MOOC 学习日志大数据管理

贺欢，郑庆华，董博
（西安交通大学陕西省天地网实验室）

摘要：由美国著名高校发起的大规模开放在线课程（MOOC）在全球掀起了前所未有的学习热潮，吸引了数以千万计的学习者参与．伴随着 MOOC 的广泛应用，源源不断产生的海量学习日志形成了 MOOC 大数据，并对相关的数据分析领域提出了重大挑战．为充分利用 MOOC 大数据的价值以满足 MOOC 发展与研究的需求，本文介绍了 MOOC 的发展状况并对现有研究的重要问题进行了总结分析，提炼了 MOOC 大数据的特点，在此基础上结合实际业务设计了 MOOC 大数据的处理框架和典型应用，最后提出了 MOOC 大数据管理面临的挑战．

关键词：大数据管理；教育数据挖掘；学习过程分析；日志分析
中图分类号：TP391

Management of MOOC Big Log

He Haun, Zheng Qinghua, Dong Bo
(Xi'an Jiaotong University SPKLSTN Lab)

Abstract: An unprecedentedly global spreading fashion of learning was initiated by MOOCs, which were launched by famous universities in America and attracted millions of learners. Along with wide application of MOOC, the MOOC big data came into being while large amount of log data was continuously generated, and it presented a significant challenge to fields related to educational data analysis. In order to satisfy the demands of development and researches of MOOC and fully utilize the value of MOOC big data, this paper introduces the development situation of MOOC, summaries important questions in current researches, extracts features of MOOC big data, designs processing framework and typical applications for MOOC big data combined with practical business, and finally proposed challenges of MOOC big data.

Key words: MOOC; big data management; educational data mining; learning analytic; log analysis

0 引言

自 2011 年开始，由斯坦福大学、哈佛大学和麻省理工学院等美国著名高校发起的大规模开放在线课程（Massive Open Online Course，MOOC）将大量优秀的教学资源以全新的形式提供给全球学习者，使任何人都能够体验国际一流高校的授课过程和教学资源，在世界范围掀起了前所未有的学习热潮[1]．教育界、工业界和政府机构都对 MOOC 投入大量资源，使 MOOC 在平台、课程和学习者规模上迅速全面发展．MOOC 开课大学和主流 MOOC 服务商 Coursera 的统计数据显示，MOOC 课程的参与人数通常以万计[2][3]，目前

基金项目：国家自然科学基金（61532015，61502379）；高等学校博士学科点专项科研基金（20130201130002）
作者简介：1984，男，博士生，主要研究方向：大数据存储管理，教育数据挖掘
通信联系人：1969，男，博士，教授，博导，主要研究领域为智能 e-Learning 环境的理论与技术、软件可信性评测. E-mail: qhzheng@mail.xjtu.edu.cn

Coursera 最受欢迎的课程有超过 24 万名学习者同期参与学习[4]，远超过任何传统课堂的同期学习人数．《纽约时报》在 2012 年 11 月发表文章“The Year of the MOOC”[5]，介绍了 MOOC 在 2012 年的发展情况，使民众了解到 MOOC 的出现以及 MOOC 给教育行业带来的重大变革．

在学术界，相关的研究也在迅速拓展．《Science》在 2013 年 4 月推出专刊“Grand Challenges in Science Education”[6]，讨论了在网络环境中使用学习者数据分析、社交网络分析以及虚拟实验室等新技术为科学教育提供支持的方法和案例，同年 10 月发表社论“Bricks and MOOCs”[7]提出了 MOOC 能够成为终生教育的一种有效媒介，并且高校应该充分利用这种创新的在线教育方式以增加人们接受高等教育的机会．2013 年 7 月，《Nature》推出专刊“Learning in Digital Age”[8]，讨论了 MOOC 在美国高校以及全球各地的应用，介绍了大数据分析技术在在线教育（e-Learning）过程中起到的作用[9]，提出了 MOOC 正在改变高等教育并将其称为 Campus 2.0[10]．2013 年 9 月，英国教育技术中心与互操作性标准组织（JISC Centre for Educational Technology and Interoperability Standards）发布白皮书《MOOCs and Open Education: Implications for Higher Education》[11]，用破坏式创新理论分析 MOOC 对当前高等教育的影响．美国的斯隆联盟（Sloan Consortium）在 2013 年和 2014 年发表的美国在线教育上一年度综述报告[12][13]，用大量篇幅介绍了对美国和全球 MOOC 发展情况的调研，分析了参与 MOOC 的高校、课程、学习者以及整体的发展趋势，并提出了 MOOC 所面临的挑战．2013 年 12 月，由 IEEE Education Society 支持的国际会议 IEEE International Conference on MOOC, Innovation and Technology in Education 召开；同期还有 IEEE Frontiers in EducationConference，International Conference on Educational Data Mining 等国际会议召开，都对 MOOC 相关问题展开了讨论．

另一方面，2012 年 10 月美国教育部发布报告“Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics”[14]，提出了 e-Learning 研究中的两个重要领域，即教育数据挖掘（Educational Data Mining，EDM）和学习过程分析（Learning Analytics，LA）都能从大数据中得到支持，并建议相关机构利用大数据技术对教育数据进行研究．2013 年 2 月美国教育部再次发布研究报告“Expanding Evidence Approaches for Measuring Learning in a Digital World”[15]，提出在未来的 e-Learning 中需要利用大数据方法，建立支持个性化学习的自适应学习系统，为学习者提供深度学习所需的学习资源．2013 年 6 月美国奥巴马政府发布了“连接教育倡议”（ConnectED Initiative）[16]，联合联邦通讯委员会（Federal Communications Commission）投入 20 亿美元用于高速宽带网络建设，为校园内的在线学习提供基础设施的保障．

MOOC 作为在线教育的一种创新形式，不仅为全世界的学习者带来了优秀的教学资源、教学理念和教学方法，同时其产生的海量学习日志数据也为研究者带来了巨大挑战．我们认为在上述背景下 MOOC 产生的学习日志数据必然是大数据，并且由于 MOOC 自身的特点以及与其它领域大数据的差别，应当针对 MOOC 研究其管理方法．本文将从分

析 MOOC 的特点与 MOOC 产生的学习日志数据数据的特点出发，对现有的 MOOC 与传统 e-Learning 的数据分析相关研究进行总结，结合西安交通大学网络教育学院的实际教学平台与分析需求，提出 MOOC 学习日志大数据的处理框架并结合实际应用场景进行分析，最后讨论 MOOC 大数据管理所面临的挑战.

1 MOOC 及 MOOC 大数据

1.1 MOOC 的兴起与现状

2011 年斯坦福大学发布了首门 MOOC 课程“人工智能导论”（Introduction Into Artificial Intelligence），吸引了全球超过 16 万名学习者加入[10]. 随后，哈佛大学、麻省理工学院等高校也推出了各自的 MOOC 课程. 之后 MOOC 迅速发展，Coursera、edX 等机构相继成立，为更多的大学提供 MOOC 的基础软硬件服务.

表 1 Cousera 和 edX 平台上参与 MOOC 的国家与学校

Tab. 1 Participations of Nations and Universities on Coursera and edX

地区	国家	学校数量	代表性学校
美洲	美国	80	哈佛大学、斯坦福大学、麻省理工学院、加州理工大学、普林斯顿大学、芝加哥大学等
	加拿大	5	多伦多大学、阿尔伯塔大学、麦克马斯特大学、麦吉尔大学、英属哥伦比亚大学
	其他	5	墨西哥：国立自治大学、蒙特雷科技大学；巴西：圣保罗大学、坎皮纳斯州立大学等
欧洲	瑞士	6	苏黎世大学、洛桑联邦理工学院、日内瓦大学、洛桑大学、苏黎世理工学院
	法国	5	巴黎高等师范大学、巴黎高等理工学院、ESSEC 商学院、巴黎高等商学院、巴黎中央学院
	英国	4	爱丁堡大学、曼彻斯特大学、伦敦大学等
	荷兰	4	阿姆斯特丹大学、莱顿大学、埃因霍温科技大学、代尔夫特理工大学
	其他	16	丹麦：哥本哈根商学院、哥本哈根大学、丹麦技术大学；德国：慕尼黑工业大学、慕尼黑大学； 西班牙：西班牙 IESE 商学院、巴塞罗那自治大学、IE 商学院；瑞典：隆德大学、卡罗林斯卡医学院； 俄罗斯: 高等经济学院、莫斯科物理科学与技术学院、圣彼得堡国立大学； 意大利：罗马大学、博科尼大学；比利时：天主教鲁汶大学
亚洲	中国	7	北京大学、清华大学、复旦大学、上海交通大学、香港中文大学、香港科技大学、国立台湾大学
	以色列	3	希伯来大学、特拉维夫大学、以色列理工学院
	日本	2	东京大学、京都大学
	其他	6	新加坡：新加坡南洋理工大学、新加坡国立大学；韩国：韩国科学技术高级研究院、首尔国立大学； 土耳其：土耳其科驰大学；印度：印度理工学院孟买分校
大洋洲	澳大利亚	5	墨尔本大学、西澳大利亚大学、澳大利亚新南威尔士大学、昆士兰大学、澳大利亚国立大学
合计		148	

随着 MOOC 的不断发展，不同国家、语言、学科的 MOOC 课程都汇聚在公共互联网学习平台上向全世界免费开放；另一方面全世界的学习者可以根据自己的爱好和需求选择课程，同世界各地的其它学习者一起学习，如此国际化的 e-Learning 活动是前所未有的．截至 2014 年 10 月 10 日，仅 Coursera 和 edX 两个国外主流 MOOC 平台就有全球各大洲的 23 个国家的 148 所大学参与其中[17][18]，如表 1 所示．2013 年起，学堂在线和中国大学 MOOC 等国内机构开始提供 MOOC 课程，截至 2014 年 10 月 10 日共有 29 所大学在这两个平台上提供了 362 门 MOOC 课程[19][20]．

每所参与 MOOC 的高校都提供了多门 MOOC 课程，涵盖遗传学[3]、护理学[21]、营养学[22]、计算机科学[23]、哲学与社会学[24]等不同学科，面向不同层次的学习者，为他们提供开拓视野的渠道．由于互联网的开放性和这些课程设计的创新性，每门课程的学习者数量都远远超过传统课堂或传统 e-Learning 课程的学习人数．根据各个大学的统计[3]，MOOC 的学习者规模极为庞大，即使参与人数较少的课程也有过万人参加，部分热门课程更有超过十万人同期学习．表 2 列举了 Coursera 和 edX 平台上部分类别的课程数量和代表性课程，以及课程的开设院校和同期选课人数．

MOOC 课程的利用 Web 交互技术实现了多种教学功能开展教学，如表 3 所示．例如麻省理工学院开设的“电路与电子学”（Circuits and Electronics）课程中，基于 HTML5 技术开发了在线电路模拟实验室功能，使学习者能够在线练习搭建电路，学习相关的物理原理[25]．

表 2 edX 和 Cousera 平台上提供的课程分类及选课人数

Tab. 2 Statistic for Course Categories and Registrants on edX and Coursera

课程分类	课程数量	课程示例	开设院校	选课人数
计算机、信息工程	437	Introduction to Artificial Intelligence[10]	斯坦福大学	>160000
物理、地球科学	222	Circuits and Electronics[26]	麻省理工学院	>154000
商业和管理	127	Introduction to Operation Management[3]	宾夕法尼亚大学	>110000
哲学与道德	22	Introduction to Philosophy[24]	爱丁堡大学	>98000
历史	54	A History of the World since 1300[27]	普林斯顿大学	>83000
人文	211	Critical Thinking[24]	爱丁堡大学	>75000
数学	96	Calculus: Single Variable[3]	宾夕法尼亚大学	>60000
艺术	68	The Ancient Greek Hero[28]	哈佛大学	>43000
教育学	121	E-Learning and Digital Cultures[24]	爱丁堡大学	>42000
统计和数据分析	85	Statistics for Public Health[29]	哈佛大学	>31000
社会科学	185	Health Policy and the Affordable Care Act[3]	宾夕法尼亚大学	>30000
化学	45	Solid State Chemistry[29]	加利福尼亚大学伯克利分校	>24000
政治金融	122	The Challenges of Global Poverty[30]	麻省理工学院	>19000
生命科学	139	Bioelectricity: A Quantitative Approach[31]	杜克大学	>12000
医学、药学、营养学	136	Rationing and Allocating Scare Medical Resources[3]	宾夕法尼亚大学	>10000

注：课程分类与数量来自 Coursera 和 edX 官方提供的选课列表，一门课程可能同时属于多个类别，故而不同类别的课程数量总和大于总课程数量

根据 Coursera 的公开数据，截至 2014 年 1 月 17 日已经有来自 190 个国家超过 2200 万学习者参与 MOOC 学习[4]．而在 edX，截至 2014 年 9 月 28 日，仅哈佛大学开设的 27 门课

程就有来自 195 个国家超过 156 万学习者注册[28]，其生源分布具体如表 4 所示，其他课程与这的统计数据相似[29].

表 3 MOOC 平台目前提供的主要教学功能

Tab. 3 Major Teaching Functions Provided by MOOC Platform

功能	功能介绍
短课件视频[32]	以知识点为中心组织的时间较短的教学视频
视频中测试[33]	在观看教学视频的过程时进行随堂测验
在线测验与考试[34]	在线组织的多种题型的试卷
讨论组[35]	按话题划分的在线讨论
在线阅读资料[36]	提供参考资料和 Wiki
虚拟实验室[37][38]	物理、化学、计算机编程等课程的在线模拟实验
作业结对评估[39]	学习者根据参考答案互相评价作业

表 4 哈佛大学 27 门 MOOC 的全球注册量分布

Tab. 4 Global Demographics of Registrants in 27 MOOCs provided by Harvard University

国家	参与人数	百分比
美国	558,557	35.6%
印度	138,035	8.8%
英国	68,118	4.3%
加拿大	55,174	3.5%
巴西	44,566	2.8%
中国	40,791	2.6%
澳大利亚	32,483	2.1%
西班牙	29,518	1.9%
德国	27,650	1.8%
法国	22,993	1.5%
其他国家	549,649	35.1%
合计 195 个国家	1,567,534 名学习者	

1.2 MOOC 学习日志数据

MOOC 课程向全世界的学习者提供开放服务，因而其不仅有机会向全世界展示该学校的教学水平与特色，同时也意味着对课程团队的巨大挑战，包括：在开课前如何编排课程结构；在课堂中如何与全球学习者交流讨论；在课下如何评估作业和最终成绩；在结课后如何从整个教学过程中发现学习者的学习模式．为了应对这些挑战，就需要对 MOOC 所产生的学习日志数据进行挖掘分析．由于 MOOC 学习者的规模性与教学过程丰富性的相乘效果，使得其产生的学习日志数据数量远远超过传统课堂以及传统 e-Learning 的日志数据数量．以麻省理工学院提供的统计数据为例[25][26][40]，2012 年秋季开设的“电路与电子学”课程在一学期内为来自全球 194 个国家超过 15 万名学习者提供了服务，从其学习管理系统、论坛、在线测验和虚拟实验室等系统中产生了超过 2 亿 3000 万条的交互记录，生成了超过 38000 个日志文件，累计约 70GB 的学习者操作原始数据，对其预处理并结构化后有依然有超过 1 亿 3859 万行记录．参考表 2 与表 4 可知，基于类似平台的其他 MOOC 课程的学习日志数据量也具有相当的规模．

根据比较有代表性的大数据 3V 定义[41]，大数据应具有规模性（Volume）、多样性（Variety）和高速性（Velocity）．结合国际数据公司（International Data Corporation，IDC）的 4V 定义

[42]和 IBM 的 4V 定义[43]，大数据还需要有价值性（Value）和真实性（Veracity）．首先，表 1、表 2 与表 4 展现的学习者规模和上述 MOOC 课程的学习日志数据量，可以说明 MOOC 的学习日志数据满足大数据定义的规模性特点；其次，根据表 2 与表 3 展现的丰富的课程类型和多种教学手段，MOOC 的学习日志数据也具有大数据定义的多样性特点；并且，学习者数量的不断增加带来了并发学习人数的增长，使得 MOOC 的学习日志数据也具有高速性特点；最后，MOOC 学习日志数据是对现实发生的学习行为的准确刻画，其必然具有真实性，并且其中蕴含了关于学习者、课程设计乃至教育方法的丰富信息，其必然也具有价值性．

综上所述，MOOC 的学习日志数据是大数据，我们称之为 MOOC 学习日志大数据（MOOC Big Data，以下简称 MOOC 大数据）．应当使用大数据的管理方法和工具对 MOOC 大数据进行处理分析．另外，MOOC 大数据除了具有上述共性的特征以外，还有一些由于 MOOC 自身特性所带来特点：

1）学习日志数据维度极为丰富．除了其他领域大数据所有具有日志信息维度以外，MOOC 大数据还蕴藏了来自知识领域、认知领域和教学领域的深层维度信息．具体而言在知识领域，MOOC 大数据包含了学科、专业、课程、知识单元等知识层次结构；在认知领域，MOOC 大数据则包括了学习动机、心理状态、情感状态、认知能力等众多要素[45][46]；在教学领域，则包括了表 3 所列举的各种手段产生的不同类型．这些深层维度不但独立存在且相互之间存在影响，再综合持续数周的学习时间维度[3]，因而虽然 MOOC 大数据的 GB 级别[26]数据规模远小于社交网络或电子商务等大数据场景中 TB 乃至 PB 级别的数据规模[44]，但是 MOOC 大数据的分析难度丝毫不亚于其他领域的大数据．

2）数据分析过程极为多样．如表 2 所示，MOOC 课程涵盖了大量不同的学科领域，一方面不同学科不同课程的知识体系和教学方法各不相同，另一方面同一门课程不同学校的教学过程也是各有特点．MOOC 大数据的数据分析需要针对具体的课程特征展开研究，课程的内在差异使得分析过程显著不同．

3）学习者差异极大．如表 4 所示 MOOC 学习者来自于世界各地，具有不同的文化背景，跨越各个年龄层次，掌握不同的知识水平，并且由于 MOOC 大数据的数据采集紧密深入学习者的全学习过程，使得学习者的数据能够在多个角度产生显著的差异，为 MOOC 大数据带来了传统 e-Learning 所不具备的多种类型的学习者个体与群体．

上述特点的存在，使得需要充分了解 e-Learning 领域数据分析研究的方法与技术，进而为设计 MOOC 大数据处理框架提供依据．

2 MOOC 大数据分析的相关研究

MOOC 大数据处理与其他大数据处理的流程基本一致，包括数据抽取、数据集成、数据分析、数据解释这 4 个组成部分[49]．其中数据分析是大数据处理流程的核心，体现了不同大数据业务的特点，因而首先需要了解 MOOC 大数据中数据分析的研究需求与分析技术．

MOOC 作为 e-Learning 的一种实现方式，具有和传统 e-Learning 学习日志数据分析相似的研究需求，传统 e-Learning 的数据分析中使用的技术也可以在 MOOC 大数据的数据分析中借鉴使用．因而本文将从 MOOC 和传统 e-Learning 两个方面总结数据分析的研究需求与分析技术，为构建 MOOC 大数据处理框架提供参考．

传统 e-Learning 的数据分析研究集中在 EDM 和 LA 这两方面．EDM 是指利用数据挖掘、机器学习和统计学等方面的技术研究教育相关的数据以发现其内在规律[50]，而 LA 则是

利用学习者产生的数据结合分析模型，用于预测或者建议学习者学习行为的相关研究[51]. EDM 与 LA 之间的差别主要在研究目标，前者侧重研究学习者相关的描述性模式，而后者则侧重研究学习过程的预测性模型[14]. 除此之外两者有很多相似性，比如在技术方面，通常都会综合利用认知科学、统计学和数据挖掘的方法分析数据；在分析的对象方面，也是来自学习者、学习系统与学习环境的相关数据[15][52].

本文参考传统 e-Learning 的 EDM 和 LA 领域的综述文献中采用的分类方法[53][54][55]，从数据分析服务对象的角度，将数据分析的研究需求分为面向学习者、指导者和管理服务的三个方面，如表 5 所示.

为了满足这些研究需求，传统 e-Learning 的数据分析主要使用了统计学与机器学习领域的分析技术，包括显著性检验[56][57][58][59][45]、主成分分析[60][61]、方差分析[62][63][64][65][48][66][34]、T 检验[23][67][65][68]、相关性分析[69][70][71][72][73]和卡方检验[74]等统计学方法，以及回归分析[75][76][77][78][79]、序列模式挖掘[80][81][82]、决策树[83][84][46][85]、关联规则挖掘[86][87][88][89]、聚类分析[90][70][81]、人工神经网络[91][92]、朴素贝叶斯[93]、支持向量机[93]和遗传算法[94]等机器学习方法. 传统 e-Learning 的数据分析通常会同时使用多种技术，并根据具体研究需要对基本分析方法进行优化和改进.

以下将按照表 5 所示的服务对象与研究目标分别阐述 MOOC 与传统 e-Learning 的数据分析相关研究内容和分析技术.

表 5 典型 e-Learning 数据分析的研究需求

Tab. 5 Typical Research Requirements in e-Learning Data Analysis

服务对象	研究目标	研究内容
学习者	学习者建模	基于认知、情感、个性、领域知识、学习者属性等方面的特征对学习者建立分类描述模型
	学习行为建模	基于学习过程中阅读、交流、查询、观看视频等一系列行为活动构建预测模型.
	学习者表现建模	基于学习者的测验成绩、交流表达等构建预测模型.
	异常分析	基于学习者模型、行为模型、表现模型研究和预测学习障碍、学业失败、退课、退学等问题.
	学习资源推荐	基于学习者模型、行为模型、表现模型研究向学习者推荐学习资料、补习活动或者调整学习过程.
指导者	课程制作	为教师制作课件、调整教学进度、安排教学计划提供参考信息和技术.
	教学反馈与评估	为教师提供学习者学习状态的统计与分析，协助教师评价学习者的作业和课程成绩.
管理服务	课程数值统计	研究为相关人员提供指定课程在学习者、学习过程、学习效果等方面的量化统计指标.
	数据可视化	研究为相关人员提供课程统计指标、学习过程分析结果的可视化方法.

2.1 学习者相关研究

MOOC 的学习日志数据以学习者为中心产生，MOOC 大数据的数据分析也以学习者的相关研究为中心展开，以学习者建模为基础，进而构建学习行为模型和学习者表现模型，为推荐学习资源和分析异常情况提供参考依据.

2.1.1 学习者建模

学习者是 MOOC 教学的基础，基于 MOOC 大数据的学习者建模通过对学习者认知能力、个性特点、情感状态等因素的研究，建立反映学习者特征的模型，为 MOOC 的课程设计、教学安排、评价考核等方面的发展提供重要的依据与推动. 学习者建模通常根据认知心理学设计评价学习者的指标，然后利用问卷调查、课程测验等方式得到学习者反馈，进而了

解学习者的学习特征．其中以学习者学习风格的建模为主要方法，包括使用 Kolb 模型、VAK/VARK 模型，Felder-Silverman 模型，LSI（Learning Style Inventory）等评估方法．

MOOC 相关研究

Kimberly 等人[95]采用先验与后验的问卷调查，基于项目相应理论（Item Response Theory，IRT）对麻省理工学院的 MOOC 课程的日志数据进行了分析，发现 MOOC 在线学习的效果略优于课堂学习，但是在交互参与方面要低一些．Michael 等人[96]基于 Rasch 模型和 IRT 对学习者建立可加因子模型（Additive Factors Model），通过赫尔辛基大学提供的 3 门 MOOC 编程导论课程的学习过程所产生的作业、学习等数据，分析学习者的学习模式和潜在的特质．Dagmar[97]讨论了学习风格对于 MOOC 参与情况的影响，以及如何将个人的学习风格融入在课程的设计实现上．Franka 等人[98]基于 Kolb 模型利用问卷调查采集的数据，建立学习者的学习风格模型，并研究不同学习风格的学习者对不同媒体类型学习资源的接受程度，进而为设计能够支持多种学习风格的 MOOC 课程提供参考．

传统 e-Learning 相关研究

Uros 等人[99]基于 Kolb 与 VAK 学习风格分类方法建立学习者模型，并设计了一个基于不同学习类型组合选择各种多媒体材料类型的学习系统．该模型显示学习者偏好结构良好的有色彩的学习文本，并且半球学习风格模型（Hemispheric Learning Style Model）是决定学习者对不同多媒体学习资料的喜好程度的最重要指标．Luke 等人[100]基于自我决定理论（Self-Determine Theory），采用在线问卷调研、访谈的统计数据，了解学习者在 e-Learning 的语言课程中的学习过程与学习困难．Richard 等人[92]提出了一个基于认知诊断和 IRT，采用人工神经网络实现的学习者模型，使用该模型可以模拟学习者在课堂学习活动中的认知过程，观察学习者的科学技能相关的批判性思维能力水平．Luuk 等人[101]利用在线写作平台中学习者的阅读与写作的过程数据，研究不同的学习风格对于完成写作任务过程的影响．Neha[102]基于 BDI 理论构建学习者的认知模型，提出使用基于专家系统的认知代理帮助学习者进行协作学习．Ezequiel 等人[89]基于 Felder-Silverman 模型和关联规则挖掘技术，研究学习者的学习风格和其学习软件工程 Scrum 敏捷方法效果的关系．

2.1.2 学习行为建模

学习行为建模是对学习过程中学习者行为特征构建的模型，利用该模型可以预测学习者行为模式、调整学习安排以适应学习者的行为趋势．在 MOOC 大数据的环境下，学习行为的复杂度明显高于其他大数据环境，因而需要针对 MOOC 学习者的学习行为进行具体分析．

MOOC 相关研究

Nabeel 等人[72]采用统计学方法和可视化分析方法分析 MOOC 课程论坛中的学习者讨论活动，了解学习者的人员组成、讨论模式以及参与情况，并分析其与最终成绩的关系．Miaomiao 等人[35]利用卡内基梅隆大学 MOOC 课程论坛中的学习者讨论行为日志，采用统计学的方法建立基于学习行为的情感模型，分别在课程和学习者个人的层面上分析了情感与课程退出率之间的关系．Jennifer 等人[48]采用统计学方法对麻省理工学院 MOOC 课程的学习日志数据和问卷调查结果进行分析，研究学习者的学习行为与其学习动机、学历水平、地域分布等因素之间的关系．Matthieu 等人[59]采用统计学方法，利用 MOOC 结对评估过程的日志数据以及调查问卷的反馈数据，研究 MOOC 结对评估过程中参与者的特征与影响评估结果的因素．

传统 e-Learning 相关研究

Juan 等人[90]基于 Felder-Silverman 学习风格模型和聚类分析，采用模拟退火算法构建了根据学习行为数据分析和预测学习风格的预测模型，解决学习偏差和认知过载等问题．Romero 等人[86]使用 Weka 和 Keel 系统对 Moodle 系统中学习者操作数据进行分析，挖掘学习者的学习模式信息．Erica 等人[66]使用随机游走模型和最大熵模型对学习者在智能教学系统中的操作进行建模，通过分析学习者在系统中的操作过程，发现学习者的行为控制能力与学习收益之间呈现正相关的趋势，控制能力更好的学习者会取得更好效果．Chien[60]提出使用自组织映射与主成分分析相结合的方法来分析学习行为日志数据，提取具有显著特性的数据模式，进而在合适的时候为学习者推荐学习资源．

2.1.3 学习者表现建模

学习者表现建模主要研究影响学习效果的因素，包括学习效率、课程完成情况、资源利用、时间利用、考试成绩等，以实现预测或者评估学习者的最终学习收益．目前 MOOC 课程的完成率整体较低[103]，意味着不同学习者的学习表现最终有明显差别，为了提高学习者的学习表现，需要建立学习者表现模型帮助学习者了解自己的状况，并协助指导者管理教学活动．

MOOC 相关研究

Suhang 等人[78]采用统计学方法和逻辑回归分类器分析学习者在 MOOC 课程中第一周课堂作业与交互讨论的日志数据，预测学习者取得证书的概率以及所获证书的等级．Julia 等人[23]利用统计先验后验问卷调查和学习行为日志数据，分析 MOOC 学习者的各种因素与最终课程完成率之间的关系．Saif 等人[104]采用统计与可视化分析的方法，研究 MOOC 课程学习者的最终成绩的分布情况，包括学习者的成绩分布，成绩与学历背景的关系等．Ignacioa[105]采用统计学方法研究“信息推动”（Information Nudge）对于学习者学习参与度的影响，进而了解学习者的最终成绩与平时参与程度之间的关系．Nikola[106]使用可视化分析的方法，对 MOOC 课程的注册学习者数量、课程设计等因素与完成率之间的关系进行定性研究，发现 MOOC 课程的时间安排也是影响学习者最终学习表现的因素．

传统 e-Learning 相关研究

Chih-Kai 等人[107]研究了 Web 讨论组协作学习过程中功能角色对于学习者学习表现的影响，并基于文本挖掘和分类技术分析构建了一个功能角色检测工具用于协助教师识别讨论组中的功能角色，进而通过调整讨论组成员结构以提高学习表现．Amelia 等人[108]将遗传算法和多示例学习方法应用在在线学习平台中，利用学习过程产生的日志数据，预测学习者能否完成课程．Nguyen 等人[75]提出使用推荐系统进行 EDM，并将其应用于预测学习者的学业表现．Rormero 等人[88]利用学习者在论坛中的讨论数据，研究学习者在论坛的表现和最终成绩之间的相关性，并基于多种数据挖掘算法和社交网络分析方法构建预测模型，在学习过程中预测学习者的最终成绩．Kotsiantis 等人[93]组合 WINNOW、朴素贝叶斯、支持向量机、1-NN 等多种机器学习算法构建集成分类器预测学习者在远程学习过程的学习表现．Laurie 等人[109]对学习者在异步讨论活动中的表现进行分析，使用数据挖掘方法研究学习进度与讨论的影响．Jung 等人[68]使用文本挖掘技术分析在线讨论区中的讨论内容，提出一种自动化生成学习者综合评价指标的计算方法，帮助学习者在学习过程中调整学习行为进而改善学习表现．Huseyin 等人[85]采用决策树分类器对学习者的个人信息与考试成绩数据进行挖掘，并开发了 MUSKUP（Mugla University Student Knowledge discovery Unit Program）软件研究影响学习者成绩的因素．

2.1.4 学习资源推荐

MOOC 平台中汇聚了大量学习资源，如何根据学习者的特点推荐适当的个性化学习资源是 MOOC 大数据需要解决的重要问题.

MOOC 相关研究

Diyi 等人[110]利用 MOOC 课程的论坛日志数据中的发帖内容、话题数量、话题相似性等要素构建推荐模型，实现根据历史论坛数据预测学习者兴趣话题并进行推荐. Dimitrios 等人[73]采用相关性分析与回归分析方法研究 MOOC 课程信息在社交媒体中的推荐过程，讨论不同社交网络用户对课程的偏好进而实现 MOOC 课程的推荐. Yunhong 等人[111]针对 MOOC 学习者提交的作业内容以及学习者的个人信息，采用文本挖掘与统计方法设计了一种结对评审者推荐系统实现为 MOOC 学习者推荐合适的评估作业. Naveen[112]基于模糊认知地图（Fuzzy Cognitive Map）理论设计了一个 MOOC 学习者知识能力模型，并利用此模型构建了适用于 MOOC 环境的自适应推荐系统，实现根据学习者的需求提供个性化的学习资源推荐.

传统 e-Learning 相关研究

Zailani 等人[71]根据学习者入学选课日志记录，采用关联规则挖掘和相关性分析方法分析所选课程是否实际符合学习者兴趣，进而为学习者提供选课的指导. Chun-Hsiung 等人[82]提出一种基于 Apriori 算法构建概念地图的方法，并开发了一个智能概念检测系统（Intelligent Concept Diagnostic System）协助指导者发现学习困难的学习者并为其提供补习路径（Remedial-Instruction Path）. Chun Fu Lin 等人[113]基于决策树方法开发了一个个性化创意学习系统来为学习者推荐个性化的学习路径. Aleksandra 等人[114]将学习者的测验成绩作为评价指标，基于学习风格索引（Index of Learning Styles）和协同过滤方法构建了一个自动适应学习者的兴趣以及知识水平的编程辅导系统推荐模型，实现为学习者推荐学习活动. Judy 等人[115]基于学习者的学习行为以及个人学习风格开发了一个在线自适应学习系统，实现为学习者个性化的安排学习过程和推荐学习资源.

2.1.5 学习异常分析

在教学活动中有多种因素会导致学习异常，而在 MOOC 这种大规模的学习环境中，学习困难、测验失败以及退课等异常情况显著增多[116]. 为了发现学习异常的原因进而帮助学习者顺利学习，就需要根据学习者模型、学习行为模型以及学习表现模型分析造成异常学习状态的原因.

MOOC 相关研究

Jeremy[117]采用统计方法对爱丁堡大学 MOOC 课程的论坛日志数据以及问卷调查数据进行分类分析，发现学习者对大规模的讨论内容存在困惑，包括：大量无意义的内容、话题过多、以及无法跟上讨论节奏，而当学习者的负面情绪达到一定程度时，学习者会退出课程与讨论. Derrick 等人[118]设计了一个面向 MOOC 论坛的衡量学习者发帖质量的声誉系统，避免无意义的发帖内容. Khe 等人[119]对 MOOC 相关的研究文章进行了统计，发现 MOOC 学习者学习失败的原因包括缺乏动力、缺少对讨论的关注、对话题的先验知识掌握不足、对课程内容理解不清、缺少帮助、缺少时间等因素.

传统 e-Learning 相关研究

Ali 等人[87]采用关联规则挖掘方法研究未能通过考试的学习者的学习特征，以找出造成学习者考试失败的因素. John 等人[79]基于逻辑回归方法和隐马尔科夫模型提出了一个用于指导中文拼音学习的统计模型，通过分析学习者的的拼音拼写过程，识别导致学习者学习拼音时的困难因素，发现容易引起拼写错误的拼音音节，进而调整学习者的学习过程，降低学

习困难．Yair[69]基于学术控制点理论（Academic Locus of Control）对 e-Learning 课程的学习日志进行分析，研究学习满意度和学习者放弃学习的关系．Carlos 等人[46]利用调研问卷采集学习者在的课程、教学、家庭、环境等方面的信息，并结合学习者的学业表现，采用数据挖掘方法对调研数据进行分类分析，研究影响学业失败、退学的因素．

2.2 指导者相关研究

MOOC 大数据不但为学习者提供支持，同时也为指导者构建更好的课程提供数据支持．通过利用上述的各种模型和分析方法，为课程设计、课件制作、课时安排、测验考试以及学习成绩评估等方面提供有价值的参考依据．

2.2.1 课程制作

课程的设计制作是 MOOC 教学活动开展的基础，需要综合考虑学习者认知能力、教学目标以及技术水平等多个方面因素，有针对性的设计合适的课程内容，包括录制教学视频、编排阅读资源、搭建虚拟实验、设置讨论话题、设计测验题目以及评估学习成绩等多个环节．

MOOC 相关研究

Philip[33][32]使用统计方法分析 MOOC 及其它在线课程的教育视频观看日志数据，发现随着视频长度的增加，学习者观看时长分布的众数先增加后减少并且均不超过 7 分钟，为指导者提供了编排教学视频的参考建议．Daniel 等人[120]对 MOOC 课程中的视频操作日志数据进行了统计分析，研究学习者的操作特点和观看习惯．Tania 等人[121]将地理信息系统与 MOOC 课程相关联并设计了三个统计分析工具对网页的访问质量进行评估，协助教师提高这种课程的设计质量．Anoush Margaryan 等人[122]采用统计方法分析两类 MOOC 课程的教学质量，提出了构建有效的 MOOC 教学过程的一些原则．Philip 等人[29]对 MOOC 课程中不同国家、不同年龄学习者的课程内容学习策略进行统计分析，帮助指导者调整教学内容的结构编排，改善在线教学手段的设计．Jon[123]对教育视频设计制作的方法与历史进行了总结，为教师提供了一些关于 MOOC 课程中所使用的视频内容的指导设计原则．

传统 e-Learning 相关研究

Jae 等人[124]基于文本挖掘技术对指定主题的文档集进行关键词抽取与词频分析，帮助指导者为 e-Learning 构建领域知识地图．Jingyun 等人[125]设计了一种课程中心的本体（Ontology）来辅助学习支持系统展现知识点和学习材料之间的关联．Le[126]等人提出一种基于云计算虚拟化的虚拟实验平台 V-Lab，实现软件安全、计算机网络安全以及信息保障等课程的上机实验要求，降低对物理机房的管理使用维护成本．Hsiu-Ping 等人[127]集成多种协作技术实现了一个在线沟通协作系统，在一个远程教学课程中为教师与学习者提供实时交互．Guangbing 等人[128]采用自动摘要技术生成学习内容的短文本概要协助学习者快速浏览，并将其用于移动学习的智能培训系统．Vikas 等人[129]对 e-Learning 中的媒体选择、学习领域、学习风格以及学习成效等因素进行统计分析，发现学习领域和学习风格会影响课程的媒体选择．Zailani 等人[130]提出临界相关支持（Critical Relative Support）测量方法用于挖掘教育数据的关联规则，为教师提供安排教学的参考信息．Enrique 等人[131]采用关联规则挖掘设计了一种的协作工具，帮助教师构建自适应的 web 课程以提高课程质量．Nawal 等人[132]使用 Web 挖掘方法对符合 SCORM（Sharable Content Object Reference Model）规范的学习资源和学习行为日志数据进行多层次的分析，协助指导者提高课程结构质量．

2.2.2 教师反馈与评估

传统在线课程的教学过程中，教师无法像传统课堂一样面对面获得学习者的反馈，进而

难以直接评估学习者的学习表现．而在 MOOC 中，凭借多种教学手段产生的日志数据，可以使指导者全面的了解学习者，但由于学习日志数据规模的庞大，需要利用 MOOC 大数据分析方法对日志进行分析．

MOOC 相关研究

Fang 等人[34]使用统计方法设计了一个分析框架，对学习者在 MOOC 课程测验中的表现进行统计分析，绘制成绩与相关因素的统计图表，协助教师了解学习者对知识点的掌握情况进而做出评估．Jan 等人[133]使用统计方法设计了一种自动评价作业的辅助系统，协助教师对 MOOC 课程中提交的大量作业进行自动化的批改，并对学习者的成绩进行重新计算．

传统 e-Learning 相关研究

Leah 等人[76]基于回归模型分析学习者在 LMS（Learning Management System）中的在线活动日志，发现影响学习者成绩的关键因素包括总讨论发帖数量，总邮件数量，总作业完成量等，并据此构建了“早期预警系统”用来帮助教师发现可能有学业问题的学习者．Fu-Ren 等人[84]基于文本挖掘、文本分类以及决策树等数据挖掘技术，提出一个类型划分系统（Genre Classification System）对论坛中学习者的讨论、问题、文章等数据进行自动化分类．Romero 等人[81]设计了一个集成在 e-Learning 系统中 Web 挖掘与推荐工具对学习者的学习数据进行在线分析与统计，协助指导者了解学习状态进而为学习者推荐合适的学习资源．Chih-Ming 等人[70]集成多种数据挖掘方法开发了一个综合评价工具用来协助指导者了解在线学习者的知识掌握情况以及分析影响学习者学习状态的主要因素．Gwo-Jen 等人[134]基于统计方法对学习者在搜索系统中的搜索行为进行分析，提出了一个 Web 搜索学习环境协助指导者分析学习者使用搜索引擎解决问题的学习行为和搜索策略，进而为学习者提供搜索建议．Divna 等人[83]采用 C4.5 分类算法，按照跨行业数据挖掘过程标准（Cross Industry Standard Process for Data Mining），利用 Moodle 中的学习者数据，实现对在线学习者的分组以帮助教师安排学习者进行小组协作．Andrea 等人[39]提出一种基于约束逻辑编程（Constraint Logic Programming，CLP）和结对评估的开放答案评估方法，并将其应用于 e-Learning 协作学习的环境，协助教师对学习者的作业进行评估．Bin-Shyan 等人[135]基于统计方法开发了一个日志检测系统对在线学习者的登录、阅读行为日志记录进行分析，协助教师评价学习者的学习态度．Gwo-Jen 等人[136]提出了一种群组决策方法用以分析概念影响关系（Concept-Effect Relationship）模型，并基于此方法实现了一个检测系统协助教师发现学习者的学习障碍并给予其学习建议．Tzu-Hua[64][56]基于统计学方法设计了一种以评估为中心的 e-Learning 学习系统，采用一系列的评估过程对学习者的学习行为进行分析以了解学习者的知识掌握水平，协助教师为学习者提供相关的补充学习材料提高学习者的学习水平．Juan 等人[91]构建了一个学习者参考模型系统，用于帮助教师了解学习者的到课情况并提供预测参考．Kenan 等人[67]综合利用统计方法和决策树、依赖网络和聚类等数据挖掘方法，设计了一个综合分析系统用以对学习系统数据库中的学习数据进行分析，进而为研究者提供研究学习规律的平台．

2.3 管理服务相关研究

MOOC 大数据除了用于学习者和指导者的数据分析支持以外，也需要面向教务管理人员等角色提供数据分析功能，其中对课程相关情况的各种指标统计与分析结果的数据可视化是主要需求．

2.3.1 课程数值统计

由于 MOOC 课程显著的规模性，使得课程各维度的数据量都产生显著增长，例如学习者的地理差异、年龄差异、学历差异等传统 e-Learning 中没有的因素都被显著放大，因而对 MOOC 课程各个方面的数值统计是 MOOC 大数据带来的一项新研究内容.

目前所有开设 MOOC 课程的大学都得到了学习者的学习日志数据，并且部分大学公开了对这些学习日志数据的数值统计. Lori 等人[25]对麻省理工学院开设的第一门 MOOC 课程“电路与电子学”进行了详细的分析. 首先从总体上统计了该课程的学习者总数以及按照年龄、地域、学习目标、学历背景等属性的分类数量，随后分别按照日期统计独立访问量、作业提交量、提问量、实验量、视频观看次数、资料阅读次数、书籍下载次数、讨论次数、Wiki 访问次数、学习者保留率. 然后统计完成课程的学习者在这些指标中的表现，并统计这些学习者对资源使用率的分布，作业、测验与考试期间的在线操作分布，参与讨论数量与人数的关系分布等. 爱丁堡大学[24]对其开设的 6 门 MOOC 课程进行了总体的数值统计，包括：6 门课程的总人数及按照年龄、教育背景、性别、工作领域、地域、学习目标的分类数量，随着教学周开展的学习人数变化，问卷调研反馈量及占总人数的比例，课程制作相关的工作人员、课时长度、视频数量/时长、教学手段等数据的数量，学习者选修课程数量分布，学习者活跃度随时间的变化，学习者发帖数量分布，学习者提交作业数量，学习者满意度分布，不同教学活动中学习者参与程度的分布，不同课程中学习者花费时间程度的分布等. 杜克大学[31]也对其开设的 MOOC 课程进行了数值统计分析，包括：课程制作相关的总工作时间、视频数量/时长、数据容量、文件数量、测验题数量等统计，周平均活跃学习者数量，课程总人数及其按照地域、学历、课程完成度、学习背景、年龄、学习动机的分类数量，论坛中自我介绍内容数量，每周的视频观看数量分布，完成不同阶段任务的学习者数量分布，学习者对不同学习动机认同程度的分布，通过认证与未通过认证的学习者在不同教学环节中参与程度的分布，学习者对学习收获的感想的分布等. 除了上述以论文形式公开的统计结果以外，哈佛大学[137]和麻省理工学院[138]采用 Web 技术将其所有 MOOC 课程的部分统计结果公开在了互联网上，提供了学习者总数量以及按照性别、教育背景、年龄等属性在地理区域上的分布情况.

2.3.2 数据可视化

与其他场景的大数据分析相似，MOOC 大数据的结果展示与数据分析也需要使用可视化技术. 不仅用于对统计数据的图形化显示，更重要的是用于对数据模式的挖掘和分析，为更深入的数据分析提供技术手段. 以下总结了目前在 MOOC 统计分析中所用到的可视化方法，如表 6 所示。除了常见的图表类型以外，在现有的 MOOC 大数据分析中也会同时使用多种图表类型综合展示数据. 如哈佛大学和麻省理工学院发布的 2013 年 MOOC 统计资料[139]中，组合使用散点图、柱状图以及折线图，展现课程学习量与获得成绩的散点图以及两个维度的数量分布图.

而在传统 e-Learning 相关研究中，基本的可视化方法即可满足较小数据规模的图形显示需求，因而相关研究相对较少. 例如 Jae 等人[124]利用折线图表现知识地图的提取率变化. Jingyun 等人[125]利用堆叠图表现学习者的学习风格分布. Yoav 等人[144]对 EDM 中聚类和序列模式挖掘的可视化进行了研究，利用状态空间图、状态序列位置图和柱状堆叠图等方式展示了序列模式挖掘和聚类等问题的可视化分析方法.

表 6 MOOC 大数据分析中的可视化方法

Table 6 Visualization Methods Applied in MOOC Big Data Analysis

图形	数据分析应用
折线图	学习者学习时间、退课率、活跃度等随课程进度的变化[25][139]；提交测验人数随课程进度的变化[3]；发帖数量随着课程进度的变化[35]；学习者观看视频时长随视频长度的变化[33][32]；
柱状图	学习者选课动机人数分布[25]；学习者成绩的人数分布[3]；学习者年龄、教育背景的人数分布[137][104][138]；提交作业量与学习者数量比例的分布[140]；各国学习者人数的排名[139]；
堆叠图	各课程注册学习者按行为、教育背景统计的数量分布[139][29]；课程问卷调研各问题回答情况分布[140]；学习者对各种教学手段功用评价的分布[98]；对比不同教学活动中学习者参与程度的分布[24]；学习者对不同学习动机认同程度的分布[31]；
地图	学习者注册量的地理分布[25][48]；活跃学习者的数量地理分布[141]；证书获得者、按性别、学历背景的地理分布[137][138]；
盒图	不同周工作量的学习者比例分布[3]；不同学历背景的学习者考试成绩分布[104]；不同课程的学习者年龄分布、注册时间分布[139]；不同长度视频的学习者观看时长分布[33]；
饼图	学习者课程满意度比例[24]；学习者的性别比例[137][104]；学习者的学历背景比例[142][48]；应用所学知识的方式比例[140]
散点图	测验提交数量与不同单元测试结果分数的关联分布[38]；课程完成率与开课时长、注册人数的关联分布[106]；测验题难度与教学效果的关联分布[143]；测验平均尝试次数与学习者数量、正确率之间的关联分布[34]
网络图	论坛内话题分组变化[72]；作业提交相似度关联关系[38]
热力图	测试提交次数与成绩关联的学习者数量分布[140]

3 MOOC 大数据处理框架

通过上述 MOOC 大数据的数据分析的总结，可以发现 MOOC 大数据的数据分析需求具有显著的多样性与复杂性，需要针对具体的 MOOC 平台和实际的研究需求设计相应的 MOOC 大数据处理框架. 为此本文依托西安交通大学网络教育学院（以下简称网络学院）的 BlueSky e-Learning 生态系统（以下简称 BlueSky）[145][146]，结合网络学院的实际需求展开 MOOC 大数据处理框架的设计与实践.

BlueSky 是一个基于云计算平台的具备知识地图导航学习、知识单元的短视频、论坛讨论组、在线测验和线下作业等多种教学手段的现代 e-Learning 生态系统，并且正在为全国各地超过 50,000 名学习者提供 13 个专业 300 余门课程的教学服务. 随着 Web 技术的发展与 MOOC 教学方式的推广，BlueSky 也在同步吸收先进的技术方法与教学理念，为学习者提供有价值的教学资源与创新的学习体验.

在技术层面上，MOOC 与 BlueSky 均基于 Web 相关技术实现的 e-Learning 信息管理系统，完整包含了基础架构层、内容层、应用层和终端层的各种组件；在教学层面上，BlueSky 实现了与 MOOC 相同的视频音频播放、在线讨论、文本阅读、作业评估和在线测验等 e-Learning 的教学功能；并且得益于网络学院开展学历教育的特点，BlueSky 还具有 MOOC 所不具备的上机实践、答疑、现场考试等线上线下结合的教学环节. 因此 BlueSky 能够产生与 MOOC 相似甚至更丰富的学习日志数据，为网络学院的习者、指导者、研究者和教务人员提供满足数据分析需求的基础.

因此我们认为 BlueSky 能够体现 MOOC 在技术教学和学习日志等方面的特点，为设计一个符合 MOOC 大数据特点的处理框架提供可靠的支持.

3.1 处理框架设计

综合文献总结与上述需求，基于 BlueSky 和业界大数据处理框架研究，本文提出一个用于管理 MOOC 大数据的处理框架，实现完整覆盖 MOOC 大数据的数据抽取、数据集成、数据分析和数据解释这四个主要处理流程，其基本结构如图 1 所示.

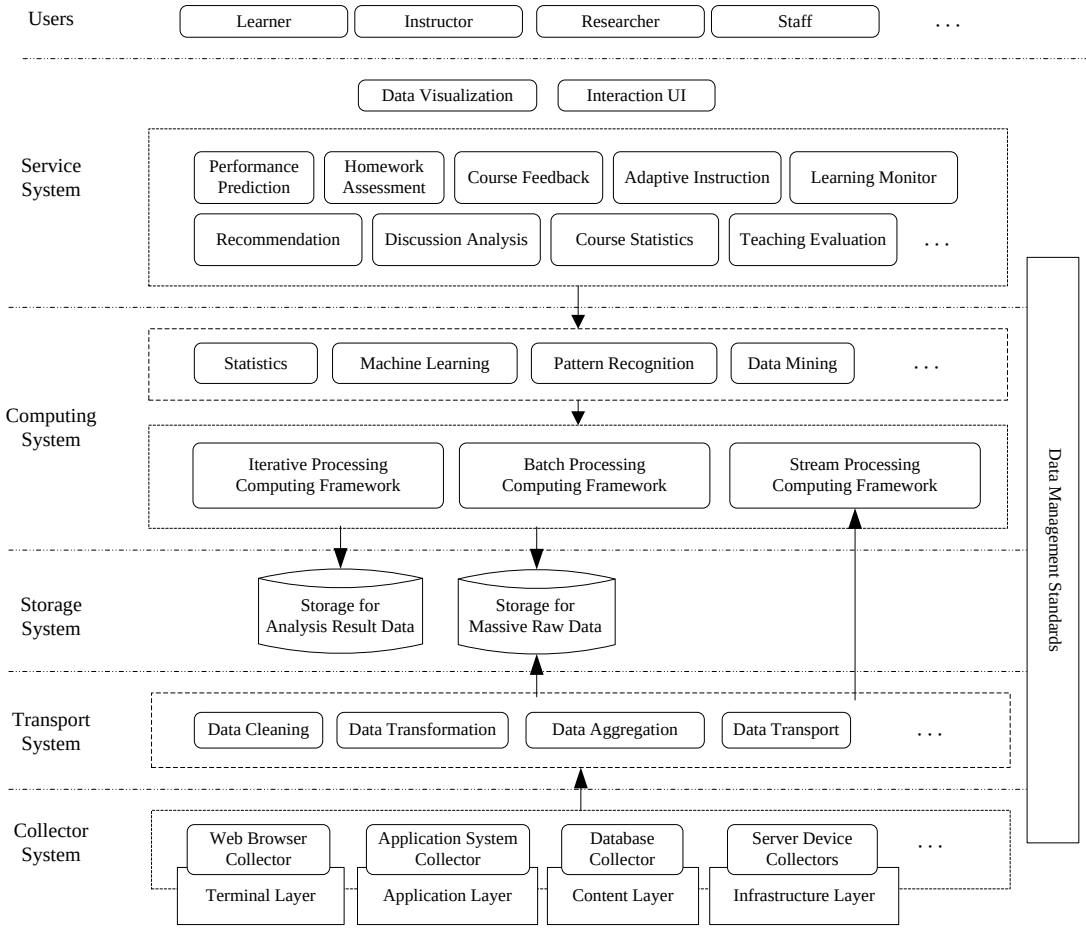


图 1MOOC 大数据处理框架

Fig. 1 Processing Framework of MOOC Big Data

MOOC 大数据处理框架由一套数据管理规范和五个系统组成，这五个系统分别是采集系统，传输系统，存储系统，计算系统和服务系统．通过这五个系统与数据管理规范的协同配合，可以实现对 MOOC 大数据全生命周期过程的支持．

1）数据管理规范是整个 MOOC 大数据处理框架的基础，为全部分析活动提供结构保障．由于 MOOC 大数据涉及到的问题域和数据类型都十分广泛，如果缺少对异构数据的统一规范，则难以整合其造成的数据孤岛，进而无法充分利用这些数据的价值．因而必须设计统一的数据管理规范，将每个数据源的内在属性及与其他数据源的关联关系统一的确立起来，为数据的采集、传输、存储和计算提供一致的数据解释，进而为各种服务奠定稳固的基础．在现有的 MOOC 数据分析相关研究中，麻省理工学院的 Kalyan 等人[26]对 MOOC 课程的学习过程设计了一种数据标准，将学习数据规范为三种模式，分别是观察模式，提交模式，协作模式，并提出基于这三种模式的数据统计分析方法，同时也兼顾考虑了数据采集和分析过程中可能暴露学习者隐私的问题．

2）采集系统由分布在 MOOC 系统各个层中的采集器组成，负责学习日志数据的自动化采集．根据 BlueSky 的设计经验，基于 Web 的 e-Learning 生态系统包含了应用层、内容层、设备层等多个层次，MOOC 作为 e-Learning 生态系统的一个典型实现，也是由这些层次组成的．学习者的学习过程即是由这些层次中的不同对象协作完成，通过在这些对象中采集学习者的使用轨迹，即可得到反映学习者学习过程的日志数据．这些日志数据不但包括了表 3 中所列举的教学手段直接产生的行为内容记录，也包括了相关服务系统与硬件间接产生的日志数据，这些日志共同构成了 MOOC 大数据的原始数据．由于教学手段与服务系统的多样性以及数据类型的不同，需要针对不同的采集对象根据数据管理规范设计采集方法，实现指

定目标的采集器．具体的日志数据采集来源、内容和类型如表 7 所示．

表 7 MOOC 学习日志数据采集来源、数据内容、数据类型

Tab. 7 The Sources, Contents and Types of Learning Log from MOOC Big Data

MOOC 层次	数据来源对象	数据内容	数据类型
终端层	个人计算机	浏览器内的鼠标、键盘操作过程日志	半结构化
	手持移动设备	触摸交互过程记录日志	半结构化
应用层	视频系统	观看数量、时长、播放器操作、视频信息等日志	半结构化
	测验系统	学习者答案、客观题目评分、尝试次数、时长、选择等操作的日志	半结构化
	虚拟实验系统	根据具体实验功能记录的学习者操作日志	半结构化
	作业系统	提交作业内容，互评分组及互评成绩等日志	非结构化，半结构化
	Wiki、阅读等系统	资源阅读时长、留言反馈等日志	半结构化
	论坛系统	发贴内容、话题线索组织、查看修改删除贴等日志	半结构化
内容层	数据库服务器	学习者与学习资源的元数据；学习过程引起的数据库查询请求日志	结构化，半结构化
	Web 服务器	Web 静态资源访问日志	半结构化
硬件层	硬件设备	电力、CPU、内存、磁盘和带宽的使用日志	半结构化

对于终端设备，需要根据终端类型按采集目标设计专用的采集器，例如在浏览器端，需要使用 JavaScript 语言设计采集学习者鼠标轨迹、键盘敲击等行为的采集器；对于服务器端系统，则根据数据的类型设计采集器，如对于数据库的结构化数据可以使用结构化查询语言检索数据库实现采集，对于各种应用系统日志文件的半结构化数据，则可以通过文件操作来采集，而对于用户提交的非结构化数据，则需要利用采集器提取内容转换为半结构化的数据以便利用；来自基础架构的数据，如虚拟化平台的数据则可以通过设备和底层系统的接口来采集．

3）传输系统负责将所采集到的数据传送到所需要的地方并进行必要的预处理操作．各种采集到的数据通常临时保存在其被采集的地方，这对充分利用数据价值显然无益：一方面此时各数据格式不同不利于分析，另一方面各数据并未建立关联进而无法挖掘内在的关联信息，而更重要的是如果在采集所在的业务设备上进行统计分析，显然会对业务系统造成额外的存储与计算的负载压力进而可能影响业务性能．所以为了充分利用所采集到的数据，需要构建传输系统，一方面对原始数据预处理并保存到存储系统的原始数据存储中，另一方面可以其传送到计算系统的流式计算框架中实现实时处理．

4）存储系统负责保存 MOOC 大数据的所有相关数据．MOOC 大数据处理框架中，不同的系统对于数据的使用方式是明显不同的，例如采集与传输系统只负责将海量的数据持续的写入而不需要读取，具体包括两种：一种是表 7 中所示的各个层次中的原始数据，作为将来进一步分析的数据基础，显然这种原始数据的规模庞大且种类较多；另一种是经过数据分析之后得到的分析结果数据，为上层的应用服务提供数据支持，这种数据的规模相对较小且有良好的结构．

由于这两种数据在规模和形式上并不相同，所以为了有效的利用其价值，存储系统至少应包含两种类型：一种用于存储 MOOC 大数据的原始数据，将其按照预先设定的存储规范有条理的持久化所有历史数据，为之后的分析提供保障．这样的存储系统必须具有良好的横向可扩展能力，并且支持数据密集型的计算操作以实现数据挖掘；另一种用于存储计算结果数据，将对原始数据统计分析和数据挖掘的计算结果持久化，为之后的数据服务提供快速的数据访问基础．因而这种存储系统需要支持快速的读写能力，同时也需要支持结构化的数据存储和数据检索．

5）计算系统负责实现 MOOC 大数据的各种数据分析需求的具体计算过程．根据之前对

MOOC 大数据的数据分析需求的总结，可以将分析过程归结为三种类型的计算：

第一种是对大量数据执行的批量计算，例如对课程的学习者特征或者学习过程的指标统计分析，需要完整计算所有日志数据进而得到统计值如总和、最值、均值、中位数、方差、偏度以及假设检验等统计指标；

第二种是对特定数据的实时指标跟踪和预测，例如对当前在线人数、课程访问量、视频点播量、讨论活跃度以及关键词趋势等指标的实时监测；

第三种是对问题数据的探索性分析，比如研究登录次数、发帖数量、观看视频时长、学习者群组划分等因素与学习者最终学习成绩之间的相关性分析、关联规则挖掘等问题，需要综合使用统计学、数据挖掘等领域的方法对数据集进行反复的深度研究.

所以，为了满足计算分析的需求，计算系统应当包含三种类型的计算框架：一种是数据密集型的批量计算框架，用于处理 MOOC 大数据的全量计算，从中统计有价值的信息；一种是流式计算框架，用于实时处理随时到来的大数据，实现快速计算跟踪指标变化；最后一种是迭代计算框架，提供对复杂的统计学方法、机器学习方法的支持，满足深度研究数据的需求．另外由于 MOOC 大数据的数据规模庞大，为了获得合理的分析效率，计算框架需要支持并行计算以提高计算效率.

6）服务系统负责为各种角色提供使用数据的接口．MOOC 中不同角色对数据分析的需求各不相同，例如学习者需要了解自己遇到的问题并获得指导，教师需要掌握课程的教学状态以便调整教学安排，系统运维人员则需要获得系统的当前情况与运行趋势．这些用户显然难以使用编程语言直接访问存储和计算系统，因而需要针对不同角色的需求设计服务接口．例如为学习者推荐学习资源参考资料和讨论话题，为教师提供可视化的课件使用统计情况与论坛讨论摘要，为系统运维人员提供服务器和应用系统的负载可视化等．总之各种角色的用户均可通过服务系统利用 MOOC 大数据，发挥 MOOC 大数据的应用价值.

实现上述 MOOC 大数据处理架构需要利用多种技术和工具，经过调研以及实践的应用，以下总结了多种免费开源软件可供选择，如表 8 所示.

表 8MOOC 大数据处理框架的开源软件

Table 8 Available Open Source Software for the Framework of MOOC Big Data

系统	开源软件
采集	Log4J, syslog
传输	ApacheFlume, Apache Kafka, Flunted, Scribe, LogStash, Rsyslog, rsync
存储	原始数据存储：Hadoop HDFS, HBase, Cassandra
	计算结果存储： MongoDB, MySQL, Redis, MariaDB, CouchDB
计算	批量计算框架：Hadoop MapReduce, Apache Mahout, Apache Hive, Apache Pig
	流式计算框架：Apache Spark Streaming, Storm, S4
	迭代计算框架：Apache Spark, R, Weka, Apache Drill, IPython Notebook
服务	Web 服务：Apache CXF, XML-RPC
	数据可视化： D3.js, ECharts, Highcharts, Matplotlib, Seaborn

3.2 典型应用

本文基于网络学院的 BlueSky，结合在研究者与管理人员在科研与系统运营等方面的数据分析需求，对 MOOC 大数据处理框架进行了原型验证，构建了一个涵盖学习日志数据全生命周期的学习日志分析平台（以下简称分析平台）.

分析平台利用分布在网络学院 BlueSky 中的采集器，实现了学习者在系统中学习过程的自动化记录．目前采集的学习日志包括学习者的学习操作日志（包括视频观看、作业提交、论坛讨论以及各种与系统的交互操作）、学习过程访问页面产生的 Web 服务器日志以及虚拟化平台的硬件日志．通过这些日志数据，可以完整的从学习者与系统两个方面刻画整个教学

过程.

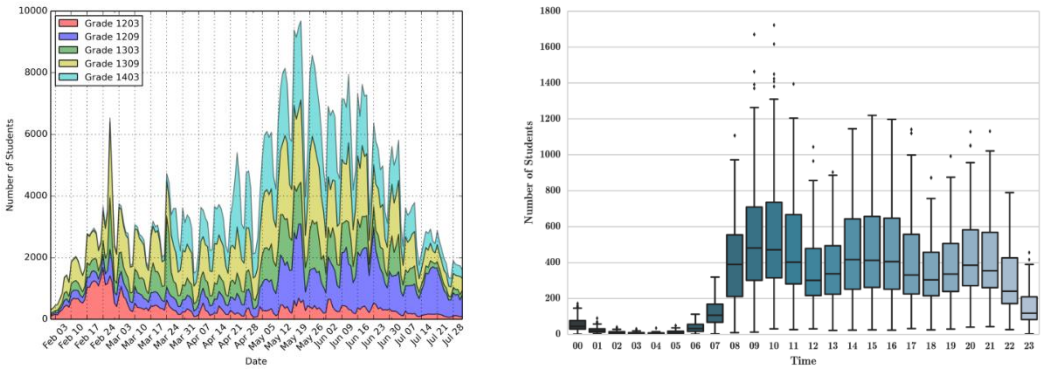
采集到日志数据之后，传输系统首先通过 HTTP 协议将日志数据按照事先制定的规范进行了数据格式规约，然后分别传输至两个不同的系统中．一个是存储系统中保存所有原始日志数据的 Hadoop HDFS；另一个计算系统中用于实时分析学习状态的 Storm 实时计算平台．对于存储在 Hadoop HDFS 的原始数据，可以采用 Hive 及 MapReduce 程序相结合的方式进行批量统计，获得对学习日志的统计结果并将这些结果保存在 MongoDB 数据库中便于教师和工作人员访问．另外，也可以使用 Spark 和 IPython notebook 相结合的方式对原始数据进行交互式的分析计算，使用统计学、机器学习等领域的方法分析日志数据，获得对学习者特性、教学过程评价等问题的更深入的分析结果.

以下将从分析平台中选取 2014 年 2 月 1 日至 8 月 1 日期间 5 个年级 46,841 名学习者共计 12,619,509 条的学习操作日志分别说明三个典型数据分析示例.

3.2.1 学习时间统计

研究者需要分析学习过程的时间统计特性，可以利用分析平台中计算系统的 Hive 计算框架对大规模的日志文件进行统计分析。

通常认为学习者在基于网络的学习过程中可能在任意时间进行学习，而实际上学习时间仍是有明显的分布特点的，如图 2(a)每日累计在线学习者人数统计显示，不同年级的学习者在访问时间上具有相似的特点．包括：工作日的学习者人数明显多于节假日，而随着考试临近，工作日与节假日的学习者人数都会迅猛增加，考试结束后则逐渐回落．图 2(b)一天各小时累计在线学习者人数统计显示，一天之中早上 7 点至晚上 23 点是学习者主要的访问时间，并且期间各时段的学习者人数分布区间范围较大，在线学习者的人数有三个明显的高峰分布，包括早上 8 点至 11 点，下午 2 点至 4 点，以及晚上 8 点至 9 点.



(a) The date distribution of learners demographics (b) The time distribution of learners demographics

图 2 基于时间区间的学习者数量分布

Fig. 2 The distribution of learners based on date and time

3.2.2 在线学习状态实时统计

为满足管理人员和教师了解当前的学习动态的需求，分析平台实现了在线学习状态实时统计的服务。利用计算系统中 Storm 流式计算框架对实时产生的学习日志进行处理，计算指定的指标观测值，并将处理结果保存在存储系统的 MongoDB 数据库中，服务系统中利用 ECharts 数据可视化工具绘制这些实时统计的处理结果的图形效果，展示给有数据需求的用户，并且可以根据用户的需求实时的与分析结果进行交互，展现合适的可视化效果.

如图 3 所示，服务系统中构建了一个基于 Web 的实时学习状态展示界面，以可交互的可视化形式提供了对当前的在线学习者数量、学习者年级与注册地分布，在线学习行为统计分布、课程学习人数分布、学习者地理分布等信息的实时状态.

3.2.3 教学分析

对于指导者，如何评价教学过程及其效果是网络学习过程中的重点与难点，需要综合利用多种数据分析方法与技术研究学习日志.

利用分析平台中 Spark 与 IPython Notebook 交互式计算框架，结合 Matplotlib 等工具提供的数据可视化界面，能够为相关研究提供支持. 如图 3 所示的线性代数课程学习相关性分析，通过对学习者视频观看行为数量与测验成绩进行分布统计与回归分析并结合可视化的图形展示，可以发现这两者之间的皮尔逊相关系数为-0.34，具有较弱的负相关性，表示该课程的视频内容可能过于困难致使部分学习者虽然多次观看视频但是仍未能掌握知识点.

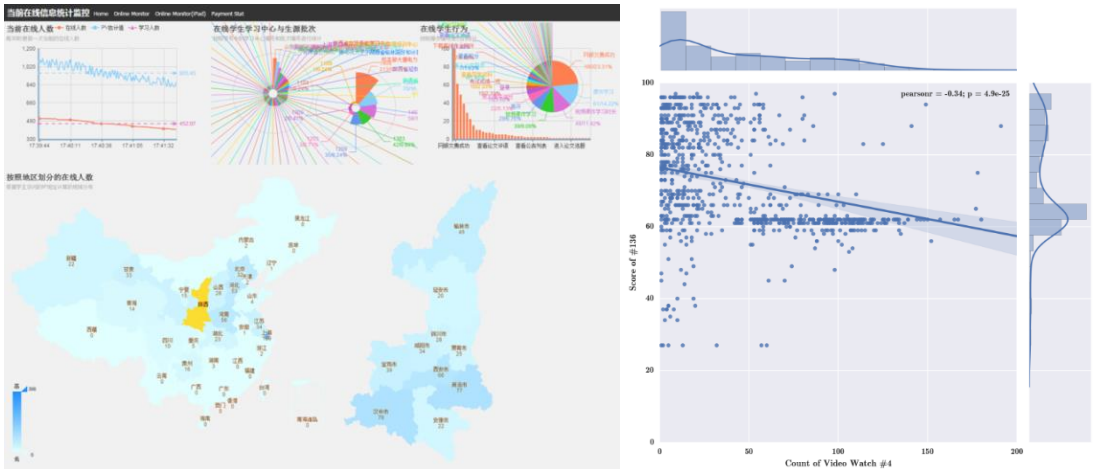


图 3 在线学习状态实时统计可视化界面与视频观看行为数量与测验成绩的分布

Fig. 3 The visualization of statistics of online learning status and the distribution of watch video and score

4 总结与挑战

本文主要总结了 MOOC 及其学习日志数据分析相关领域的研究现状，提出了使用 MOOC 大数据处理框架解决 MOOC 产生海量学习日志数据分析需求的设计思路，介绍了基于网络学院实际运行平台的学习日志数据分析平台的原型应用. 另外 MOOC 大数据也为相关的研究带来了新的挑战.

1）数据采集. MOOC 大数据采集的学习日志有两个方面，一是刻画学习者个人特性的详细记录，包括学习学习习惯、认知能力、情绪态度等；另一个是教学过程的关联信息，包括课程结构、内容表达、学习轨迹等. 对于前者，需要开发专用的采集工具获得学习者的准确信息，比如拍摄学习者的学习过程、记录键盘敲击序列、跟踪实验过程的眼动区域等明显涉及学习者个人隐私的数据. 而对于后者，则需要设计反映教学过程的规范将学习行为、学习资源、知识地图网络等多个层次的数据紧密关联. 因而如何设计合理有效的采集规范、方法与工具，使得既能获取更多有价值的数据为 MOOC 大数据提供分析基础，同时又能保护学习者隐私不被泄漏，是 MOOC 大数据的数据采集需要解决的重要问题.

2）数据分析与服务应用的鸿沟. MOOC 大数据能够从学习日志中挖掘有价值的信息，从理论上提供改善教学过程、提高学习效率、评价教学效果等多方面的参考. 如果进一步将这些研究成果转化为 MOOC 教学过程中的具体服务，则能够为学习者和指导者提供帮助，为 MOOC 发展产生正向推动力，并为 MOOC 大数据反馈更有价值的数据与问题，形成良好的发展循环. 而如何跨越从数据分析到服务应用之间的鸿沟使 MOOC 大数据落到实处则是 MOOC 大数据所面临的一个重要挑战.

[参考文献] (References)

[1] http://en.wikipedia.org/. Massive open online course [EB/OL].2014[2014-10-10]. http://en.wikipedia.org/wiki/Massive_open_online_course

[2] Carolyn King, Andrew Robinson, James Vickers. Online education: Targeted MOOC captivates students [J]. Nature, 2014, 505(7481): 26

[3] Laura Perna, Alan Ruby, Robert Boruch Nicole Wang, Janie Scull, Chad Evans, Seher Ahmad. The Life Cycle of a Million MOOC Users [EB/OL].2013[2014-10-10]. http://www.gse.upenn.edu/pdf/ahead/perna_ruby_boruch_moocs_dec2013.pdf

[4] Coursera.org. Coursera.org Community [EB/OL]. 2014[2014-10-10]. https://www.coursera.org/about/community

[5] Pappano, Laura. The Year of the MOOC [N]. The New York Times, 2012

[6] Science Editors. Grand Challenges in Science Education [J]. Science, 2013, 340(1): 291

[7] Marcia McNutt. Bricks and MOOCs [J]. Science, 2013, 342(6157): 402

[8] Nature Editors. Learning in a Digital Age [EB/OL].2013[2014-10-10]. http://www.nature.com/nature/focus/digitallearning/index.html

[9] Seth Fletcher. How Big Data Is Taking Teachers Out of the Lecturing Business [EB/OL]. 2013[2014-10-10]. http://www.scientificamerican.com/article/how-big-data-taking-teachers-out-lecturing-business/

[10] M. Mitchell Waldrop. Online learning: Campus 2.0 [J]. Nature, 2013, 495(7442): 160-163

[11] Li Yuan, Stephen Powell. MOOCs and Open Education: Implications for Higher Education - A white paper [C] //http://publications.cetis.ac.uk/2013/667. Bolton, England: Ce, 2013

[12] I. Elanine Allen, Jeff Seaman. Grade Change: Tracking Online Education in the United States, 2013 [EB/OL]. 2014[2014-10-10]. http://onlinelearningconsortium.org/publications/survey/grade-change-2013

[13] I. Elanine Allen, Jeff Seaman. Changing Course: Ten Years of Tracking Online Education in the United States, 2012 [EB/OL]. 2014[2014-10-10]. http://onlinelearningconsortium.org/publications/survey/changing_course_2012

[14] Marie Bienkowski, Mingyu Feng, Barbara Means. Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief [EB/OL]. 2012[2014-10-10]. http://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

[15] U.S. Department of Education Office of Educational Technology. Expanding Evidence Approaches for Measuring Learning in a Digital World [EB/OL].2013[2014-10-10]. http://tech.ed.gov/files/2013/02/Expanding-Evidence-Approaches.pdf

[16] White House. ConnectED Initiative [EB/OL].2013[2014-10-10]. http://www.whitehouse.gov/issues/education/k-12/connected

[17] Coursera. Courses of Coursera [EB/OL]. [2014-10-10]. https://www.coursera.org/courses

[18] edX. Courses of edX [EB/OL]. [2014-10-10]. https://www.edx.org/course-list

[19] http://www.xuetangx.com/. Partners of xuetangx.com [EB/OL]. [2014-10-10]. http://www.xuetangx.com/partners

[20] http://www.icourse163.org/. Partners of icourse163.org [EB/OL]. [2014-10-10]. http://www.icourse163.org/university/view/all.htm#/

[21] Karen Goldschmidt, Jane Greene-Ryan. Massive Open Online Courses in Nursing Education [J]. Journal of Pediatric Nursing, 2014, 29(2): 184-186

[22] Christina M. Stark, Jamie Pope. Massive Open Online Courses: How Registered Dietitians Use MOOCs for Nutrition Education [J]. Journal of the Academy of Nutrition and Dietetics, 2014, 114(8): 1147-1151, 1153, 115

[23] Wilkowski, Julia, Amit Deutsch, et al. Student skill and goal achi-evement in the mapping with google MOOC [C] //Proceedings of the first ACM conference on Learning@ scale conference. Atlanta:ACM, 2014

[24] University of Edinburgh. MOOCs @ Edinburgh 2013 - Report #1 [EB/OL].2013[2014-10-10]. https://www.era.lib.ed.ac.uk/bitstream/1842/6683/1/Edinburgh_MOOCs_Report2013_no1.pdf

[25] Lori Breslow, David E. Pritchard, Jennifer DeBoer, Glenda S. Stump, Andrew D. Ho, Daniel T. Seaton, . Studying Learning in the Worldwide Classroom Research into edX's First MOOC [EB/OL]. 2013[2014-10-10]. http://www.rpajournal.com/dev/wp-content/uploads/2013/05/SF2.pdf

[26] Kalyan Veeramachaneni, Franck Dernoncourt, Colin Taylor, et al. MOOCdb: Developing Data Standards for MOOC Data Science [C] //Proceedings of the 1st Workshop on Massive Open Online Courses at the 16th Annual Conference on Artificial Intelligence in Education. Memphis:TN, 2013

[27] Chris Parr. Mooc completion rates 'below 7%' [EB/OL].2013[2014-10-10]. http://www.timeshighereducation.co.uk/news/mooc-completion-rates-below-7/2003710.article

[28] http://harvardx.harvard.edu/. World Map of Enrollment [EB/OL].2014[2014-10-10]. http://harvardx.harvard.edu/harvardx-insights/world-map-enrollment

[29] Philip J. Guo, Katharina Reinecke. Demographic differences in how students navigate through MOOCs [C] //Proceeding L@S '14 Proceedings of the first ACM conference on Learning. New York:ACM, 2014

[30] MITx. All MITx Offerings Estimated Registration by Country [EB/OL]. 2014[2014-10-10]. http://odl.mit.edu/mitx-insights/world-map-enrollment/

[31] Duke University. Bioelectricity: A Quantitative Approach [EB/OL]. 2013[2014-10-10]. http://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/6216/Duke_Bioelectricity_MOOC_Fall2012.pdf

[32] Philip Guo. Optimal Video Length for Student Engagement [EB/OL].2014[2014-10-10].

https://www.edx.org/blog/optimal-video-length-student-engagement
[33] Philip J. Guo, Juho Kim, Rob Rubin. How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos [C] //Proceeding L@S '14 Proceedings of the first ACM conference on Learning. New York: ACM, 2014
[34] Fang He, Kalyan Veeramachaneni, Una-May O'Reilly. Analyzing Millions of Submissions to Help MOOC Instructors Understand Problem Solving [EB/OL]. 2014[2014-10-10]. http://groups.csail.mit.edu/EVO-DesignOpt/groupWebSite/uploads/Site/ProblemAnalytics.pdf
[35] Miaomiao Wen, Diyi Yang,Carolyn Penstein Rosé. Sentiment Analysis in MOOC Discussion Forums: What does it tell us? [C] //Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[36] Revathi Viswanathan. Teaching and Learning through MOOC [EB/OL]. [2014-10-10]. http://www.academia.edu/download/30401977/4-_Viswanathan__2012.pdf
[37] M. Mitchell Waldrop. Education online: The virtual lab [J]. Nature, 2013, 499(7458): 268-270
[38] Jonathan Huang, Chris Piech, Andy Nguyen, et al. Syntactic and Functional Variability of a Million Code Submissions in a Machine Learning MOOC [C] //AIED 2013 Workshops Proceedings. Berlin: Springer, 2013
[39] Andrea Sterbini ,Marco Temperini . Analysis of open answers via mediated peer-assessment [C] //System Theory, Control and Computing (ICSTCC), 2013 17th International Conference. Sinaia:IEEE, 2013
[40] Daniel T. Seaton, Yoav Bergner, Isaac Chuang, Piotr Mitros, David E. Pritchard . Who Does What in a Massive Open Online Course? [J]. Communications of the ACM, 2013, 57(4): 58-65
[41] Grobelnik M. Big-data computing: Creating revolutionary breakthroughs in commerce, science, and society [EB/OL]. [2014-10-10]. http://videolectures.net/eswc2012_grobelnik_big_data/r
[42] Barwick H. The "four Vs" of Big Data. Implementing Infomation Infrastructrue Symposium [EB/OL]. [2014-10-10]. http://www.computerworld.com.au/article/396198/iiis_four_vs_big_data/
[43] IBM. What is big data? [EB/OL]. [2014-10-10]. https://www-01.ibm.com/software/data/bigdata/
[44] Lior Abraham, John Allen, Oleksandr Barykin, et al. Scuba: Diving into Data at Facebook [C] //Proceedings of the VLDB Endowment. New York: ACM, 2013
[45] Hale Ilgaz , Arif Altun , Petek Aşkar. The effect of sustained attention level and contextual cueing on implicit memory performance for e-learning environments [J]. Computers in Human Behavior, 2014, 39(1): 1-7
[46] Carlos Márquez-Vera, Cristóbal Romero Morales, Sebastián Ventura Soto. Predicting School Failure and Dropout by Using Data Mining Techniques [J]. IEEE JOURNAL OF LATIN-AMERICAN LEARNING TECHNOLOGIES, 2013, 8(1): 7-14
[47] Yan Li, Maria Ranieri. Educational and social correlates of the digital divide for rural and urban children: A study on primary school students in a provincial city of China [J]. Computers & Education, 2013, 60(1): 197-209
[48] Jennifer DeBoer, Glenda S. Stump, Daniel Seaton, Lori Breslow. Diversity in MOOC Students' Backgrounds and Behaviors in Relationship to Performance in 6.002x [EB/OL].2013[2014-10-10]. http://tll.mit.edu/sites/default/files/library/LINC%20'13.pdf
[49] Meng Xiaofeng, Ci Xiang. Big Data Management: Concepts, Techniques and Challenges [J]. Journal of Computer Research and Development, 2013, 50(1): 146-169
[50] http://en.wikipedia.org/. Educational data mining [EB/OL].2014[2014-10-10]. http://en.wikipedia.org/wiki/Educational_data_mining
[51] wikipedia. Learning analytics [EB/OL].2014[2014-10-10]. http://en.wikipedia.org/wiki/Learning_analytics
[52] George Siemens, Ryan S J.d. Baker . Learning analytics and educational data mining: towards communication and collaboration [C] //Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. New York: ACM, 2012
[53] Siti Khadijah Mohamada, Zaidatun Tasir. Educational data mining: A review [C] //The 9th International Conference on Cognitive Science. Beijing,China:Elsevi, 2013
[54] Cristobal Romero, Sebastian Ventura. Educational Data Mining: A Review of the State of the Art [J]. IEEE Transactions on Systems, Man, and Cybernetics - part C: Applications and Reviews, 2010, 40(6): 601-618
[55] Alejandro Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works [J]. Expert Systems with Applications, 2014, 41(4 part 1): 1432-1462
[56] Tzu-Hua Wang. Web-based dynamic assessment: Taking assessment as teaching and learning strategy for improving students' e-Learning effectiveness [J]. Computers & Education, 2010, 54(4): 1157-1166
[57] Francisco J. Veredas , Esperanza Ruiz-Bandera , Francisca Villa-Estrada, Juan F. Rufino-González, Laura Morente. A web-based e-learning application for wound diagnosis and treatment [J]. Computer Methods and Programs in Biomedicine, 2014, 116(3): 236-248
[58] Brent Morgan, William Baggett ,Vasile Rus. Error Analysis as a Validation of Learning Progressions [C] //Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[59] Matthieu Cisel, Rémi Bachelet , Eric Bruillard . Peer assessment in the first French MOOC : Analyzing assessors' behavior [C] //Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[60] Chien-Sing Lee. Diagnostic, predictive and compositional modeling with data mining in integrated learning environments [J]. Computers & Education, 2007, 49(3): 562-580
[61] Nian-Shing Chen, Kinshuk, Chun-Wang Wei, Hong-Jhe Chen. Mining e-Learning domain concept map from academic articles [J]. Computers & Education, 2008, 50(3): 1009-1021
[62] Dominique Verpoorten, Wim Westera, Marcus Specht. Using reflection triggers while learning in an online course [J]. British Journal of Educational Technology, 2012, 43(6): 1030-1040
[63] Bin-Shyan Jong, Chien-Ming Chen, Te-Yi Chan, Tsong-Wuu Lin, Yen-Teh Hsia. Effect of knowledge

complementation grouping strategy for cooperative learning on online performance and learning achievement [J]. Computer Applications in Engineering Education, 2014, 22(3): 541-550
[64] Tzu-Hua Wang. Developing an assessment-centered e-Learning system for improving student learning effectiveness [J]. Computers & Education, 2014, 73(1): 189-203
[65] Sharona T. Levy , Uri Wilensky. Mining students' inquiry actions for understanding of complex systems [J]. Computers & Education, 2011, 56(3): 556-573
[66] Erica L. Snow, Mathew E. Jacovina, Laura K. Allen, et al. Entropy: A Stealth Measure of Agency in Learning Environments [C] // Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[67] Kenan Zengin, Necmi Esgi, Ergin Erginer, et al. A sample study on applying data mining research techniques in educational science: Developing a more meaning of data [C] //3rd World Conference on Educational Sciences - 2011. Amsterdam: Elsevier, 2011
[68] Jung-Lung Hsu , Huey-Wen Chou , Hsiu-Hua Chang. EduMiner: Using text mining for automatic formative assessment [J]. Expert Systems with Applications, 2011, 38(4): 3431-3439
[69] Yair Levy. Comparing dropouts and persistence in e-learning courses [J]. Computers & Education, 2007, 48(2): 185-204
[70] Chih-Ming Chen , Ming-Chuan Chen. Mobile formative assessment tool based on data mining techniques for supporting web-based learning [J]. Computers & Education, 2009, 52(1): 256-273
[71] Zailani Abdullah, Tutut Herawan, Noraziah Ahmad, et al. Extracting highly positive association rules from students' enrollment data [C] //Social and Behavioral Sciences WCETR 2011 . Amsterdam: Elsevier, 2011
[72] Nabeel Gillani, Rebecca Eynon. Communication patterns in massively open online courses [J]. The Internet and Higher Education, 2014, 23(1): 18-26
[73] Dimitrios Kravvaris, Georgios Ntanis,Katia L. Kermanidis. Studying massive open online courses: recommendation in social media [C] //Proceedings of the 17th Panhellenic Conference on Informatics. New York: ACM, 2013
[74] Gayle Christensen, Andrew Steinmetz, Brandon Alcorn, Amy Bennett, Deirdre Woods,Ezekiel J Emanuel. The MOOC Phenomenon: Who Takes Massive Open Online Courses and Why? [EB/OL].2013[2014-10-10]. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2350964&download=yes
[75] Nguyen Thai-Nghe, Lucas Drumond, Artus Krohn-Grimberghe, Lars Schmidt-Thieme. Recommender system for predicting student performance [J]. Procedia Computer Science, 2010, 1(2): 2811-2819
[76] Leah P. Macfadyen, Shane Dawson. Mining LMS data to develop an "early warning system" for educators: A proof of concept [J]. Computers & Education, 2010, 54(2): 588-599
[77] Michael V. Yudelson, Stephen E. Fancsali, Steven Ritter, et al. Better Data Beat Big Data [C] //Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[78] Suhang Jiang, Adrienne E.Williams, Katerina Schenke, et al. Predicting MOOC Performance with Week 1 Behavior [C] //Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[79] John Kowalski, Yanhui Zhang, Geoffrey J. Gordon. Statistical Modeling of Student Performance to Improve Chinese Dictation Skills with an Intelligent Tutor [J]. Journal of Educational Data Mining, 2014, 6(1): 3-27
[80] Nilesh N Wani, Shrikant Lade. Comparative Analysis on the Application of Data Mining in the Field of Teaching Evaluation [EB/OL].2013[2014-10-10].
[81] Cristóbal Romero, Sebastián Ventura, Amelia Zafra, Paul de Bra. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems [J]. Computers & Education, 2009, 53(3): 828-840
[82] Chun-Hsiung Lee, Gwo-Guang Lee, Yungho Leu. Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning [J]. Expert Systems with Applications, 2009, 36(2 part 1): 1675-1684
[83] Divna Krpan, Slavomir Stankov. Educational Data Mining for Grouping Students in E-learning System [C] //Proceedings of the ITI 2012 34th Int. Conf. on Information Technology Interfaces. Zagreb: Srce-Unive, 2012
[84] Fu-Ren Lin, Lu-Shih Hsieh, Fu-Tai Chuang. Discovering genres of online discussion threads via text mining [J]. Computers & Education, 2009, 52(2): 481-495
[85] Huseyin Guruler, Ayhan Istanbullu, Mehmet Karahasan. A new student performance analysing system using knowledge discovery in higher educational databases [J]. Computers & Education, 2010, 55(1): 247-254
[86] Cristóbal Romero, Sebastián Ventura, Enrique García. Data mining in course management systems: Moodle case study and tutorial [J]. Computers & Education, 2008, 51(1): 368-384
[87] Ali Buldu, Kerem Üçgün. Data mining application on students' data [J]. Procedia - Social and Behavioral Sciences, 2010, 2(2): 5251-5259
[88] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, Sebastián Ventura. Predicting students' final performance from participation in on-line discussion forums [J]. Computers & Education, 2013, 68(1): 458-472
[89] Ezequiel Scott, Guillermo Rodríguez, Álvaro Soria, Marcelo Campo. Are learning styles useful indicators to discover how students use Scrum for the first time? [J]. Computers in Human Behavior, 2014, 36(1): 56-64
[90] Juan Yang, Zhi Xing Huang, Yue Xiang Gao, Hong Tao Liu. Dynamic Learning Style Prediction Method Based on a Pattern Recognition Technique [J]. IEEE Transactions on Learning Technologies, 2014, 7(2): 165-177
[91] Juan A. Lara, David Lizcano, María A. Martínez, Juan Pazos, Teresa Riera. A system for knowledge discovery in e-learning environments within the European Higher Education Area - Application to student data from Open University of Madrid, UDIMA [J]. Computers & Education, 2014, 72(1): 23-26
[92] Richard L. Lamb, David B. Vallett, Tariq Akmal, Kathryn Baldwin. A computational modeling of student cognitive processes in science education [J]. Computers & Education, 2014, 79(1): 116-125
[93] S. Kotsiantis, K. Patriarcheas, M. Xenos. A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education [J]. Knowledge-Based Systems, 2010, 23(6): 529-535

[94] C.J. Carmona, P. González, M.J. del Jesus, et al. Evolutionary algorithms for subgroup discovery applied to e-learning data [C] //Education Engineering (EDUCON), 2010 IEEE. Madrid :IEEE, 2010
[95] Kimberly F. Colvin, John Champaign, Alwina Liu, et al. Comparing Learning in a MOOC and a Blended On-Campus Course [C] //Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[96] Michael Yudelson, Roya Hosseini, Arto Vihavainen, et al. Investigating Automated Student Modeling in a Java MOOC [C] //Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[97] Dagmar El-Hmoudova. MOOCs Motivation and Communication in the Cyber Learning Environment [J]. Procedia - Social and Behavioral Sciences, 2014, 131(15): 29-34
[98] Franka Gr¨unewald, Christoph Meinel, Michael Totschnig, et al. Designing MOOCs for the Support of Multiple Learning Styles [C]//Scaling up Learning for Sustained Impact. Berlin: Springer, 2013
[99] Uroš Ocepek, Zoran Bosnić, Irena Nančovska Šerbec, Jože Rugelj. Exploring the relation between learning style models and preferred multimedia types [J]. Computers & Education, 2013, 69(1): 343-355
[100] Luke K. Fryer, H. Nicholas Bovee, Kaori Nakao. E-learning: Reasons students in language learning courses don't want to [J]. Computers & Education, 2014, 74(1): 26-36
[101] Luuk Van Waes, Daphne van Weijen, Mariëlle Leijten. Learning to write in an online writing center: The effect of learning styles on the writing process [J]. Computers & Education, 2014, 73(1): 60-71
[102] Neha Gulati. Framework for cognitive agent based expert system for metacognitive and collaborative E-Learning [C] //2013 IEEE International Conference in MOOC Innovation and Technology in Education (MITE). Jaipur: IEEE, 2013
[103] Sanjeeva Srivastava. Online education: E-learning booster in developing world [J]. Nature, 2013, 501(216): 316-316
[104] Saif Rayyan, Daniel T. Seaton, John Belcher, David E. Pritchard, Isaac Chuang. Participation And performance In 8.02x Electricity And Magnetism: The First Physics MOOC From MITx [EB/OL]. 2014[2014-10-10]. http://arxiv.org/pdf/1310.3173.pdf
[105] Ignacio Martinez. The Effects of Nudges on Students' Effort and Performance: Lessons from a MOOC [EB/OL]. 2014[2014-10-10]. http://curry.virginia.edu/uploads/resourceLibrary/19_Martinez_Lessons_from_a_MOOC.pdf
[106] Nikola Stevanović. Effects of Motivation on Performance of Students in MOOC [EB/OL].2014[2014-10-10]. http://portal.sinteza.singidunum.ac.rs/paper/137
[107] Chih-Kai Changa, Gwo-Dong Chenb, Chin-Yeh Wang. Statistical model for predicting roles and effects in learning community [J]. Behaviour & Information Technology, 2011, 30(1): 101-111
[108] Amelia Zafra, Crist´obal Romero, Sebasti´an Ventura. Predicting Academic Achievement Using Multiple Instance Genetic Programming [C] //International Conference on Intelligent Systems Design and Applications. Pisa:IEEE, 2009
[109] Laurie P. Dringus, Timothy Ellis. Using data mining as a strategy for assessing asynchronous discussion forums [J]. Computers & Education, 2005, 45(1): 141-160
[110] Diyi Yang, Mario Piergallini, Iris Howley, et al. Forum Thread Recommendation for Massive Open Online Courses [C] //Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[111] Yunhong Xu, Ru Wang. Peer Reviwer Recommendation in Online Social Learning Context: Integration Information of Learners and Submissions [EB/OL]. 2014[2014-10-10]. http://aisel.aisnet.org/pacis2014/295
[112] Naveen Bansal. Adaptive Recommendation System for MOOC [EB/OL].2014[2014-10-10]. http://www.it.iitb.ac.in/frg/wiki/images/7/74/11305R010-Naveen_Bansal-Stage1Report.pdf
[113] Chun Fu Lin, Yu-chu Yeh, Yu Hsin Hung, Ray I Chang. Data mining for providing a personalized learning path in creativity: An application of decision trees [J]. Computers & Education, 2013, 68(1): 199-210
[114] Aleksandra Klašnja-Milićevića, Boban Vesina, Mirjana Ivanovićb, Zoran Budimac. E-Learning personalization based on hybrid recommendation strategy and learning style identification [J]. Computers & Education, 2011, 56(3): 885-899
[115] Judy C.R. Tsenga, Hui-Chun Chub, Gwo-Jen Hwangb, Chin-Chung Tsai. Development of an adaptive learning system with two sources of personalization information [J]. Computers & Education, 2008, 51(2): 776-786
[116] Thanasis Daradoumis, Roxana Bassi, Fatos Xhafa, et al.A review on massive e-learning (MOOC) design, delivery and assessment [C] //2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. Compiegne: IEEE, 2013
[117] Jeremy Knox. Digital culture clash: "massive" education in the E-learning and Digital Cultures MOOC [J]. Distance Education, 2014, 35(2): 164-177
[118] Derrick Coetzee, Armando Fox, Marti A. Hearst, et al. Should Your MOOC Forum Use a Reputation System? [C] //Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. New York: ACM, 2014
[119] Khe Foon Hew, Wing Sum Cheung. Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges [J]. Educational Research Review, 2014, 12(1): 45-58
[120] Daniel T. Seaton, Albert J. Rodenius, Cody A. Coleman, et al. Analysis of video use in edX courses [C] //AIED 2013 Workshops Proceedings. Berlin: Springer, 2013
[121] Tania Calle-Jimenez, Sandra Sanchez-Gordon ,Sergio Luján-Mora . Web accessibility evaluation of massive open online courses on Geographical Information Systems [C] //IEEE Global Engineering Education Conference (EDUCON). Istanbul: IEEE, 2014
[122] Anoush Margaryan, Manuela Bianco, Allison Littlejohn. Instructional quality of Massive Open Online Courses (MOOCs) [J]. Computers & Education, 2015, 80(1): 77-83

[123] Jon Baggaley. The Sudden Revival of Educational Video [C] // IEEE 63rd Annual Conference International Council for Educational Media (ICEM) . Singapore: IEEE, 2013
[124] Jae Hwa Lee, Aviv Segev. Knowledge maps for e-learning [J]. Computers & Education, 2012, 59(2): 353-364
[125] Jingyun Wang, Takahiko Mendori, Juan Xiong. A language learning support system using course-centered ontology and its evaluation [J]. Computers & Education, 2014, 78(1): 278-293
[126] Le Xu, Dijiang Huang, Wei-Tek Tsai. Cloud-Based Virtual Laboratory for Network Security Education [J]. IEEE Transactions on Education, 2014, 57(3): 145-150
[127] Hsiu-Ping Yueh, Weijane Lin, Yi-Lin Liu, Tetsuo Shoji, Michihiko Minoh. The Development of an Interaction Support System for International Distance Education [J]. IEEE Transactions on Learning Technologies, 2014, 7(2): 191-196
[128] Guangbing Yang, Nian-Shing Chen, Kinshukc, Erkki Sutinen, Terry Andersond, Dunwei Wen. The effectiveness of automatic text summarization in mobile learning contexts [J]. Computers & Education, 2013, 68(1): 233-243
[129] Vikas Sahasrabudhe, Shivraj Kanungo. Appropriate media choice for e-learning effectiveness: Role of learning domain and learning style [J]. Computers & Education, 2014, 76(1): 237-249
[130] Zailani Abdullah, Tutut Herawan, Noraziah Ahmad, et al. Mining significant association rules from educational data using critical relative support approach [C] //World Conference on Educational Technology Researches. Amsterdam: Elsevier, 2011
[131] Enrique García, Cristóbal Romero, Sebastián Ventura, Carlos de Castro. A collaborative educational association rule mining tool [J]. The Internet and Higher Education, 2011, 14(2): 77-88
[132] Nawal Sael, Abdelaziz Marzak, Hicham Behja. Web Usage Mining data preprocessing and multi level analysis on Moodle [C] //AICCSA. Ifrane: IEEE, 2013
[133] Jan Renz, Thomas Staubitz, Christian Willems, et al. Handling re-grading of automatically graded assignments in MOOCs [C] //IEEE Global Engineering Education Conference (EDUCON). Istanbul: IEEE, 2014
[134] Gwo-Jen Hwanga, Pei-Shan Tsaib, Chin-Chung Tsaic, Judy C.R. Tseng. A novel approach for assisting teachers in analyzing student web-searching behaviors [J]. Computers & Education, 2008, 51(2): 926-938
[135] Bin-Shyan Jong, Te-Yi Chan, Yu-Lung Wu. Learning Log Explorer in E-Learning Diagnosis [J]. IEEE Transactions on Education, 2007, 50(3): 216-228
[136] Gwo-Jen Hwang, Patcharin Panjaburee, Wannapong Triampo, Bo-Ying Shih. A group decision approach to developing concept-effect models for diagnosing student learning problems in mathematics [J]. British Journal of Educational Technology, 2013, 44(3): 453-468
[137] HarvardX. HarvardX Insights [EB/OL]. 2014 [2014-10-10]. http://harvardx.harvard.edu/harvardx-insights
[138] MITx. MITx Insights [EB/OL].2014[2014-10-10]. http://odl.mit.edu/insights/
[139] HarvardX, MITx. HarvardX and MITx: The First Year of Open Online Courses Fall 2012-Summer 2013 [EB/OL]. 2014[2014-10-10]. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263
[140] Heather Miller, Philipp Haller, Lukas Rytz, et al. Functional Programming For All! Scaling a MOOC for Students and Professionals Alike [C] //Proceedings of ICSE' 14. New York: ACM, 2014
[141] Rene F. Kizilcec. Collaborative Learning in Geographically Distributed and In-person Groups [C] //AIED 2013 Workshops Proceedings. Berlin: Springer, 2013
[142] Tawanna Dillahunt, Bingxin Chen, Stephanie Teasley. Model Thinking: Demographics and Performance of MOOC Students Unable to Afford a Formal Education [C] //Proceedings of the first ACM conference on Learning@ scale conference. New York: ACM, 2014
[143] Arto Vihavainen, Matti Luukkainen and Jaakko Kurhila. Multi-faceted Support for MOOC in Programming [C] //The 13th Annual Conference on Information Technology Education. Alberta: ACM, 2012
[144] Yoav Bergner, Zhan Shu, and Alina A. von Davier. Visualization and Confirmatory Clustering of Sequence Data from a Simulation-Based Assessment Task [C] //Proceedings of the 7th International Conference on Educational Data Mining. London: IEDMS, 2014
[145] B. Dong, Q. H. Zheng, J. Yang, et al. An e-Learning ecosystem based on cloud computing infrastructure [C] //Proc. 9th IEEE International Conference on Advanced Learning Technologies. Beijing: IEEE Compute Society, 2009
[146] B. Dong, Q. H. Zheng, M. Qiao, et al. BlueSky cloud framework: an e-Learning framework embracing cloud computing [C] //Proceedings of the 1st international conference on cloud computing. Berlin: Springer, 2009