3的记录数的2.2倍。

学习者学习事件跨度字段，共有101234条记录有数据，最长时间跨度天数为167天，最短时间跨度为1天，以54天为界限，54-167天之间虽然仍然有记录但是都小于10，多为1或2这种特殊情况。跨度集中在10天以内，随着事件跨度天数增加，记录数整体呈递减趋势，在事件跨度10天以内急剧递减。跨度为1天的记录数最多（43 232），占42.71%。

学习者发帖数字段，共37091条记录有数据，学习者在一门课程中最大发帖数为436，最小发帖数为1，发帖数集中在10以内，随着发帖数的增加，记

录数整体呈递减趋势，发帖数在5条以内的记录数急剧递减；发帖数为1的记录数最多（16 313），占43.98%。

学习者交互数和学习程度字段记录形式为0或1，交互数中1表示学习者课程内交互数大于等于1，学习程度字段中1表示学习者与课程模块交互浏览比例大于等于50%，交互数和学习程度字段0和1值记录数分布情况如表5。

表5 学习者交互数和学习程度分布表

数值	交互数	学习程度
0	245 724	315 468
1	79 475	9 731

由表5可知，交互数中共有245 724条记录为0，剩余79 475条记录为1，即没有交互的学习者人数是有交互的学习者人数的3倍多，绝大部分学习者课程内交互数小于1。学习程度中共有315 468条记录为0，剩余9 731条记录为1，即与课程模块交互浏览比例小于50%的学习者人数是大于等于50%的学习者人数的32倍多。

学习者课程模块浏览比例字段，共91 312条记录有数据。随着学习者课程模块浏览比例的增加，记录数呈上下波动趋势；课程模块浏览比例为100%的记录数最多（16 016），占据17.5%，其它比例最高不超过2.3%。

学习者课程模块完成比例字段，共26 661条记录有数据，课程模块完成比例为0.077和1的记录数较多，分别为2 804、1 011；在0.5以内分布较密集。

学习者的成绩比例字段，共82 002条记录有数据。成绩比例呈现两极分化的趋势，成绩比例是0或1的记录占了绝大部分，成绩比例是0的记录数为24 150，占29.45%；成绩比例是1的记录数为4 844，占5.91%。成绩比例在0.5以内分布较密集。

四、不同类型学习者差异检验分析

（一）不同类型学习者学习目的差异分析

Canvas Network开放数据集里学习者共四类：主动型、顺便走访型、观察型和被动型。学习者学习目的共10种，具体参见表4。我们将10种学习者学习目的分为两类：兴趣类和提升类。兴趣类包括：1.对MOOCs感到好奇；4.喜欢成为社区学习者中的一员；5.喜欢学习那些吸引我的主题；9.喜欢在线学习的形式；10.想要尝试Canvas Network平台上的课程。提升类包括：2.准备考大学；3.准备返回学校；6.希望能获得新的职业技能；7.希望能获得促进工作提升的技能；8.希望能获得在工作中使用的技能。

在表4学习者学习目的分布表中，我们可以看出以兴趣为学习目的共27 291条记录，占78.2%。下面将四类学习者分别进行学习目的统计，然后与学习者总体进行比较，探寻不同学习者类型学习目的差异性。具体比较方法为：将该类学习者的某个学习目的人数占比与总体学习者对应学习目的人数占比进行比较，如果结果是大于，则说明该类学习者的这个学习目的高于一般水平，我们记录为1；如果比较结果是等于，则说明该类学习者的这个学习目的属于一般水平，我们记录为0；如果比较结果是小于，则说明该类学习者的这个学习目的低于一般水平，我们记录为-1。最后计算各类型学习者兴趣类和提升类学习目的的总得分，比较两类学习目的总得分，不同类型学习者的学习目的属于总得分高的那类，结果如表6。

表6 不同类型学习者不同学习目的倾向打分统计表

目的类型	学习目的	主动型	顺便造访型	观察型	被动型
兴趣类	对MOOCs感到好奇	−1	1	1	1
	喜欢成为社区学习者中的一员	1	1	−1	−1
	喜欢学习那些吸引我的主题	−1	−1	−1	1
	喜欢在线学习的形式	−1	1	1	1
	想要尝试Canvas Network平台上的课程	−1	1	1	1
提升类	第一次准备考大学	−1	1	−1	1
	准备返回学校	1	−1	−1	−1
	希望能获得新的职业技能	1	−1	−1	−1
	希望能获得促进工作提升的技能	1	1	−1	−1
	希望能获得在工作中使用的技能	1	−1	−1	1

从表6可以得出主动型学习者兴趣类目的得分为-3，提升类目的得分为3分；顺便造访型学习者兴趣类目的得分为3，提升类目的得分为-1；观察型学习者兴趣类目的得分为1，提升类目的得分为-5；被动型学习者兴趣类目的得分为3，提升类目的得分为-1。即主动型学习者以提升自我为主要学习目的，他们的学习更多的是为了提升自我能力；顺便造访和观察型学习者以兴趣为主要学习目的，但是兴趣类目的得分均为1，说明他们的学习目的不强；

被动型学习者主要以兴趣为主要学习目的，他们的学习更多是基于自己的兴趣。

（二）不同类型学习者每周期望学习时间差异分析

学习者每周期望学习时长共6种，分别为低于1小时、1-2小时、2-4小时、4-6小时、6-8小时、多余8小时。我们将期望学习时间按时长分为高、中、低三个时间投入层次，高时间投入对应6小时及以上，中时间投入对应2-6小时，低时间投入对应2小时及以下。

将四类学习者分别进行期望学习时间统计，然后与学习者总体进行比较，探寻不同学习者类型期望学习时间差异性。比较方法与学习目的相似，具体为：将该类学习者的某个等级时间投入人数占比与总体学习者对应等级时间投入人数占比进行比较，如果结果是大于，则说明该类学习者的这个等级时间投入比一般水平高，记录为1；如果结果是等于，则说明该类学习者的这个学习等级时间投入属于一般水平，记录为0，如果结果是小于，则说明该类学习者的这个学习时间投入比一般水平低，记录为-1。最后计算不同时间投入等级总得分，总得分高的每周期望学习时间等级即为该类学习者每周期望学习时间。结果如表7。

表7 不同类型学习者每周期望学习时间倾向打分统计表

时间投入等级	期望学习时间	主动型	顺便造访型	观察型	被动型
低时间投入	少于1小时	−1	1	1	1
	1−2小时	−1	1	1	1
中等时间投入	2−4小时	1	−1	−1	−1
	4−6小时	1	−1	−1	−1
高时间投入	6−8小时	1	−1	−1	−1
	多于8小时	1	−1	−1	−1

从表7可以得出主动型学习者低时间投入得分为-2，中等时间投入得分为2分，高时间投入得分为2；顺便造访型学习者低时间投入得分为2，中等时间投入得分为-2分，高时间投入得分为-2；观察型学习者低时间投入得分为2，中等时间投入得分为-2分，高时间投入得分为-2；被动型学习者低时间投入得分为2，中等时间投入得分为-2分，高时间投入得分为-2。即主动型学习者期望进行中高时间投入，顺便造访、观察和被动型学习者均期望低时间投入。

（三）不同类型学习者学习行为差异检验分析

采用SPSS 20软件对学习者的学习事件数、学习事件跨度、发帖数、模块浏览比例、模块完成比例五个学习者学习行为变量进行分析，不同类型学习者有效人数分别为：主动型学习者633人，顺便造访型学习者120人，观察型学习者52人，被动型学习者310人。对这部分学习者的学习行为采用单因素方差分析（ANOVA）方法分析，以对比不同类型学习者的学习行为差异，结果如表8所示。

表8 不同类型学习者学习行为单因素方差分析（ANOVA）结果

		平方和	df	均方	F	p
学习事件数	组间	4649972.049	3	1549990.683	0.565	0.638
	组内	3046612300.543	1111	2742225.293		
	总数	3051262272.592	1114			
学习事件跨度	组间	549.975	3	183.325	3.551	0.014
	组内	57362.101	1111	51.631		
	总数	57912.075	1114			
发帖数	组间	602.380	3	200.793	8.768	0.000
	组内	25444.020	1111	22.902		
	总数	26046.400	1114			
模块浏览比例	组间	23759.734	3	7919.911	6.376	0.000
	组内	1379988.822	1111	1242.114		
	总数	1403748.556	1114			
模块完成比例	组间	0.060	3	0.020	1.563	0.197
	组内	14.259	1111	0.013		
	总数	14.319	1114			

从表8可以发现，不同类型的学习者在学习事件跨度、模块浏览比例、发帖数方面存在显著差异（$p<0.05$），在学习事件数、模块完成比例方面不存在显著差异（$p>0.05$）。

在单因素方差分析的基础上，采用LSD方法检验不同类型学习者两两之间的学习行为差异，结果如表9所示。

表9 不同类型学习者学习行为LSD检验结果

	序列	均值差	p
学习事件数	主动型vs顺便造访型	94.334	0.567
	主动型vs观察型	153.269	0.521
	主动型vs被动型	136.268	0.235
	顺便造访型vs观察型	58.935	0.830
	顺便造访型vs被动型	41.934	0.814
	观察型vs被动型	−17.001	0.945
学习事件跨度	主动型vs顺便造访型	0.156	0.828
	主动型vs观察型	1.892	0.068
	主动型vs被动型	1.445*	0.004
	顺便造访型vs观察型	1.736	0.146
	顺便造访型vs被动型	1.289	0.095
	观察型vs被动型	−0.447	0.678
发帖数	主动型vs顺便造访型	0.634	0.184
	主动型vs观察型	0.999	0.148
	主动型vs被动型	1.687*	0.000
	顺便造访型vs观察型	0.365	0.646
	顺便造访型vs被动型	1.053*	0.041
	观察型vs被动型	0.688	0.338
模块浏览比例	主动型vs顺便造访型	7.949*	0.024
	主动型vs观察型	15.453*	0.002
	主动型vs被动型	−3.643	0.136
	顺便造访型vs观察型	7.504	0.200
	顺便造访型vs被动型	−11.592*	0.002
	观察型vs被动型	−19.096*	0.000

续表9

	序列	均值差	p
模块完成比例	主动型vs顺便造访型	0.013056	0.247
	主动型vs观察型	0.020817	0.203
	主动型vs被动型	−0.007470	0.342
	顺便造访型vs观察型	0.007761	0.680
	顺便造访型vs被动型	−0.020526	0.092
	观察型vs被动型	−0.028287	0.096
*．均值差的显著性水平为 0.05。			

不同类型学习者学习行为LSD检验结果发现：学习事件数方面，四种类型学习者学习事件数平均值从大到小依次为主动型学习者、顺便造访型学习者、被动型学习者、观察型学习者。不同类型学习者之间均无显著差异（p>0.05）。

学习事件跨度方面，四种类型学习者事件跨度平均值从大到小依次为主动型学习者、顺便造访型学习者、被动型学习者、观察型学习者。主动型学习者与被动型学习者之间存在显著差异（p<0.01），即主动型学习者的事件跨度明显大于被动型学习者。主动型学习者与顺便造访型学习者、观察型学习者之间不存在显著差异（p>0.05）。顺便造访型学习者、观察型学习者、被动型学习者两两之间不存在显著差异（p>0.05）。

发帖数方面，四种类型学习者发帖数平均值从大到小为主动型学习者、顺便造访型学习者、观察型学习者、被动型学习者。主动型学习者与被动型学习者之间存在显著差异（p=0.000），即主动型学习者的发帖数显著多于被动型学习者。主动型学习者与顺便造访型学习者、观察型学习者之间不存在显著差异（p>0.05）。顺便造访型学习者与被动型学习者之间存在显著差异（p=0.041），即顺便造访型学习者的发帖数显著多于被动型学习者。观察型学习者与顺便造访型学习者、被动型学习者之间不存在显著差异（p>0.05）。

模块浏览比例方面，四种类型学习者模块浏览比例平均值从大到小依次为被动型学习者、主动型学习者、顺便造访型学习者、观察型学习者。主动型学习者与顺便造访型学习者、观察型学习者之间存在显著差异（p<0.05），即主动型学习者的模块浏览比例显著高于顺便造访型学习者和观察型学习者。主动型学习者与被动型学习者之间不存在显著差异（p>0.05）。被动型学习者与顺便造访型学习者、观察型学习者之间存在显著差异（p<0.01），即被动型学习者的模块浏览比例显著高于顺便造访型学习者和观察型学习者。顺便造访型学习者与观察型学习者之间不存在显著差异（p=0.200）。

模块完成比例方面，四种类型模块完成比例平均值从大到小为被动型学习者、主动型学习者、顺便造访型学习者、观察型学习者。不同类型学习者之间均无显著差异（p>0.05）。

（四）不同类型学习者学习成绩差异检验分析

采用SPSS 20软件对成绩比例变量进行分析，不同类型学习者有效人数分别为：主动型学习者13 312人，顺便造访型学习者2 261人，观察型学习者1 914人，被动型学习者11 114人。对这部分学习者的学习成绩采用单因素方差分析（ANOVA）方法分析，以对比不同类型学习者的学习成绩差异，结果如表10所示。

表10 不同类型学习者学习成绩单因素方差分析（ANOVA）结果

	平方和	df	均方	F	p
组间	88.005	3	29.335	208.906	0.000
组内	4015.636	28597	0.140		
总数	4103.641	28600			

从表10可以发现，不同类型学习的成绩比例存在显著差异（p=0.000）。

由于不同类型学习者的成绩比例方差不齐，因此，在单因素方差分析的基础上，采用Tamhane方法检验不同类型学习者两两之间的成绩比例差异，结果如表11所示。

表11 不同类型学习者成绩比例Tamhane检验结果

序列	均值差	p
主动型vs顺便造访型	0.1394097*	0.000
主动型vs观察型	0.1686726*	0.000
主动型vs被动型	0.0800374*	0.000
顺便造访型vs观察型	0.0292629*	0.024
顺便造访型vs被动型	−0.0593722*	0.000
观察型vs被动型	−0.0886351*	0.000

不同类型学习者成绩比例Tamhane检验结果发现：四种类型学习者成绩比例平均值从大到小依次为主动型学习者、被动型学习者、顺便造访型学习者、观察型学习者。不同类型学习者之间均存在显著差异（p<0.05）。

五、结论思考

（一）MOOC学习者学习现状主要发现

MOOC学习者学历水平多为硕士和本科，Canvas Network平台上学习者人数最多的学历为硕

士学位（或同等学历）和完成4年大学学位的学习者，这两类学习者人数共占59.32%，这说明MOOC学习者学历主要以本科和硕士及同等学历为主，这与中国MOOC学习者学历情况相似[1]。

本次Canvas Network由于去身份识别去掉了18岁以下学习者信息，在19岁以上学习者三个年龄层（{19-34}，{34-54}和{55 or old}）中，34-54岁人数最多，占44.77%，并且55岁及以上学习者也占了14.41%，这与中国MOOC学习者年龄集中在20-30岁相比年龄提升一个等级，这说明世界范围内MOOC学习者年龄普遍比中国学习者年龄偏高，另一个层面说明中老年阶层人员也有巨大的MOOC学习潜力。

Canvas Network上学习者的学习目的表明，56.07%的学习者“喜欢学习那些吸引我的主题”，由此推断，课程教师在开设MOOC课程前，课程主题的选择对于吸引学习者比较重要。

在Canvas Network提供的数据集中学习者的学习行为数据包括两类：数量类和比例类。数量类学习行为包括学习事件数、学习事件跨度和发帖数，在这三种学习行为上随着数量增加，学习者人数均呈现递减趋势，并且在最小范围内递减趋势明显，这说明大多数MOOC学习者学习参与交互程度较低；学习行为数量跨度较大，这说明MOOC学习者的参与交互潜力巨大，MOOC学习者学习参与交互程度还有很大的提升空间；学习事件跨度集中在10天以内，这说明大多数MOOC学习者难以坚持学习，如何设计课程以促进学习者坚持学习是网络学习通常面临的难题。

比例类学习行为包括课程模块浏览比例、课程模块完成比例和学习成绩比例，在这三种学习行为分析中，学习者课程模块浏览比例人数最多的是100%，而学习者课程模块完成比例人数最多的比例是7.7%，学习者课程模块浏览比例显著高于模块完成比例。这在一定程度上反映了MOOC学习者的学习深入程度较低，对于大部分的课程模块仅仅做到了浏览层次，而很少有完成的行为。学习者成绩比例呈现明显的两极分化的特点，即MOOC学习者成绩比例大多是0或1，这在一定程度上反映当前世界范围内MOOC课程形成性测试存在的问题。成绩为0的可能原因有两种，一是测试对于学习者来说太难，二是学习者在测试中不愿意完成测试而直接“交白卷”，由于第二种原因存在的可能性，导致难以正确评价MOOC学习者的学习效果，这也是MOOC学习评价面临的难题。

（二）不同类型MOOC学习者学习差异特点

不同类型学习者的学习目的不同，其中主动型、被动型和顺便造访型学习者学习目的较为明确，主动学习者主要是为了提升能力，被动和顺便造访型学习者主要是基于学习兴趣，而观察型学习者则没有明确的学习目的。基于不同的学习目的，不同类型学习者每周期望学习时间投入也各不相同，主动型学习者学习时间投入明显区别于另外三类学习者，主动学习者每周期望投入更多的学习时间，而其它三类学习者时间投入上表现一致，都期望投入最少的学习时间。

对MOOC学习者的学习事件数、学习事件跨度、发帖数、模块浏览比例和模块完成比例五种学习行为进行不同类型学习者差异检验分析结果显示在学习事件数、学习事件跨度和发帖数三种学习行为上，主动型学习者平均值最大，在模块浏览比例和完成比例方面被动型学习者平均值最大，主动学习者平均值虽然第二，但是与被动学习者之间不存在显著差异，这说明主动型学习者会进行全方位更多的学习行为投入，被动型学习者相对主动型学习者会更加有选择性地进行投入，侧重在与成绩相对直接相关的学习行为上，比如模块浏览和完成比例这些可能作为对MOOC学习者进行形成性评价指标的行为。观察型和顺便造访型学习者在五种学习行为上均不存在显著差异，在五种学习行为平均值上都具有相对一致的低投入行为表现，说明这两类学习者类型比较相似。

对不同类型MOOC学习者成绩比例进行差异检验发现，Canvas Network上四种类型学习者成绩比例平均值从大到小依次为主动型、被动型、顺便造访型、观察型学习者，这与两个综合类的行为指标（模块浏览比例和模块完成比例）表现一致，并且不同类型学习者之间均存在显著差异。学习成绩差异分析结果再次说明MOOC学习者学习行为与学习成绩显著相关，这为通过一些综合性的学习行为指标对不同类型MOOC学习者学习成绩进行预测提供了事实依据。通过对被动学习者的学习行为和学习成绩进行具体分析可以发现，被动型学习者在单方面行为指标（学习事件数、学习事件跨度和发

帖数）上表现和顺便造访型及观察型学习者表现一致，甚至表现更糟，但是在一些综合性行为指标上表现很好依然可以取得较好的学习成绩，这一方面可能说明对MOOC学习者进行单方面行为指标评价意义不是很大，另一方面也可能说明现在世界范围内MOOCs中对学习者的学习评价存在一定的不足之处，需要更加完善的综合评价指标来对MOOC学习者进行真实学习效果评价。

六、总结

继edX公开发布第一个MOOC数据集之后，Canvas Network于今年提供了一份更为全面的MOOC数据集，这为研究者提供了宝贵的研究资源。本研究基于该数据集现状，在对MOOC学习者学习现状数据分析的基础上，对不同类型学习者在学习目标、每周期望学习时间、学习行为和学习成绩方面进行差异性分析发现不同类型MOOC学习者由于注册学习课程的目标和学习投入期望的差异性导致他们学习行为表现出显著差异性。当然该数据集由于去身份识别等安全隐私方面的保密措施，去掉了一些关键学习者信息，如学习者性别和国家信息，以及缺乏时间、点击流等重要学习者学习行为要素信息，因此还无法对学习者的学习进行更加深入的分析。MOOC 应用产生的海量数据为在教育领域进行挖掘分析提供了基础，大数据时代，我们需要将大数据与传统的控制数据集（小数据）结合起来，创建对人类行为更深入、更准确的表达（Lazer et al,2014）。Canvas Network平台在开放该数据集（个人-课程）的同时，公布了一份2014年4月至2015年9月份其平台上学习者课程、活动和更为详细的学习者学习行为过程数据集（Canvas,2016），本研究后续会持续跟进研究，以期发现更多有价值的MOOC学习行为现象，进而为通过学习分析对MOOC学习者提供个性化学习支持提供更多借鉴参考。

参考文献

[1]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.

[2]姜桂兴. 全球开放数据运动蓬勃发展[N]. 学习时报,2015-03-30(7).

[3]Dataverse.Deidentified student level data from the first year of HarvardX and MITxcourses[DB/OL].(2014-06-09)[2016-09-26].http://thedata.harvard.edu/dvn/dv/mxhx.

[4]Canvas Network平台[OL].[2016-09-26].https://www.canvas.net/

[5]Canvas Network 个人-课程开放数据集[DB/OL].（2016-03-01)[2016-09-26].https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/1XORAL.

[6]Lazer D, Kennedy R, King G, et al. The Parable of Google Flu: Traps in Big Data Analysis[J]. Science, 2014, 343（6176):1203-1205.

[7]Canvas Network 学习者学习行为过程开放数据集[DB/OL].（2016-03-30)[2016-09-26].https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XB2TLU.

作者简介：

胡红梅，硕士，中央民族大学附属中学丰台实验学校信息技术教师。研究方向：信息技术与课程整合，学习分析。

宗阳，硕士，北京师范大学远程教育研究中心。研究方向：在线学习分析，教育数据挖掘。

MOOC Learning Analysis Based on Canvas Network Open Data Set

Hu Hongmei[1] and Zong Yang[2]

（1. Fengtai Experimental School of the High School Affiliated to Minzu University of China, Research Information Center, Beijing,100074; 2. Beijing Normal University, Research Center of Distance Education, Beijing,100875.）

Abstract: The large-scale learner participation makes the MOOC platform record a mass of data about learner behavior, which is the base of learning analysis. Based on the characteristics of open data set published by Canvas Network, the statistical analysis was used to analyze the data set field in this study, attempting to fully demonstrate MOOC learners' actual learning status. On the basis of statistical analysis, a sample of learners' data with self-evaluation data was selected to analyze the differences among different types of learners in learning objectives, expected learning time per week, learning behavior and learning achievement.

Keywords:MOOC; Learning Analysis; Learning Behavior; Open Data Set