MOOC 在线学习行为的人类动力学分析

樊 超[1] 宗利永[2]

(1. 山西农业大学 文理学院,山西太谷 030801;2. 上海出版印刷高等专科学校 文化管理系,上海 200093)

[摘要] 近年来,随着互联网技术的深入推进,在线学习模式和 MOOC 平台得到了日益普及和发展,与传统课堂教学模式形成互补。同时,在线学习也得到了教育学界的普遍关注。为了深入了解用户在线学习的行为模式和特点,本研究采用人类动力学研究方法,对国内 MOOC 平台"学堂在线"的用户学习抽样数据进行定量分析发现,用户的在线学习行为具有明显的异质性(即非均匀性),表现在:1)用户的选课量和课程的选课人数有很大差异;2)用户在线学习时间间隔分布和持续时间分布都表现为幂律分布,因而在线学习行为具有阵发和重尾的特征;3)对用户学习行为的活跃性研究发现,用户学习的活跃天数大多较短,在线学习时间和次数也服从幂律分布,且随着时间的推移,用户的平均学习时间会上升,在课程学习后期的讨论环节会投入更多时间和精力。

[关键词] 人类动力学;大数据;时间标度律;活跃性

[中图分类号] G434 [文献标识码] A [文章编号] 1007-2179(2016)02-0053-06

一、引言

当前,在线学习领域内有不同平台下的多种学习形式,例如国际上享有盛誉的 edX、Coursera 和 Udacity 三大 MOOC 平台,国内有规模较大的中国大学视频公开课、MOOC 中国、MOOC 学院等。这些平台汇聚了大量源自世界名校的优秀课程,且大多数免费,也有部分是可以颁发认证的收费课程。此外,很多教育培训类网站还提供有偿收费的学习和培训课程。正是由于在线学习具有学习泛在化以及资源系统化和海量化的优点,这种新型的学习方式能够迅速风靡全世界,并对传统教学方式以及图书资料等管理方式产生巨大冲击,因此也受到了教育学者的普遍关注,大量关于在线学习的研究成果相继问世。例如,陈肖庚等(2013)、袁松鹤等(2014)分别介绍了 MOOC 的诞生过程及其在国内的发展状况;胡勇等(2006,2015)、肖为胜(2009)、杨素娟(2011)等从理论上阐述了在线学习行为的教育学特征和模式;李松(2010)、胡小勇(2015)等结合具体技术对在线学习平台设计进行了探讨;王红艳(2013)、龙三平(2014)等对期刊引文数据库中关于在线学习研究的论文进行了分析,总结了在线学习领域的研究现状。高地(2014)、蔡文芳(2015)对 MOOC 平台突出的高退课率和低通过率现象进行了探讨和反思。

关于在线学习行为的现有研究大多是从教育学理论角度进行定性分析,而基于定量计算的研究并不多见,这在很大程度上是因为人们难以获取和处理在线学习相关数据。而如今,随着大数据时代的来临,人们逐渐能够借助信息技术和程序算法获取并分析处理学习过程中产生的数据。最近,用定量方法分析在线学习行为的研究陆续面世,这些研究采用教育统计学或者数据分析方法对 MOOC 平台在线学习数据进行统计、分析和预测。方旭(2015)基于 TAM3 模型建立了 MOOC 学习行为影响因素模型,计算了主观规范、计算机自我效能感等十余种变量对学习行为中有用性感知和易用性感知因素的影响程度,并提出相应教育学对策;李曼丽等

[收稿日期]2016-01-07 [修回日期]2016-02-29 [DOI 编码]10.13966/j.cnki.kfjyyj.2016.02.007

[基金项目]山西农业大学科技创新基金"人类行为的分形特征研究"(201208);上海出版传媒研究院招标课题"在线学习行为及动力学研究"(SABY1403);上海市教委科研创新项目"文化创意产业社会化商业模式运行机制研究"(15ZS093)。

[作者简介]樊超,博士,山西农业大学讲师,上海出版传媒研究院研究员,研究方向:人类行为动力学、社会网络分析、数据挖掘(fanchao.cn@gmail.com);宗利永,博士后,上海出版印刷高等专科学校副教授,研究方向:社会化媒体用户行为。

(2015)以“学堂在线”平台上的一门课程为例,使用Tobit和Logit两个定量分析模型分析影响课程参与和完成的因素及其影响方式,包括课程注册时间、课程参与度、学习时长、学生来源等;蒋卓轩等(2015)对MOOC数据中的学习者进行分类,分析每一类学习者的学习行为特征,并以此为基础预测学习效果。这类研究代表着用数据分析的思路分析学习行为成为可行。

研究用户的在线学习行为有助于理解人类学习行为机制,了解每个用户行为特征和学习方式,从而为提高课程完成率、给用户提供个性化服务奠定基础。在大数据的契机下,人类动力学这一新兴的研究方向为定量分析用户的在线学习行为提供了全新视角。该方向采用定量计算的方法分析人类行为在时间和空间上表现出的统计特征、标度规律和动力学机制(樊超等,2011;周涛等,2013)。由于其在理论和实践中的双重价值,问世不久便得到学者的广泛关注,并逐步应用于流行病传播控制、交通疏导等。在该领域研究中,常用异质性度量概率分布中各个样本偏离中心的程度。若样本服从正态分布,则理论上99.7%的样本会落在$\mu\pm3\sigma$范围内,即样本方差较小(同质或均匀),偏离均值极大的样本出现概率极低,可以忽略。而大量人类动力学研究的成果揭示人类行为的很多统计量偏离传统假设的正态分布或者泊松分布,表现为概率密度函数为$p(x)=Pr(X=x)=Cx^{-\lambda}$的幂律分布的统计规律。幂律分布的衰减速度较上述两种分布慢,使得样本方差非常大(即异质或非均匀),且大数值样本的出现成为必然且不可被忽略,例如社会中的高收入人群。

本文采用人类动力学的研究方法对一组国内MOOC平台的抽样数据进行分析。研究结果表明,用户的在线学习行为具有明显的异质性、表现出阵发和重尾的特征,即用户学习的活跃性会随时间的推移发生变化,一方面会有大量用户退课,另一方面坚持下来的学习者却有更多的在线时间。

二、数据描述与统计特征

2013年10月,清华大学面向全球正式发布在线学习平台“学堂在线”(www.xuetangx.com)。经过两年的发展,截至到2015年9月23日,该平台注册用户数破百万,选课数超251万人次,成为国内最大的中文MOOC平台之一。2015年5月,“学堂在线”与国际计算机学会的数据挖掘及知识发现会议合作,为2015国际知识发现和数据挖掘竞赛提供用户在线学习数据,开展以在线学习行为中退课现象的研究与预测为目的的大数据竞赛。本文所采用的数据即为该竞赛的公开数据。

该匿名化的抽样数据包括了“学堂在线”平台有过学习记录的部分活跃用户的选课记录和行为记录,包括112,448名用户,39门课程,选课记录总计200,905次,时间跨度为278天(从2013年10月27日至2014年8月1日,精确到秒)。每个用户至少选过一门课程,而每门课程至少被选过一千次。同时数据还包含用户每次选课过程中的详细学习行为记录,涉及七个类别:access(访问课程内容)、discussion(参与课程论坛讨论)、navigate(访问课程的其它部分)、problem(完成课程作业)、video(观看视频)、wiki(访问课程wiki)和page_close(关闭网页)。详细行为记录多达13,545,124条,使我们能够全面了解用户使用在线学习系统的行为方式(见图1)。

在选课系统中,如果将用户和课程分别看作网络中的两类节点,选课关系看作连边,那么可以构建出用户与课程之间的二部图网络。在网络中,将节点拥有连边的数量称为度,而所有节点的度的概率分布称为网络的度分布(吴金闪等,2004),是衡量网络结构的核心指标。如果网络各个节点度的差异很大,则称网络具有异质性,即非均匀性。

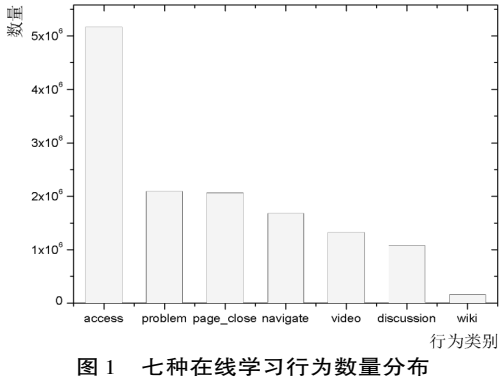

图1 七种在线学习行为数量分布

在选课网络中,用户节点的度表示用户选过课程的数量,而课程节点的度表示该课程被多少用户选过(用户不会多次选择同一门课程)。在用户度的概率分布图中(见图2),曲线可以用带指数截尾的幂律分布函数来拟合,即$y=0.869exp(-0.285x)x^{-1.456}$。因此用户度分布具有明显的异质性,即大部

分用户的选课量非常少,选择1门或2门课程学习的用户占总数的83.4%。约11万用户中只有937个用户选过至少10门课程,80个用户选过多达20门及以上课程,6个用户选过多达30门以上课程。由于样本数据涉及的课程数量较少,统计发现39门课程的选课量都不相同,因此本研究采用直方图给出课程的度分布(见图3)。由此可知,不同课程选课人数之间的差异显著。39门课程中,16门课程的选课次数不足3000次,而选课量超过10000次的5门课程的选课量总和占到了选课总量的39.8%。显而易见,如果进一步考虑那些选课量不足1000的冷门课程,"学堂在线"平台课程的热门程度差异会更明显,即少数课程占据了大多数的选课量,而很多课程鲜有人问津。

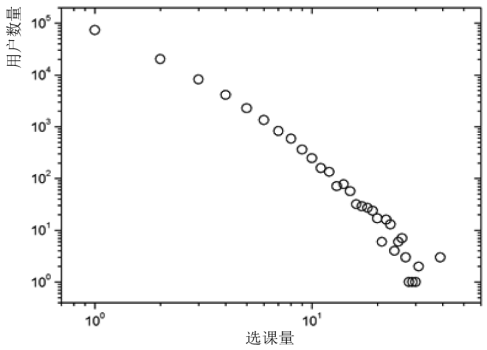

图2 用户度分布

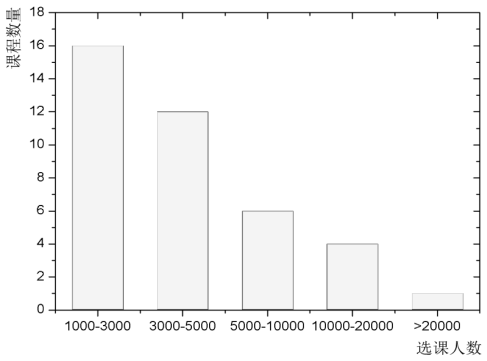

图3 课程度分布

总之,用户和课程的度分布表现出明显的异质性,符合经济学的长尾理论,因而在线学习系统既要精心开发热门课程,又要重点服务深度用户。

三、在线学习行为的时间标度律

(一)时间间隔分布

时间间隔指某事件或者行为在两次发生期间间隔的时间,是人类动力学研究的核心统计量。该分布律可以度量人类行为产生的频率,为理解人类行为的产生机制并为设计随机服务系统提供理论依据。本文从系统和行为两个层面统计了在线学习行为的时间间隔分布。前者指不区分用户及其选课,从学习系统的角度将全部用户的所有学习行为按发生时间顺序排列计算相邻行为之间的间隔时间;后者指一次选课在同一天中出现的各个行为之间的间隔时间。二者的结果可见图4和图5。可以看出,不论是在系统层面还是学习层面,时间间隔分布曲线的主体部分在双对数坐标下都近似为一条直线,可以由幂律函数 $y=cx^{-\lambda}(\lambda>0)$ 拟合,幂律指数 $\lambda\approx2.090$(系统层面)和 $\lambda\approx1.440$(学习层面)。这说明人类行为的发生具有阵发和重尾特征,即短时间内的密集爆发和长时间的沉寂交织。并且由幂指数的大小可知,行为层面的异质性更强,即分布更宽广、行为发生更不均匀。这是由于对同一用户来说,各种形式的学习行为必然会耗费一定的时间,因此各个行为之间会存在较多的空闲期;而当个体行为汇聚到群体层面时,这些空闲期会被其他用户行为所填充,因此较长时间间隔出现的概率会降低,故分布的异质性会弱一些,即幂律指数更小。

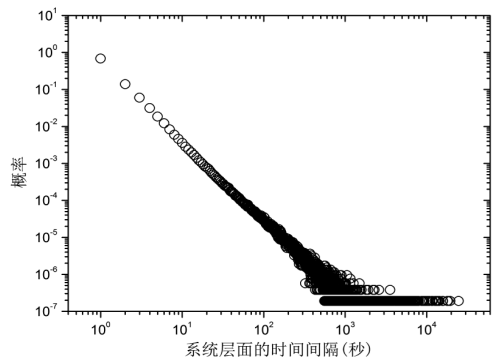

图4 系统层面的时间间隔分布

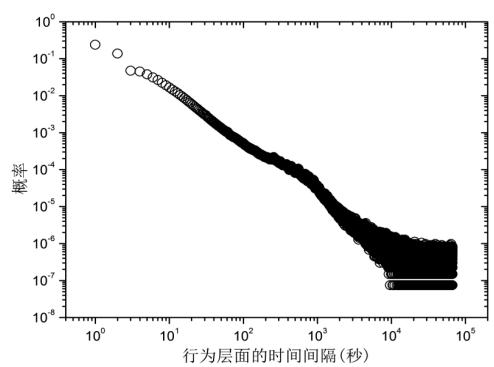

图5 行为层面的时间间隔分布

(二)持续时间分布

如果将在线学习系统看成随机服务系统的服务台,则用户可以视为到达的客户,持续时间是系统给

每个学习行为提供服务的时间。该指标会对服务系统的运行效率产生重要影响。当持续时间较短,用户等待时间和系统队列长度会较短。本文分别对七种学习行为在同一天中的持续时间进行计算(不区分用户和课程),发现它们均服从幂律分布,即大部分学习行为的持续时间很短,只有少数行为持续时间较长。限于篇幅,图6只给出discussion行为的持续时间概率分布,幂律指数$\lambda \approx 1.546$。因而,持续时间分布也具有很强的异质性,81.2%的讨论时间不超过5分钟(不足1分钟的计为1分钟)。

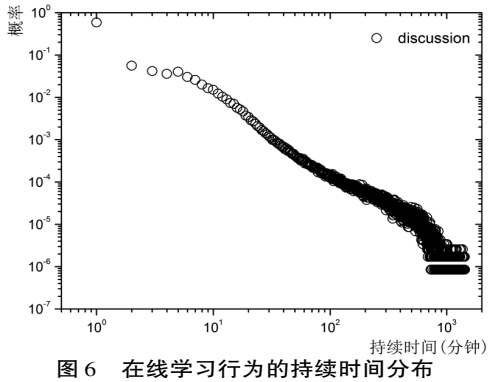

图6　在线学习行为的持续时间分布

四、在线学习行为的活跃性

(一)活跃性的静态特征

“学堂在线”平台的课程对上课时间的安排分为两类:一类课程具有固定的开课时间,每周若干次,另一类课程没有固定时间,可以随时参与或退出。因此,用户学习行为的活跃性在一定程度上受到课程模式的影响。由于数据没有对开课方式进行说明,由此不作区分地统计了每次选课行为中活跃天数的分布,即用户在该门课程的学习过程中有多少天曾经上线学习过。统计结果如图7所示,活跃天数分布在线性对数坐标下呈一条直线,可用指数函数$y=0.152exp(-0.266x)$拟合,因而活跃天数分布相对前文所述的幂律分布来说更具均质性,即绝大多数的选课行为持续天数非常短(95.1%的课程的上课天数不超过10天)。这一现象由两方面原因造成,一是那些不限定时间的课程持续天数较短,用户短时间内密集学完课程内容;二是学习的退率现象比较严重,大多数用户没有坚持学完全部课程。

为了进一步佐证这一猜想,本研究计算每次选课的学习时间分布,即用户在线参与学习的总时长,由用户在同一天上线学习的最早时间和最迟时间之差计算得到,不足1小时的计为1小时,结果如图8所示。该曲线以$\lambda \approx 0.969$的幂律函数形式下降,说明学习时间分布有明显的非均匀性,78.0%的学习时间只有5小时或更少。随后,曲线在超过约15小时后加速衰减,说明较长时间发生的概率极低但真实存在,表明大量学习者听过几天课程后退出学习。这样大的差异也是在线学习行为异质性表现。

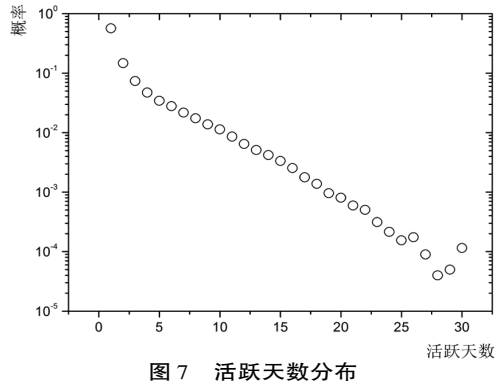

图7　活跃天数分布

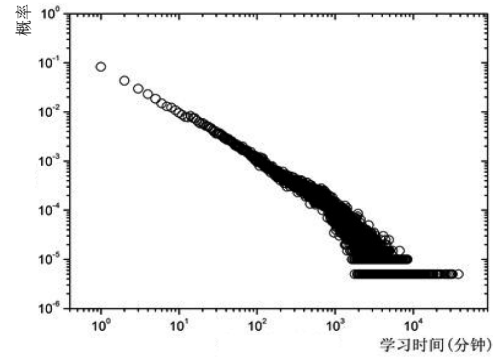

图8　在线学习时间分布

考虑到有些用户会在一天的不同时段分散参与学习,使得上述算法产生偏差。对此,我们以一次选课中所有七种学习行为的次数总和来代替学习时间。结果显示,行为次数的概率分布同样表现为与图8类似的幂律分布,$\lambda \approx 1.217$。进一步对比行为次数和学习时间,发现两者的皮尔逊相关系数高达0.7679,即两个统计量高度线性相关。因此两者对于描述学习行为的活跃性来说有类似的统计意义。

(二)活跃性的动态演化

由上面分析可知在线学习行为在整体层面上有着极大差异,那么活跃性在时间维度上又有哪些特征?随着时间的推移用户学习的活跃性如何变化?本文用相对时步τ_i代替真实时间描述平均在线时间随时步推移的演化规律。相对时步τ_i表示用户i在这次选课行为的第τ个活跃日;在线时间以分钟为单位,对这一天出现的所有在线时间求平均值。

这样变换可以将真实时间不同的所有学习行为标度到同一虚拟时间尺度上(见图9)。虽然我们已经知道课程后期阶段会有更多用户退出学习,但平均在线时间是随着时间推移而单调上升,即用户在课程后期的学习花费更多的时间。这一现象也符合教学过程的一般规律,即课程后期特别是临近结课和考试时需要学习者投入更多的精力和时间。

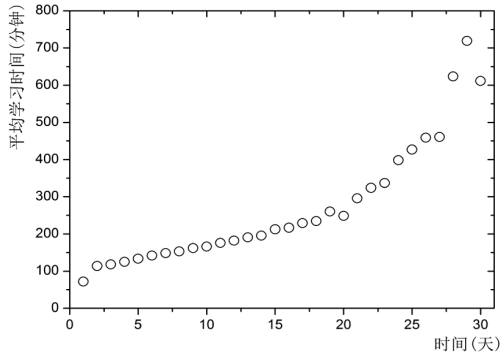

图9 在线时间推移演化规律

为了进一步挖掘这一现象的原因,我们统计了每次选课中每种学习行为在第 τ_i 天发生的次数(见图10)。排除曲线尾部由于样本点较少而出现的较大波动,多数学习行为在整个学习过程中次数没有明显变化,只有discussion行为的数量出现大幅度上升。discussion曲线以指数函数 $y = 6.110\exp(0.068x)$ 的形式上升,其它曲线均符合 $y=ax+b(a<1)$ 的亚线性增长,其中access曲线斜率 $a\approx0.250$,而其余曲线增长率均小于0.150。由此认为,学习阶段后期在线时间的上升是由于参与讨论和访问次数增多所致。

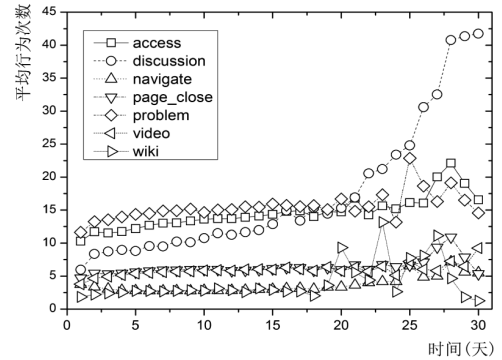

图10 七种在线学习行为次数随时间推移演化规律

五、总结与讨论

从宏观角度看,在线学习行为表现出复杂和矛盾的特征,即多数用户所选课程不多,但有大比例的用户投入的学习时间较少,甚至没有坚持完成整个学习过程。这说明与传统课堂教学相比,在线学习方式仍然缺乏足够的监督与激励机制。但是,对于那些对课程内容有真正需求的学习者来说,他们不仅能够坚持学完课程,还会在后半段投入更多的时间和精力,这与传统学习方式一致。

本研究揭示了在线学习行为的统计规律和动力学机制,有助于理解人类学习行为的产生机制,人们利用在线学习平台指导用户的学习行为,做好课程规划和设置,并且为进一步分析在线学习行为的退课现象打好基础。当然,由于数据获取渠道的局限性,本文无法针对用户的个人属性展开分析。在大数据时代到来的契机下,越来越多的教育数据被保存,使得我们有机会将学生的课堂表现、考试成绩、课外表现甚至毕业后的发展统一起进行宏观考量。教育大数据作为一个新兴的研究方向必然会为教育学研究带来新思路、途径和机遇。

[参考文献]

[1] Barabasi,A. L. (2005). The origin of bursts and heavy tails in human dynamics[J]. Nature, 435: 207-211.

[2] Blanchard,P., & Hongler,M. O. (2007). Modeling human activity in the spirit of Barabasi's queueing systems[J]. Physical Review Letters, 75(2): 026102.

[3] 蔡文芳(2015). 我国MOOC课程完成情况分析及其对策——基于果壳网、问卷星等网络调查[J]. 职业技术教育,36(17): 21-23.

[4] 陈肖庚,王顶明(2013). MOOC的发展历程与主要特征分析[J]. 现代教育技术, 23(11): 5-10.

[5] 樊超,郭进利,韩筱璞,汪秉宏(2011). 人类行为动力学研究综述[J]. 复杂系统与复杂性科学, 8(2):1-17.

[6] 方旭(2015). MOOC学习行为影响因素研究[J]. 开放教育研究,21(3): 46-54.

[7] Guo,J. L., Fan,C., & Guo Z. H. (2011). Weblog patterns and human dynamics with decreasing interest[J]. European Physical Journal B, 81(3):341-344.

[8] 高地(2014). MOOC热的冷思考——国际上对MOOCs课程教学六大问题的审思[J]. 远程教育杂志,(2):39-47.

[9] Han,X. P., Zhou,T., & Wang,B. H. (2008). Modeling human dynamics with adaptive interest[J]. New Journal of Physics, 10(7): 073010.

[10] 胡小勇,李丽娟,郑晓丹(2015). 在线环境下学习者协作解决问题的策略研究[J]. 远程教育与网络教育,(1):44-50.

[11] 胡勇,王陆(2006). 在线学习者的自我评价与反思研究[J]. 开放教育研究,12(2):69-73.

[12] 胡勇,赵凤梅(2015). 在线学习成效的理论分析模型及测

量[J]. 网络教育与远程教育,(10):37-45.

[13] 蒋卓轩,张岩,李晓明(2015). 基于MOOC数据的学习行为分析与预测[J]. 计算机研究与发展,52(3):614-628.

[14] 李曼丽,徐舜平,孙梦嫽(2015). MOOC学习者课程学习行为分析——以"电路原理"课程为例[J]. 开放教育研究,21(2):63-69.

[15] 李楠楠,周涛,张宁(2008). 人类动力学基本概念与实证分析[J]. 复杂系统与复杂性科学,5(2):15-24.

[16] 李松,张进宝,徐玲(2010). 在线学习活动设计研究[J]. 现代远程教育研究,(4):68-72.

[17] 龙三平,张敏(2014). 在线学习理论研究的现状与趋势——基于SSCI数据库(1994-2013年)的科学计量分析[J]. 远程教育杂志,(3):64-70.

[18] Malmgren, R. D., Stouffer, D. B., Campanharo, A. S. L. O., et al (2009). On univ ersality in human correspondence activity[J]. Science, 325: 1696- 1705.

[19] Malmgren, R. D., Stouffer, D. B., Motter, A. E. (2008). A Poissonian explanation for heavy tails in email communication[J]. PNAS, 105(47): 18153- 18158.

[20] Vazquez, A., Oliveira, J. G., Dezso, Z. (2006). Modeling bursts and heavy tails in human dynamics[J]. Physical Review E, 73(3): 036127.

[21] 王红艳,胡卫平(2013). 中国在线学习研究现状与启示[J]. 中国远程教育,(8):30-34.

[22] 吴金闪,狄增如(2004). 从统计物理学看复杂网络研究[J]. 物理学进展,24(1): 18:46.

[23] 肖为胜,方志军(2009). 在线学习模式浅探[J]. 教育学术月刊,(6):108-109.

[24] 杨素娟(2011). 在线学习社区的组织气氛探析[J]. 中国远程教育,(6):64-68.

[25] 袁松鹤,刘选(2014). 中国大学MOOC实践现状及共有问题——来自中国大学MOOC实践报告[J]. 现代远程教育研究,(4):3-12.

[26] 周涛,韩筱璞,闫小勇(2013). 人类行为时空特性的统计力学[J]. 电子科技大学学报(自然科学版),42(4):481-540.

(编辑:李学书)

Human Dynamics Analysis on MOOC Online Learning Behaviors

FAN Chao[1] & ZONG Liyong[2]

(1. *College of Arts and Sciences, Shanxi Agricultural University, Taigu* 030801, *China*
2. *Department of Art Management, Shanghai Publishing and Printing College, Shanghai* 200093, *China*)

Abstract: *Online learning is such a developing field. utilizing learning platforms such as MOOCs, users can carry out learning at any time and any places. but due to the deficiency of supervision and motivation, lots of learners quit the learning process. Therefore, studies on behavior patterns of online learners can provide great help to further understand online learning behaviors, design learning platforms and make learning rules. the emergence of big data era provides us an opportunity to study online learning behaviors from a new perspective. Taking advantages of the complete and accurate dataset, we are able to explore deeper into learning mechanism and regularity. This paper conducted a quantitative analysis using the methodology of human dynamics, which is used to study the statistical property, scaling law and dynamical mechanism of human temporal-spatial behaviors. Via analyzing an open dataset from a Chinese MOOC platform, we found a heterogeneity or inhomogeneity in online learning behaviors, specifically from the following three perspectives: both the number of courses one user selects and the number of users who select one certain course show great discrepancy; the distributions of both the inter-event time and duration time of online learning follow power-law decay, indicating the burst and heavy tail property of online learning behaviors; the activeness of online learning behaviors shows a discrepancy between different users and different stages. Specifically, users spend different days and time on the learning process. Lots of users drop out from learning although the users who persist spend more time and put more effort on the courses The conclusions hopes can facilitate a better understanding of the statistical characteristics of online learning behaviors and provide explanations for people who drop out from learning.*

Key words: *human dynamics; Big Data; temporal scaling law; duration time; activitiveness*