基于数据挖掘技术的在线学习行为研究综述*

柴艳妹，雷陈芳
（中央财经大学 信息学院，北京 100081）

摘 要：随着慕课快速发展为当下最新、最潮的学习形式，在线学习平台积累了大量学习行为数据，数据挖掘技术被引入在线学习行为的研究，从而涌现出大量的研究成果。为了深入分析在线学习行为研究中数据挖掘技术的整体应用情况，从国内外公认的 Web of Science 数据库收集 2008—2017 年 3 月相关文献进行了统计和可视化分析，介绍了利用数据挖掘技术进行在线学习行为研究的一般流程，并将数据挖掘技术在在线学习行为研究中的应用总结归纳为五类，详细介绍了相关研究成果及代表文献。最后总结并讨论了未来可能的研究方向。

关键词：慕课；在线学习行为；数据挖掘；可视化分析
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2018)05-1287-07
doi:10.3969/j.issn.1001-3695.2018.05.002

Survey of online learning behavior research applying data mining technology

Chai Yanmei, Lei Chenfang
(*School of Information, Central University of Finance & Economics, Beijing* 100081, *China*)

Abstract: With online learning platforms such as MOOCs becoming the latest and the most popular form of learning, the online learning platform had accumulated a large volume of learning data. Consequently data mining was widely used in research on online learning behavior, and massive outstanding achievements related to this had emerged. In order to understand the general situation of online learning research applying data mining, this paper first analyzed the relative literatures using visualization method. It searched the literatures data from Web of Science from 2008 to 2017. Secondly it introduced the general process of applying data mining to online learning behavior research. Then it discussed the data mining technology in the application of five categories in the study of online learning behavior, and summarized the relevant research results and literature. Finally it summarized and discussed the possible future research directions.

Key words: massive open online courses(MOOC); online learning behavior; data mining; visualization analysis

0 引言

近年来互联网技术的高速发展给传统教育注入了新鲜血液，催生了在线学习(e-learning)这一新兴的学习方式。广义的在线学习是指运用电子技术进行的学习行为，包括基于电视会议、CDROM 和网络的学习行为，而目前的在线学习概念，则专指基于网络的学习[1]。作为远程教育的重要形式，在线学习为学习者提供了大量的在线资源，打破了学习时间和空间的限制，满足了学习者多样化的学习需求。从 1989 年美国凤凰城大学推行在线学位计划起，众多学校和公司进行了大量的尝试，使大学公开课风靡一时。随着互联网的普及，2012 年，一种基于网络、针对大众人群的大规模开放在线课程(massive open online courses，MOOC)井喷式涌现。Coursera、EdX、Udacity 三大 MOOC 平台迅速发展，参与在线学习的人数不断增长。截至 2013 年 11 月 18 日，Coursera 的注册人数达到 540 万，Udacity 的注册人数超过 100 万，EdX 的注册人数也超过 90 万[2]，在线学习成为一种深受欢迎的学习方式。在线学习平台将教学者和学习者的行为完整记录，产生了大量连续的教—学互动信息。这些信息表征了学习者零散、无意识的学习行为，是深入研究学习行为和学习心理的新素材，通过对其深入分析，能反映出学生最真实的思维和学习情况。在理论方面有助于研究学习的本质、学习者的学习心理和学习行为，在实际应用方面有助于跟踪学习者的学习过程、评价学习效果、准确把握学习者的学习状态，以便及早进行干预，从而提高学习效率和学习质量。因此，在线学习行为研究引起了越来越多研究者的关注。

由于在线学习过程中产生的数据庞杂且包含大量非结构信息，使用简单的统计分析方法不容易发现其隐藏的知识和规律，所以研究者将视线投向了数据挖掘技术。数据挖掘是从大型数据集中挖掘隐含在其中的人们事先不了解、对决策有用的知识的过程，具有允许数据集不完全、有噪声、不确定、包含各种存储形式的优势，至今已广泛应用于各个领域。一些研究人员也开始尝试使用数据挖掘技术对在线学习数据进行深度分析，并取得了较好的成果。本文旨在综述数据挖掘技术在在线学习行为分析中的应用现状，并进一步展望未来的研究趋势。a)使用可视化分析方法从定量角度了解 2008—2016 年此研究

收稿日期：2017-04-17；**修回日期**：2017-06-08 **基金项目**：中央财经大学教改项目(020650514003)；中央财经大学课程教学团队建设项目(011459014008/032)

作者简介：柴艳妹(1978-)，女，河南焦作人，副教授，博士，主要研究方向为图像处理与模式识别(chai-4@163.com)；雷陈芳(1992-)，女，福建人，硕士，主要研究方向为在线学习行为分析.

领域文献发表的大致情况;b)介绍使用数据挖掘技术进行在线学习行为分析的流程、常用数据挖掘方法及常用工具;c)针对数据挖掘技术所能解决的在线学习行为研究问题进行了归纳总结和深入分析;d)展望了该领域的研究和发展前景。

1 文献发表情况分析

为了了解数据挖掘技术在在线学习分析领域中的应用现状,本文在 Web of Science 核心合集数据库中进行了主题检索,时间限制为 2008—2017 年 3 月。并对检索结果进行了简单统计分析,同时利用 Citespace 可视化软件生成科学知识图谱。

1.1 “在线学习行为研究”文献分析

在 Web of Science 核心合集数据库中以“e-learning”为主题进行检索,共检索到论文 12 260 篇,以“learning behavior”为主题进行精炼,检索结果为 875 篇。2008—2016 年每年发文量如图 1 所示。从检索结果可知,该领域从 2008 年开始每年发文量逐年增加,意味着在线学习行为研究逐渐引起研究者的关注。2012 年有所减少,这可能是由于 2012 年三大 MOOC 平台上线并迅速流行引起广泛关注,导致了在线学习研究领域内的研究方向的转变。但 2013 年又大幅上升,并超过下降前的发文量,说明 MOOC 重新激起了研究者的热情。

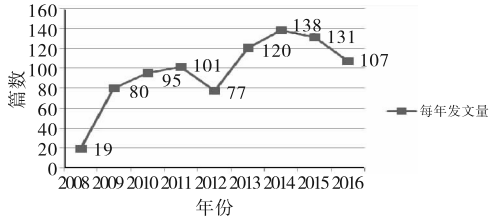

图1 2008—2016年在线学习行为研究每年发文量

通常文献的关键词可以反映出相关研究的关键问题和核心思想,因此本文将 2008—2016 年的文献数据导入 Citespace 中,选择“keyword”节点类型,得到研究热点领域图谱如图 2 所示。图中每个节点代表一个关键词,节点大小表示关键词出现的频数,“年轮”的颜色表示论文发表的年份,紫色圆环层厚薄表明该节点的重要程度。从图 2 中可看出,该领域的研究热点关键词主要有教育、行为、在线学习、信息技术、互联网、用户接受度、技术接受模型、建模、学生等,由此可知当前在线学习行为研究热点主要集中于使用者对在线学习这一新兴学习方式或技术的接受程度、对学习者学习行为和学习表现的建模以及在线学习系统设计等方面。

为了探究 2012 年前后该领域研究热点的转变情况,本文使用 Citespace 分别对 2008—2011 年和 2012—2016 年的文献进行分析,结果如图 3、4 所示,图中没有显示共同关键词“e-learning”,并且标注出了频率大于 10 的关键词。从分析结果可知,与 2008—2011 年间相比,2012—2016 年间研究者依然关注学习者对在线学习的接受程度,这一期间新增的研究热点包括学习者的学习表现、学习态度、学习动机以及学习风格等。除此之外,2012—2016 年间在线学习行为研究热点领域图谱中出现关键词“教育数据挖掘”,说明数据挖掘技术开始受到研究者的重视。

1.2 “基于数据挖掘技术的在线学习行为研究”文献分析

在线学习平台积累的大量学习数据使研究者萌生了将数据挖掘技术引入在线学习行为研究的想法,研究者开始尝试利用数据挖掘技术发现在线学习数据中隐藏的知识和规律,取得了较好的成果。数据挖掘技术的应用促进了在线学习行为研究的发展,并催生了新的研究方向和研究热点。为了探究当前数据挖掘技术与在线学习行为研究结合的情况,本文分别以 e-learning + learning behavior、e-learning + learning behavior + data mining、e-learning + data mining 为主题在 Web of Science 核心合集数据库中进行检索,检索结果如表 1 所示。从检索结果可知明确使用数据挖掘方法进行在线学习行为分析的论文较少,这一方向仍处于刚刚起步阶段,有待更深入的研究。由于多次精炼可能导致相关文献被忽略,且文献数量过少将不利于分析,因此本文采用检索主题 e-learning + data mining 的检索结果进行后续分析。

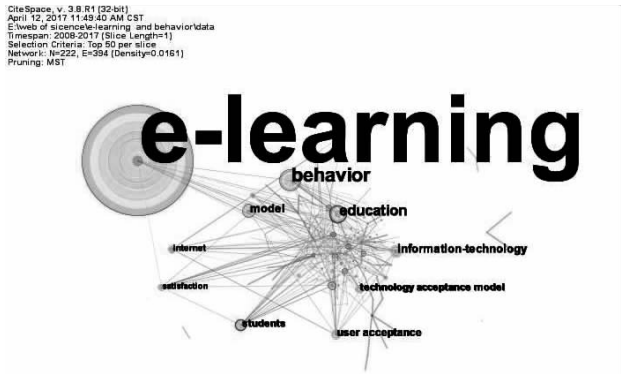

图2 2008—2016年在线学习行为研究热点领域图谱

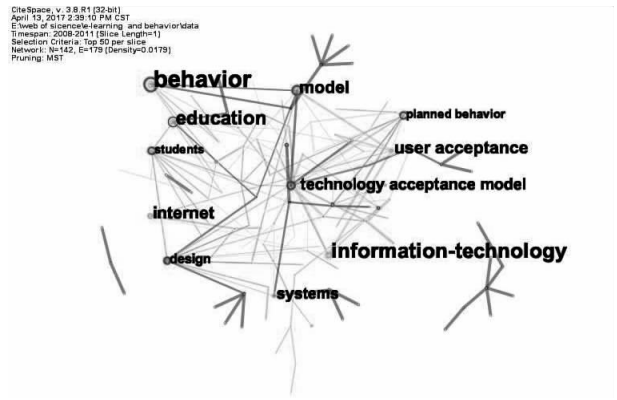

图3 2008—2011年在线学习行为研究热点领域图谱

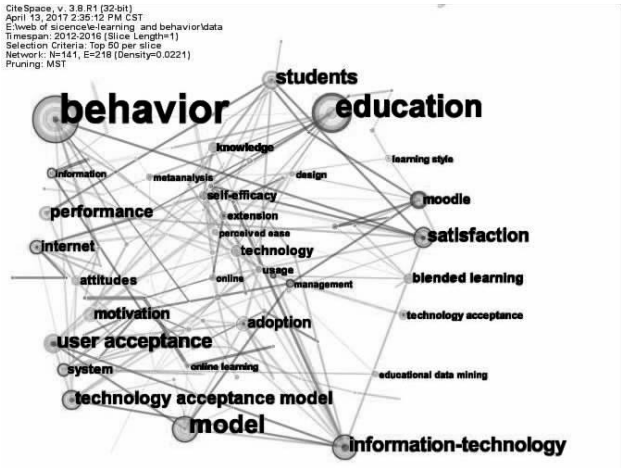

图4 2012—2016年在线学习行为研究热点领域图谱

2008—2016 年基于数据挖掘技术的在线学习行为研究每年发文量如图 5 所示。大体上,该研究方向的发文量变化趋势

与在线学习行为研究的相似,但数量明显较少。说明该研究方向还是新兴的研究领域,研究者不多,每年发文量较少。

表 1　2008—2017 年 3 月 Web of Science 相关主题文献数量

检索主题	文献数
e-learning	12 260
e-learning + learning behavior	875
e-learning + learning behavior + data mining	70
e-learning + data mining	247

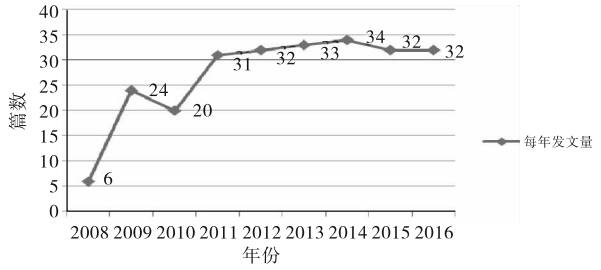

图5　2008—2016年基于数据挖掘技术的在线学习行为研究每年发文量

将文献数据导入 Citespace 分析基于数据挖掘技术的在线学习行为研究热点,结果如图 6 所示。表 2 截取了频数和中介中心度排序前 10 的关键词。

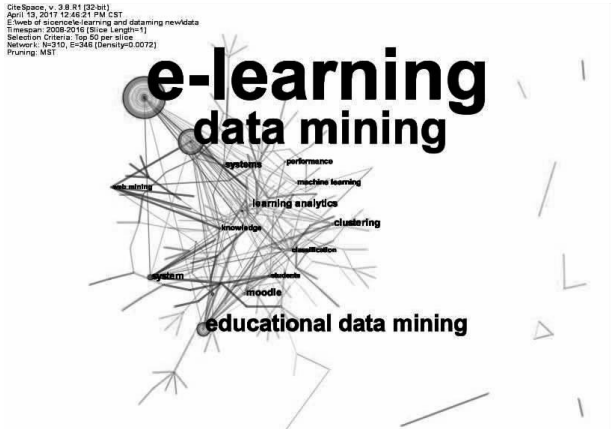

图6　2008—2016年基于数据挖掘技术的在线学习行为研究热点领域图谱

表 2　2008—2016 年"基于数据挖掘技术的在线学习行为研究"高频及高中介中心度关键词

排名	高频关键词		高中介中心度关键词	
	关键词	频数	关键词	中介中心度
1	e-learning	106	data mining	0.39
2	data mining	67	e-learning	0.38
3	educational data mining	34	system	0.24
4	clustering	14	educational data mining	0.23
5	moodle	14	knowledge	0.11
6	system	13	students	0.11
7	learning analytics	13	courses	0.11
8	knowledge	10	classification	0.09
9	machine learning	10	clustering	0.08
10	performance	10	moodle	0.07

综合图 6 和表 2,得到热点关键词包括教育数据挖掘、知识、系统、聚类、Moodle、学习分析、分类、Web 挖掘、机器学习等。其中,教育数据挖掘包含使用数据挖掘技术对在线学习行为进行分析,其含义更广;Web 挖掘是研究者常用于进行在线学习行为分析的方法,因为当前大量的在线学习平台采用网站形式,在线学习行为具体表现为对网站资源的访问并记录于日志文件中,通过对网站访问数据的挖掘可达到分析在线学习行为的目的;聚类和分类是常用的数据挖掘方法;Moodle 是研究中常使用的学习行为数据来源。

2　利用数据挖掘技术进行在线学习行为研究的一般流程

Romero 等人[3]认为数据挖掘技术在在线学习系统中的应用是一个循环迭代的过程,挖掘得到的知识应该进入系统的循环中,支持决策、改进学习系统、改善学习者的学习,而不是仅将数据转换成知识就结束。在线学习行为研究中应用数据挖掘的流程包含四阶段,分别为数据收集、数据预处理、应用数据挖掘方法以及解释、评估和应用[4],如图 7 所示。

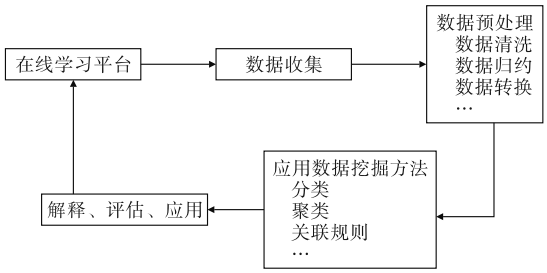

图7　在线学习行为研究中应用数据挖掘技术的一般流程

2.1　数据收集

数据收集一般由在线学习平台自动完成。学习者对学习平台的使用信息、互动信息将被记录于数据库中或以日志形式记录。在线学习行为研究的数据来源除了 EdX、Coursera、Udacity 等 MOOC 平台这一新兴数据来源外,还包括学习管理系统(LMS)和智能辅导系统(ITS)等,其中学习管理系统(如 Moodle)是当前在线学习行为研究的最主要数据来源。

2.2　数据预处理

数据预处理是对数据进行清洗并转换成适合数据挖掘的形式的过程,主要包括数据清洗、数据转换和丰富、数据整合以及数据归约等任务。此阶段还可以从已有数据中筛选进行挖掘分析的数据,整合特征,对数据进行离散化处理等。

2.3　数据挖掘方法

将原始数据处理好后,使用数据挖掘方法对其进行分析,寻找对教师、学生和平台管理者有意义的知识。当前研究者进行在线学习行为研究常使用的数据挖掘方法有分类、聚类、关联规则挖掘、序列模式挖掘、文本挖掘等。分类主要应用于预测学习者成绩、是否辍学等重要信息,以及根据分类标准抽取各类学习者的特征进行个性化教育。聚类可用于对学生进行分组,发现具有相似学习特点和行为模式的学生群体,以改善基于小组的协作学习[5]。聚类亦可用于孤立点分析,检测学习者的异常学习行为、欺诈行为,以便及早提醒教师作出相应处理。此外,聚类还可以作为其他挖掘方法的预处理步骤,将聚类的结果用于进一步的数据挖掘分析,得到每类学习者更深层的、未知的特征,从而提高精确度和挖掘效率;或将进一步分析得到的结果进行类间比较,发现不同群体间的差异。关联规

则和序列模式挖掘多被应用于发现学习者在线学习过程中的学习和使用习惯等,例如经常出现的网页浏览路径和学习活动安排次序,进而实施个性化的学习推荐和学习安排。目前应用文本挖掘进行在线学习行为研究的文献不多,主要是对论坛的发言内容和聊天记录进行分析。除此之外,马尔可夫模型、社会网络分析、模糊集理论和遗传算法等方法也曾被研究者应用于在线学习行为研究领域。表3总结了在线学习行为研究常用的数据挖掘方法及其主要应用和代表算法。

表3 在线学习行为研究常用数据挖掘方法

数据挖掘方法	主要应用	代表算法
分类	预测; 抽取各类学习者特征,进行个性化教育	决策树方法[6,7]、贝叶斯方法[8,9]、神经网络方法[10]、SVM、KNN等
聚类	发现相似群体,促进协作学习; 发现学习者特征、行为模式; 孤立点分析,检测异常学习行为; 其他挖掘方法的预处理步骤	K-means算法[11,12]、EM算法[13]、层次聚类法[14]等
关联规则挖掘	发现学习者的学习习惯、使用习惯;个性化教育	Apriori算法[7,15]、RP-tree算法等
序列模式挖掘		AprioriAll算法[16]、GSP算法[17]、PrefixSpan算法
文本挖掘	分析论坛或讨论版的发言内容和聊天记录	抽取关键词[18]

在应用数据挖掘技术进行在线学习行为研究的过程中,除了少部分研究者编程实现数据挖掘算法,大多数研究者都是利用现有的数据挖掘工具所提供的算法。考虑到研究成本问题,相比DBMiner、SPSS Clementine、DB2 Intelligent Miner等商业软件,研究人员更常使用Weka和KEEL等开源的软件进行在线学习行为研究。此外,有研究者还开发了专门进行教育数据挖掘的工具,如the Mining tool[19]、MultiStar[20]、GISMO[21]以及Sequential mining tool[22]等。

2.4 解释、评估和应用

最后,需要结合在线学习情境对挖掘结果进行解释,分析其背后的实际含义,并利用获取的知识进行决策。a)教师及学习平台管理者可以利用得到的学习规律、认知规律等理论知识和学习者的特征及习惯对教学过程与学习平台的设计等进行调整,了解学习者学习现状,及时发现表现不佳的学习者,以便进行及时干预;b)学习者可以通过分析结果更加了解自身特点、学习状态和知识掌握程度等,以便作出自我调整,以提高学习效率和学习效果。

3 数据挖掘技术在在线学习行为研究中的应用

在线学习行为研究的最终目的是改善学习,提高学习者的学习效率和效果。因此本文按照数据挖掘技术所能解决的问题将当前的研究总结归纳为五类,分别是发现学习规律、分析学习特征及习惯、评估学习现状、预测学习效果和个性化学习服务。

3.1 发现学习规律

研究人员使用数据挖掘技术对在线学习平台记录的数据进行分析,期望发现客观的学习规律和认知规律。

1)影响学习的因素研究

影响学习的因素是研究人员最为关注的主题,众多研究者探究了各种因素与学习成绩之间的关系。如Natek等人[6]使用决策树算法对学生分类,得到不同成绩等级学生在个人信息和教学环节表现等方面的特征,分析发现影响课程最终成绩的因素有学习类型、学习过程中的活动表现和平时测试成绩等。Chanchary等人[7]使用关联规则挖掘和决策树分类方法发现学习者在学习管理系统上的使用行为与最终成绩之间存在关系。He Wu[18]使用文本挖掘技术对学习者在线学习过程中的问答和聊天记录进行分析,最终发现学生提问次数与期末成绩之间的关系。Vaessen等人[11]建立离散马尔可夫模型,并应用K-means聚类算法和Logistics回归等方法分析了智能辅导系统中学生求助策略与成绩之间的关系。Cho等人[23]使用社会网络分析方法分析了计算机支持协作学习社区中交流风格、社会网络和学习成绩之间的关系。Ding等人[24]利用计算机支持协作学习平台进行实验,使用逐步回归等方法对学生协作解决物理问题过程中产生的会话进行分析,得出性别或性别组合对协作学习模式及学习效果的影响。李曼丽等人[25]使用Tobit回归和Logistics回归分析影响学生课程完成度的因素。

2)学习行为模式研究

学习行为模式研究是当前最受研究者关注的热点话题之一。如Yu等人[26]使用在线时间、阅读文献数、提问数量等特征描述学习者行为,应用模糊关联规则挖掘方法探索了每种学习行为模式间的关系。Talavera等人[13]利用最大期望(EM)算法进行聚类分析,发现学习者在协作学习中的行为模式。Araya等人[27]对一个大型多人在线数学游戏的数据进行聚类分析,发现了学生团队协作中的规律。Perera等人[17]利用软件开发项目中常用的内容管理、任务管理和代码管理工具,运用聚类和序列模式挖掘技术,研究学生参加软件开发项目时表现优秀和较差的小组在使用工具时的行为差异。Chen等人[15]使用关联规则挖掘方法发现学生对某些知识点的误解会导致后续其他特定知识点的误解。吴淑苹[28]将学习者的学习行为动作进行统一编码后使用序列模式挖掘算法,得到学习者的学习路径网状图,并分析其中的高频度学习路径发现有效学习模式。李爽等人[29]使用聚类方法将学习者根据其行为序列特征划分成五类,通过分析各类学习者在线学习产生的行为序列和行为转换模式将这五类分别定义为低投入式、浅层投入式、绩效投入式、循序渐进式以及随机参与式。

3.2 分析学习特征及习惯

了解和分析学生的学习特征和学习习惯,可更好地改进课程安排、实施个性化教学、提升用户体验、提高学习效率。

1)学习特征研究

当前研究者关注较多的是认知方式和学习风格两方面的特征,如Feldman等人[8]使用学生在某游戏上的尝试次数、持续时间和最终等级的数据训练朴素贝叶斯分类器,判断学生的学习风格。Damez等人[30]使用模糊决策树的方法对学生建模,通过分析学生与学习系统的互动行为所体现出的认知特征,区分有在线学习经验者和新手。Ayers等人[14]使用基于层级的聚类、K-means和基于模型的聚类三种方法进行认知诊断。Jovanovic等人[31]使用K-means算法依据学生行为将学生分为特别好、好、差三类,并结合MBTI量表收集的学生认知方式信息分析各类学生的认知方式特点。吴青等人[32]借助Kolb学习风格量表获得学生的学习风格,并使用关联规则挖掘方法获得各种学习风格的行为特征。

2)学习习惯研究

当前主要是探究学习者浏览学习资料的习惯,即学习者浏览学习资料的先后顺序(序列关系)及哪些学习资料常被一同浏览(相关关系)等。Wang[33]实现了基于关联规则和序列模式挖掘技术的分析工具,展示学习内容的动态浏览结构,发现学习资料间的关联关系和序列关系。Romero 等人[16]使用 AprioriAll、GSP、PrefixSpan 等序列模式挖掘算法对动态超媒体学习系统 AHA!上的学习数据进行分析,获得知识点间的序列关联关系。

3.3　评估学习现状

评估学生学习现状包括评估学生对知识的掌握情况、检测不良学习心态及行为等,有利于教师及时纠正补全错漏知识、修改后续的课程安排、干预学习者的异常行为等。

目前已有部分研究者尝试使用数据挖掘技术评估学生的学习现状,如 Baker 等人[34]使用机器学习的潜在相应模型,检测学生对智能辅导系统的误使用,并训练分类器识别学习者是否欺骗系统。Pahl 等人[35]利用序列模式挖掘方法获得在线学习者的学习路径,并与教师预期路径进行比较,进而发现偏离预期者及时干预。Chen 等人[9]以 Twitter 上标签为“Engineering Problem”的微博为样本,借助文本处理技术,训练朴素贝叶斯多标签分类器,并使用该分类器分析美国普渡大学附近发表的微博,推测学生当前面临的问题。Reffay 等人[36]为了研究小组协作学习状况,应用社会网络分析方法计算协作学习中学习小组的内聚性,并识别孤立的人群、活跃的次级小团体,以及分析组内交流中各成员的角色。Rajendran 等人[37]结合心理学、教育学理论,认为追求目标的过程中遭遇的阻碍导致挫折,通过分析学生使用智能辅导系统的目标,研究达成目标的阻碍因素,以具体的阻碍因素为自变量建立线性模型,判断学习者在使用智能辅导系统时的情感状态(是否感到挫折),提醒教师及时作出干预化解学生不良情绪,改善学习者学习状态。王卓等人[38]依据 MOOC 的特点对贝叶斯知识跟踪模型进行改进,提出了基于知识点和测试提交历史的贝叶斯知识跟踪模型,从而更准确地推断学习者是否掌握各个知识点。

3.4　预测学习效果

研究者使用学习者已修课程的历史数据建立模型,单纯依赖数据相关关系或结合教育理论,预测学习者的学习效果。

Anozie 等人[39]对学习者每月在线学习的学习日志进行线性回归分析,预测期末考试成绩。Arroyo 等人[40]利用智能辅导系统中的学习数据,从分析变量间的相关关系开始,先使用极大似然方法学习条件概率,构建贝叶斯网络,推断学生的态度是积极还是消极,再建立动态贝叶斯网络对学生知识行为进行建模,预测其未来的表现。宗阳等人[41]使用 Logistics 回归建模预测模型,依据 MOOCs 学习者的学习行为预测其学习成绩。蒋卓轩等人[42]利用北京大学在 Coursera 上开设的六门慕课的学习行为数据分析了学习者类型,并分别使用线性判别分析、Logistics 回归和线性核支持向量机等方法预测学习者的学习结果。

除了学习效果,研究者还十分关心在线学习者未来的行为,如 Pedro 等人[43]使用 Logistics 回归预测使用 ASSISTment 智能辅导系统的中学生未来是否进入大学。Arnold 等人[44]使用多元线性回归预测学习者花费多长时间进行在线学习。Cocea 等人[45]利用分类方法识别出缺乏学习动力的学生,及时采取补救行动降低辍学率。

一些研究者在建立预测模型的基础上,探究如何提高预测准确度,为此尝试多种数据挖掘方法,比较各种预测方法的性能。如 Feng 等人[46]使用逐步回归分析探究导致对学习者考试成绩预测性能不佳的错误来源。Romero 等人[48]使用 25 种流行的分类器对科尔瓦多大学学生进行分类,比较分类器的各项性能。为了提高分类或预测的准确度,Minaei-Bidgoli 等人[46]使用遗传算法对二次型贝叶斯分类器、1-最近邻、K-最近邻、Parzen 窗估计、多层神经网络和决策树等六种不同分类方法进行组合得到最佳分类器。

3.5　个性化学习服务

个性化学习服务,即根据学生的特点、当前学习情况,向其推荐课程、学习活动、学习资料以及学习方法等,提供学习建议,动态调整学习安排,是当前在线学习行为研究的热点问题之一。

Chu 等人[49]利用 Apriori 算法设计了一个基于 Web 的课程推荐系统,用以为面临选课问题的学习者提供建议。Lu[50]利用模糊规则推测学生对教材的需求,实现个性化学习材料推荐。Teng 等人[51]依据学习行为的相似性对学习者进行聚类,分析聚类结果后为每类学习者提供针对性建议。Wang 等人[10]利用 BP 神经网络方法实现了一个自适应英语学习系统,可根据学习者的性别、性格和学习焦虑程度等推荐不同难度的学习材料。Aher 等人[12]使用在 K-means 算法聚类的基础上,应用 Apriori 算法对各类学生的课程学习记录进行关联规则分析,得到各类学生偏好的课程学习顺序,从而向学生推荐合适的课程。

4　结束语

本文通过对 Web of Science 核心数据库中 2008—2017 年在线学习行为相关领域的研究文献进行统计和可视化分析发现,从 2012 年起,应用数据挖掘进行在线学习行为研究的文献逐渐增多,数据挖掘技术正逐渐受到该领域研究者的青睐。但从数量上看,这一领域的文献数量仍然很少,说明其正处于新兴阶段,有待深入研究。从研究内容上看,数据挖掘技术和在线学习行为研究的结合已经取得一些成果,研究者使用数据挖掘技术解决的在线学习行为研究问题涵盖发现学习规律、分析学习特征及习惯、评估学习现状、预测学习效果和个性化学习服务等主题,涉及范围广泛,表明数据挖掘技术在在线学习行为研究中的应用前景广阔。

随着 MOOC 等新兴在线学习环境的兴起和流行,积累了海量的在线学习数据,使得基于数据挖掘技术的在线学习行为研究受到越来越多研究者的关注,并将逐步走向成熟。未来可能的研究方向和需要解决的问题有:

a)缺少公开的研究数据集是制约该领域发展的瓶颈之一。大多研究论文中的数据是来源于高校内部的在线学习系统或远程教育平台的私有数据,外部研究者无法获取。虽然 MOOC 平台的广泛使用创造了大量的在线学习记录,但由于隐私保护、相关利益等原因,用户使用数据并没有公开。目前已公开的数据有哈佛大学和麻省理工学院联合发布的 edX 平台

16 门课程数据以及 UCI 数据库中发布的少量教育数据。由于各在线学习系统记录的信息不尽相同,大多数研究算法都是应用于小数据集,且不具有可移植性,普适性不高。此外,很多研究使用的数据内容单一,多为某些学习活动的发生次数、发生时间,相比在线学习系统记录的用户日志信息,信息利用度不高。如何利用现有信息丰富数据内涵、有效提取特征,并进行深入数据挖掘分析有待进一步探索。

b)数据挖掘算法的应用创新将会是新的研究方向。目前基于数据挖掘技术的在线学习行为研究多采用封装好的成熟算法,如何针对在线学习的应用场景进行改进、设计专业的学习数据分析工具有待研究。例如,在预测学习效果时,除了直接调用 Logistics 回归、决策树、神经网络、SVM、AdaBoost 等传统的分类工具外,还可以对这些算法的过程和参数设定方法加以改进,使其能更好地解决在线学习应用中出现的问题。另外,针对具体问题,也可将其他学科中的算法加以改进并和数据挖掘算法结合,提出新的解决方案。如在学习效果影响因素分析方面,可以利用信息论中的信息增益、最小描述长度互信息等度量各个影响因素的影响程度,还可以利用特征选择理论中基于样本距离的 ReliefF 法[52]、基于关联性度量的 CFS 法[53]、基于最小相关性的 MRR[54] 等算法来解决各影响因素的冗余性问题。

c)数据挖掘技术将会和教育学、心理学等学科理论共同驱动在线学习行为的研究。当前大多数基于数据挖掘技术的在线学习行为研究很少与教育学、心理学等学科的理论深入联系,但这些学科在学习行为等方面积累了很多先进的理论。例如,行为主义学者斯金纳提出的操纵条件反射理论[55],认知学派布鲁纳的认知—发现说[56] 以及构建学习主义的支架式教学和抛锚式教学[57] 等理论都可以在在线学习平台上进行新的研究和验证。借助数据挖掘这一工具,可以全面有效地进行学习行为分析,发现学习规律、了解学习过程受哪些因素影响以及如何才能有效学习。此外,美国心理学家奥苏贝尔的动机理论[58] 和耶基斯—多德森定律[59] 也可引入在线学习的行为研究,使用数据分析的方法研究学习动机和学习效果之间的关系。总之,如何既利用已知客观规律的巨人肩膀,又发挥技术挖掘由数据驱动打破常规发现新模式、新知识的优势,将教育学、心理学等理论知识与数据挖掘技术结合建立知识与数据共同驱动模型将会成为未来的研究趋势。

参考文献:

[1] 杨为民. 在线学习的现状与发展研究[D]. 兰州:西北师范大学, 2007.

[2] Laura pappano. The year of the MOOC [EB/OL]. (2012-11-04). http://www. nytimes. com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace. html? pagewanted = all&_r = 0.

[3] Romero C, Ventura S. Data mining in education[J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2013, 3(1): 12-27.

[4] Romero C, Ventura S, García E. Data mining in course management systems: moodle case study and tutorial[J]. Computers & Education, 2008, 51(1): 368-384.

[5] Tang T Y, McCalla G. Smart recommendation for an evolving e-learning system[J]. International Journal on E-Learning, 2005, 4(1): 105-129.

[6] Natek S, Zwilling M. Student data mining solution-knowledge management system related to higher education institutions[J]. Expert Systems with Applications, 2014, 41(14): 6400-6407.

[7] Chanchary F H, Haque I, Khalid M S. Web usage mining to evaluate the transfer of learning in a Web-based learning environment[C]// Proc of International Workshop on Knowledge Discovery and Data Mining. Piscataway, NJ: IEEE Press, 2008: 249-253.

[8] Feldman J, Monteserin A, Amandi A. Detecting students' perception style by using games[J]. Computers & Education, 2014, 71(2): 14-22.

[9] Chen Xin, Vorvoreanu M, Madhavan K P C. Mining social media data for understanding students' learning experiences[J]. IEEE Trans on Learning Technologies, 2014, 7(3): 246-259.

[10] Wang Y H, Liao H C. Data mining for adaptive learning in a TESL-based e-learning system[J]. Expert Systems with Applications, 2011, 38(6): 6480-6485.

[11] Vaessen B E, Prins F J, Jeuring J. University students' a chievement goals and help-seeking strategies in an intelligent tutoring system[J]. Computers & Education, 2014, 72(11): 196-208.

[12] Aher S B, Lobo L M R J. Combination of machine learning algorithms for recommendation of courses in e-learning system based on historical data[J]. Knowledge-Based Systems, 2013, 51(1): 1-14.

[13] Talavera L, Gaudioso E. Mining student data to characterize similar behavior groups in unstructured collaboration spaces[C]//Proc of European Conference on Artificial Intelligence. 2004: 17-23.

[14] Ayers E, Nugent R, Dean N. A comparison of student skill knowledge estimates[C]//Proc of International Conference on Educational Data Mining. 2009: 1-10.

[15] Chen C M, Hsieh Y L, Hsu S H. Mining learner profile utilizing association rule for Web-based learning diagnosis[J]. Expert Systems with Applications, 2007, 33(1): 6-22.

[16] Romero C, Ventura S, Zafra A, *et al*. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems[J]. Computers & Education, 2009, 53(3): 828-840.

[17] Perera D, Kay J, Koprinska I, *et al*. Clustering and sequential pattern mining of online collaborative learning data[J]. IEEE Trans on Knowledge & Data Engineering, 2008, 21(6): 759-772.

[18] He Wu. Examining students' online interaction in a live video streaming environment using data mining and text mining[J]. Computers in Human Behavior, 2013, 29(1): 90-102.

[19] Zaïane O R. Web usage mining for a better Web-based learning environment[C]//Proc of Conference on Advanced Technology for Education. 2001: 357-360.

[20] Silva D, Vieira M. Using data warehouse and data mining resources for ongoing assessment of distance learning[C]//Proc of IEEE International Conference on Advanced Learning Technologies. Washington DC: IEEE Computer Society, 2002: 40-45.

[21] Mazza R, Milani C. Exploring usage analysis in learning systems: gaining insights from visualizations [C]//Proc of Workshop on Usage Analysis in Learning Systems at the 12th International Conference on Artificial Intelligence in Education. 2005: 65-72.

[22] Morales C R, Soto S V, Pérez A R P, *et al*. Using sequential pattern mining for links recommendation in adaptive hypermedia educational systems[C]//Current Developments in Technology-Assisted Educa-

tion. 2006:1016-1020.

[23] Cho H, Gay G, Davidson B, *et al.* Social networks, communication styles, and learning performance in a CSCL community[J]. Computers & Education,2007,49(2):309-329.

[24] Ding N, Bosker R J, Harskamp E G. Exploring gender and gender pairing in the knowledge elaboration processes of students using computer-supported collaborative learning[J]. Computers & Education, 2011,56(2):325-336.

[25] 李曼丽,徐舜平,孙梦嫽. 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,2015,21(2):63-69.

[26] Yu P, Own C, Lin L. On learning behavior analysis of Web based interactive environment[C]//Proc of Implementing Curricular Change in Engineering Education. 2001:1-10.

[27] Araya R, Jiménez A, Bahamondez M, *et al.* Teaching modeling skills using a massively multiplayer online mathematics game[J]. World Wide Web: Internet & Web Information Systems,2014,17(2): 213-227.

[28] 吴淑苹. 基于数据挖掘的教师网络学习行为分析与研究[J]. 教师教育研究,2013,25(3):49-57.

[29] 李爽,钟瑶,喻忱,等. 基于行为序列分析对在线学习参与模式的探索[J]. 中国电化教育,2017(3): 88-95.

[30] Damez M, Bouchon-Meunier B, Ha T, *et al.* Fuzzy decision tree for user modeling from human-computer interactions[C]//Proc of the 5th International Conference on Human System Learning. 2005:287-302.

[31] Jovanovic M, Vukicevic M, Milovanovic M, *et al.* Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study[J]. International Journal of Computational Intelligence Systems,2012,5(3):597-610.

[32] 吴青,罗儒国,王权于. 基于关联规则的网络学习行为实证研究[J]. 现代教育技术,2015,25(7):88-94.

[33] Wang F H. On using data-mining technology for browsing log file analysis in asynchronous learning environment[C]//Proc of Conference on Educational Multimedia, Hypermedia and Telecommunication. Berlin:Springer,2002:2005-2006.

[34] Baker R, Corbett A, Koedinger K. Detecting student misuse of intelligent tutoring systems[C]//Proc of the 7th International Conference on Intelligent Tutoring Systems. 2004:531-540.

[35] Pahl C, Donnellan D. Data mining technology for the evaluation of Web-based teaching and learning systems[C]//Proc of Congress E-learning. 2003:1-7.

[36] Reffay C, Chanier T. How social network analysis can help to measure cohesion in collaborative distance-learning[M]. Berlin: Springer, 2003:343-352.

[37] Rajendran R, Iyer S, Murthy S, *et al.* A theory-driven approach to predict frustration in an ITS[J]. IEEE Trans on Learning Technologies,2013,6(4):378-388.

[38] 王卓,张铭. 基于贝叶斯知识跟踪模型的慕课学生评价[J]. 中国科技论文,2015,10(2):241-246.

[39] Anozie N, Junker B. Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system [C]//Proc of Educational Data Mining AAAI Workshop. Palo Alto, CA:AAAI Press,2006:1-6.

[40] Arroyo I, Murray T, Woolf B P, *et al.* Inferring unobservable learning variables from students' help seeking behavior[C]//Proc of International Conference on Intelligent Tutoring Systems. Berlin:Springer, 2004:782-784.

[41] 宗阳,孙洪涛,张亨国,等. MOOCs 学习行为与学习效果的逻辑回归分析[J]. 中国远程教育,2016,36(5):14-22.

[42] 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,2015,52(3):614-628.

[43] Pedro M O Z S, Baker R S J D, Bowers A J, *et al.* Predicting college enrollment from student interaction with an intelligent tutoring system in middle school[J]. Langmuir the Acs Journal of Surfaces & Colloids,2011,27(11):6897-6904.

[44] Arnold A, Scheines R, Beck J E, *et al.* Time and attention: students, sessions, and tasks[C]//Proc of AAAI Workshop Educational Data Mining. Palo Alto,CA:AAAI,2005:62-66.

[45] Cocea M, Weibelzahl S. Can log files analysis estimate learners' level of motivation? [C]//Proc of the 14th Workshop Adaptivity and User Modeling in Interactive Systems. 2006:32-35.

[46] Feng Mingyu, Heffernan N, Koedinger K. Looking for sources of error in predicting student's knowledge[C]//Proc of AAAI Workshop on Educational Data Mining. Palo Alto,CA:AAAI Press,2005:1-8.

[47] Romero C, Ventura S, Espejo P G, *et al.* Data mining algorithms to classify students[C]//Proc of the International Conference on Educational Data Mining. 2008:8-17.

[48] Minaei-Bidgoli B, Kashy D, Kortemeyer G, *et al.* Predicting student performance: an application of data mining methods with an educational Web-based system[C]//Proc of ASEE/IEEE Frontiers in Education Conference. 2003:T2A-13.

[49] Chu K, Chang M, Hsia Y. Designing a course recommendation system on Web based on the students' course selection records[C]// Proc of World Conference on Educational Multimedia, Hypermedia and Telecommunications. 2003:14-21.

[50] Lu J. A personalized e-learning material recommender system[C]// Proc of International Conference on Information Technology for Application. 2004:374-379.

[51] Teng Chaiwen, Lin C, Cheng S, *et al.* Analyzing user behavior distribution on e-learning platform with techniques of clustering[C]// Proc of Society for Information Technology & Teacher Education International Conference. Chesapeake,VA:AACE,2004:3052-3058.

[52] Sarrafzadeh A, Atabay H A, Pedram M M, *et al.* ReliefF based feature selection in content-based image retrieval[C]//Proc of International Multi Conference of Engineers and Computer Scientists. 2012: 19-22.

[53] Lu Xinguo, Peng Xianghua, Liu Ping, *et al.* A novel feature selection method based on CFS in cancer recognition[C]//Proc of IEEE International Conference on Systems Biology. Washington DC: IEEE Computer Society,2012:226-231.

[54] Li Biqing, Hu Lele, Chen Lei, *et al.* Prediction of protein domain with mRMR feature selection and analysis[J]. PLoS One,2012,7 (6):e39308.

[55] 阮晓钢,武璇. 斯金纳自动机:形成操作性条件反射理论的心理学模型[J]. 中国科学:技术科学,2013,43(12):1374-1390.

[56] 李俊霞. 布鲁纳的认知—发现说在成人教学中的应用[J]. 成人教育,2003(11):3-5.

[57] 吕雪晴,王华清,刘满芝. 基于建构主义学习理论的网络教学模式构建[J]. 教学与管理,2006(2):77-78.

[58] 韩亚梅. 奥苏贝尔学习教学理论及其对教学实践的启示[J]. 陕西广播电视大学学报,2008,10(3):29-32.

[59] 于连科. 学习动机理论在远程教育中的应用[J]. 现代远距离教育,2009(3):16-20.