软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(11):3026−3042 [doi: 10.13328/j.cnki.jos.004887]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

教育数据挖掘研究进展综述*

周 庆[1,2], 牟 超[2], 杨 丹[3]

[1](信息服务社会可信服务计算教育部重点实验室(重庆大学),重庆 400044)

[2](重庆大学 计算机学院,重庆 400044)

[3](重庆大学 软件学院,重庆 400044)

通讯作者: 周庆, E-mail: tzhou@cqu.edu.cn, http://www.cqu.edu.cn

摘 要: 教育数据挖掘(educational data mining,简称 EDM)技术运用教育学、计算机科学、心理学和统计学等多个学科的理论和技术来解决教育研究与教学实践中的问题.在大数据时代背景下,EDM 研究将迎来新的转折点.为方便读者了解 EDM 的研究进展或从事相关研究和实践,首先介绍 EDM 研究的概貌、特点和发展历程,然后重点介绍和分析了 EDM 近年来的研究成果.在成果介绍部分,选取的研究成果大部分发表于 2013 年以后,包括以往较少涉及的几种新型教育技术.在成果分析部分,对近年来的典型案例作了分类、统计和对比分析,对 EDM 研究的特点、不足及发展趋势进行了归纳和预测.最后讨论了大数据时代下 EDM 面临的机遇和挑战.

关键词: 大数据;教育环境;交叉学科;MOOCs;ITS

中图法分类号: TP311

中文引用格式: 周庆,牟超,杨丹.教育数据挖掘研究进展综述.软件学报,2015,26(11):3026−3042. http://www.jos.org.cn/1000-9825/4887.htm

英文引用格式: Zhou Q, Mou C, Yang D. Research progress on educational data mining: A survey. Ruan Jian Xue Bao/Journal of Software, 2015,26(11):3026−3042 (in Chinese). http://www.jos.org.cn/1000-9825/4887.htm

Research Progress on Educational Data Mining: A Survey

ZHOU Qing[1,2], MOU Chao[2], YANG Dan[3]

[1](Key Laboratory of Dependable Service Computing in Cyber Physical Society of Ministry of Education (Chongqing University), Chongqing 400044, China)

[2](College of Computer Science, Chongqing University, Chongqing 400044, China)

[3](School of Software Engineering, Chongqing University, Chongqing 400044, China)

Abstract: Educational data mining (EDM) focuses on solving theoretical and practical problems in education by applying principles and techniques from educational science, computer science, psychology, and statistics. It is believed that EDM will become more mature and promising in the Age of Big Data. This paper aims to help readers to understand or engage EDM research. First, the basic concepts, characteristics and research history of EDM are introduced. Then some latest results of EDM are presented and analyzed. Most results were published in 2013 and later, including some studies on several educational techniques that were rarely investigated before. Those results are also analyzed via classification, statistics and comparison, and based on which strength and weakness of EDM is discussed. Finally, opportunities and challenges facing EDM are discussed.

Key words: big data; educational environment; interdisciplinary research; MOOCs; ITS

数据挖掘技术可以从大量的数据中发现隐藏的模式与知识[1],目前已成功应用在生物、金融和电子商务等

* 基金项目: 国家自然科学基金(61472464, 61402020); 中央高校基本科研业务费(CDJZR12.18.55.01, 106112015CDJSK04JD02); 重庆市前沿与应用基础研究计划(cstc2013jcyjA40017)

收稿时间: 2015-02-12; 修改时间: 2015-05-11, 2015-07-14, 2015-08-11; 定稿时间: 2015-08-26

© 中国科学院软件研究所 http://www.jos.org.cn

广泛的领域.近年来,在教育信息化、远程教育和 Web 2.0 等应用的带动下,教育数据挖掘(educational data mining,简称 EDM)开始受到越来越多的研究者的关注[2].

教育数据挖掘技术综合应用教育学、计算机科学、心理学和统计学等多个学科的理论和技术来解决教育研究与教学实践中的问题.通过分析和挖掘教育相关的数据,EDM 技术可以发现和解决教育中的各类问题,如辅助管理人员做出决策、帮助教师改进课程以及提高学生的学习效率等.教育问题的复杂性和多学科交叉的性质,使 EDM 在数据来源、数据特点、研究方法和应用目的等方面均表现出其独特性.

在过去几年中,教育领域和信息领域都发生了革命性的变化,在线学习系统、智能手机应用和社交网络为 EDM 研究提供了大量的应用和数据.以在线学习系统 MOODLE[3]为例,截至 2013 年,已为全球超过 6 000 万名学生和老师提供服务[4].截至 2012 年 6 月,全球智能手机用户人数超过 10 亿人[5],社交媒体 Facebook 的用户数超过 22 亿人[6].大规模公开在线课程(massive open online courses,简称 MOOCs)是近两年兴起的新型教学模式.截至 2014 年底,在 MOOCs 网站 Coursera 上注册的用户人数已超过 1 000 万[7].显然,EDM 也正处于一个“大数据”的时代.这一特殊的背景,预示着 EDM 研究将在近几年内迅速发展.与以往的 EDM 综述性论文相比,本文的主要贡献如下:

(1) 从教育环境的角度对 EDM 研究进行分类介绍.以往的 EDM 综述性论文一般按技术或应用目的对研究成果进行分类,本文按教育环境进行分类,以体现了 EDM“从教育中来,回到教育中去”的理念.

(2) 介绍了近两年的 EDM 研究进展.现有的 EDM 综述论文主要分析了 2012 年以前的研究成果,本文则以 2013 年~2014 年的研究成果为主,使读者了解这一领域的最新研究进展.特别地,增加了对一些新型教育技术(如 MOOCs 和移动计算)的研究成果的介绍.以往的 EDM 综述性论文很少涉及这些内容,本文对其作了介绍和总结.

(3) 对 EDM 研究的现状及发展趋势作了分析与评价.本文对近年来 EDM 的重要研究案例进行了分类、统计和对比分析,对当前 EDM 研究的特点与不足进行了归纳,同时预测了该领域的研究趋势.

(4) 展望了大数据时代下 EDM 的研究前景.大数据技术对教育的发展有着深远的影响,最新的 EDM 研究也印证了这一趋势.本文对这一时代背景下 EDM 研究面临的挑战和机遇进行了分析和展望.

本文首先介绍 EDM 的基本知识和一般研究过程.之后,重点对 EDM 近年的研究成果作分类介绍.然后对这些研究成果作分析与评价.最后,对大数据时代下的 EDM 研究进行总结与展望.

1 EDM 研究概述

1.1 EDM的特点

与 EDM 联系最紧密的学科分别是计算机科学、教育学和统计学,如图 1 所示[8].从图中可以看到,这三大学科两两交叉分别产生了数据挖掘与机器学习(data mining and machine learning,简称 DM&ML)、基于计算机的教育(computer-based education,简称 CBE)以及学习分析(learning analytics,简称 LA).通过与这 3 个领域的对比可以看出 EDM 的特点.

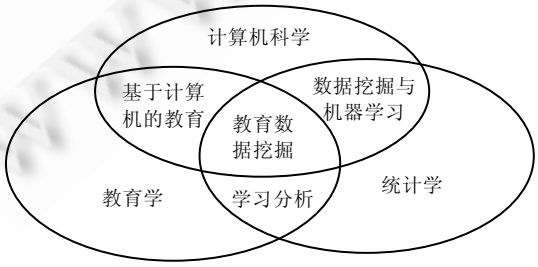

Fig.1 Main areas related to educational data mining[8]

图 1 EDM 涉及的主要学科[8]

© 中国科学院软件研究所 http://www.jos.org.cn

EDM 与一般的 DM&ML 研究的主要区别在于其数据的教育学科特色,表现在以下几个方面:

- 多学科:EDM 数据通常涉及教育学、心理学和社会学的概念和技术,如教学目的、学习经验、教学评估、兴趣、动机、团队协作、人际关系和家庭背景等.对这一类数据,研究者既要能理解其概念,也要掌握测量和评价它们的技术.
- 多层次:EDM 数据的多层次特性来自于教育机构和教学材料的结构性,如学生可按学区、学校、院系和班级进行组织,而教学内容可按课程、章节、知识点和概念进行组织.
- 多精度:EDM 数据通常包含时间刻度,一项教学研究可能跨越几年甚至一生,也可能以毫秒的精度进行记录.这使研究者可按不同的时间精度分析数据.
- 多情景:EDM 数据的多情景特性来自于教育学科本身的特点.一个学生获得知识的经验与教学的时间、地点、教师和环境相关,也与学生自身的动机、能力和情绪相关,以上任意要素的改变可能会导致不同的学习经验.
- 多语义:EDM 数据的多语义特性来自于几个方面,如师生的行为存在多义性、师生使用的自然语言存在多义性、教育环境中的噪声数据或缺失的数据会带来歧义,甚至不同教育理论对同一数据的解释也会导致多义性.

EDM 与一般 CBE 研究的主要区别在于应用目的的不同,后者的目标是辅助或替代传统的教学过程,而 EDM 则致力于实现传统教学缺少或难以完成的功能.表 1 总结了不同角色使用 EDM 的目的.

Table 1 Application purposes of EDM for different stakeholders

表 1 不同角色使用 EDM 的目的

角色	使用 EDM 的目的
学生	了解自己的性格、兴趣、能力和学习风格 了解自己的学习效率、学习效果和学习进展 向其推荐课程、学习资源和学习策略
教师	了解教学的效率,改进教学材料 了解学生的个体和总体情况 预测学生的学习成绩
管理人员	了解教育机构的历史与现状 提供决策支持,改进管理制度,科学分配教育资源 对教师和课程进行评价
教育研究者	验证现有教育理论,发现新规律 为教育实验提供数据和论据 对教学材料、课程或教学系统进行评价

EDM 与一般 LA 研究的主要区别在于采用的技术:后者多采用统计,而 EDM 多采用机器学习和数据挖掘技术.从另一角度来看,LA 侧重于描述已发生的事件或其结果,而 EDM 侧重于发现新知识与新模型[8].

1.2 EDM的发展历程

EDM 的发展大致可分为两个时期:

- 第 1 个时期是 20 世纪 80 年代~20 世纪末,研究者开始将数据挖掘技术用于教育领域,但研究方法比较简单,研究成果很少.受当时的技术水平的限制,这一时期的数据一般来自于调查问卷和信息管理软件,采用的数据挖掘技术主要是统计分析和关联规则算法.
- 第 2 个时期则是从本世纪初至今,EDM 的研究方法与研究成果快速发展.进入 21 世纪以来,互联网的普及引发了教育技术的变革,这一时期的 EDM 数据主要来自于开放和智能的在线学习系统,采用的数据挖掘技术更加多样化.2012 年,美国教育部发布的蓝皮书《通过教育数据挖掘和学习分析促进教与学》标志着 EDM 已受到广泛关注[9].

国内的EDM研究起步较晚,与国外相比在研究广度和深度上均有较大的差距[10].近10年以来,国内对EDM的研究取得了一些进展[11−13],但总体上仍存在不足,主要体现在3个方面:一是创新性不强,研究成果多为对国外

© 中国科学院软件研究所 http://www.jos.org.cn

研究的评论、跟踪和改进;二是技术深度不够,研究成果多发表在教育类期刊而非技术类期刊;三是研究范围较窄,研究成果主要集中在智能导学系统[14]和个性化学习[15]两个领域.

近几年来,教育技术领域发生了巨大的变化:一是许多新型的信息技术开始用于教育领域并取得了巨大的成功,如增强现实、移动计算和云计算技术;二是一些相对成熟的信息技术同教育结合产生了新的教学形态,如基于游戏的学习、基于社交网络的教学以及 MOOC 等.这些新的教育技术和教学形态为 EDM 的研究提供了海量数据,而大数据技术又为分析和挖掘这些数据提供了支持.可以预见,在大数据时代背景下,EDM 将更加成熟和繁荣;另一方面,随着我国对教育改革和大数据的日益重视,国内的 EDM 研究也将迎来新的转折点.

1.3 EDM的学术组织与成果总结

目前,与 EDM 最相关的两个国际学术组织分别是成立于 2011 年的 Int'l Educational Data Mining Society (http://www.educationaldatamining.org)以及成立于 2012 年的 IEEE Task Force of Educational Data Mining (http://datamining.it.uts.edu.au/edd).

与 EDM 相关的学术会议最早于 20 世纪 80 年代开始举办,目前已经有多个与 EDM 密切相关的会议(参见表 2).国际人工智能协会在 2005 年和 2006 年连续举办了两届 EDM 专业研讨会,即 AAAI workshop on Educational Data Mining.自 2008 年开始,EDM 的专业会议 Int'l Conf. on Educational Data mining 每年举办一次,截至 2014 年 7 月已经举办 7 届.刊登 EDM 研究成果的期刊数量更多,表 3 列出了与 EDM 相关的部分知名期刊.

Table 2 Related conferences about EDM

表 2 EDM 相关学术会议

会议名称	缩写	类型	首届年份
Int'l Conf. on Artificial Intelligence in Education	AIED	Biannual	1982
Int'l Conf. on Intelligence Tutoring Systems	ITS	Biannual	1988
Int'l Conf. on Educational Data mining	EDM	Annual	2008
Int'l Conf. on User Modeling, Adaption, and Personalization	UMAP	Annual	2009
Int'l Conf. on Learning Analytics and Knowledge	LAK	Annual	2011

Table 3 Related journals about EDM

表 3 EDM 相关的期刊

期刊名称	缩写	ISSN 号	影响因子(2013 年)
Journal of Engineering Education	JEE	2168-9830	2.717
Computers & Education	CAE	0360-1315	2.630
Expert System with Applications	ESWA	0957-4174	1.965
IEEE Trans. on Learning Technologies	TLE	1939-1382	1.22
Journal of Educational Data Mining	JEDM	2157-2100	–

在 EDM 发展的各个时期,均有相应的综述性论文发表[8,16−20].例如,文献[19]对 1995~2005 年的 EDM 研究成果进行了总结;文献[17]重点剖析了 2004 年~2012 年发表的 9 篇典型的 EDM 论文;文献[8]发表于 2013 年,对 EDM 的概况、技术和发展历程做了较全面的介绍.

除此之外,还可以通过两个途径更详细地了解 EDM 技术:一是 2010 年 Romero 等人编写的第一本关于 EDM 技术的专业书《Handbook of Educational Data Mining》[21],共有 36 章,详细阐述了 EDM 的概念、主要技术和典型案例;二是 2014 年 Baker 在 MOOCs 网站 Coursera(https://www.coursera.org)上开设的课程《Big Data in Education》,讲授了 EDM 的基础知识和技术.

2 EDM 的工作流程

图 2 显示了 EDM 正常的工作流程.从数据挖掘的角度来看,处理流程包含了预处理、数据挖掘和评估 3 个阶段[22];从教育的角度来看,这是一个从教育环境产生的数据中发现知识,再利用这些知识来改善教育环境的循环过程.

© 中国科学院软件研究所 http://www.jos.org.cn

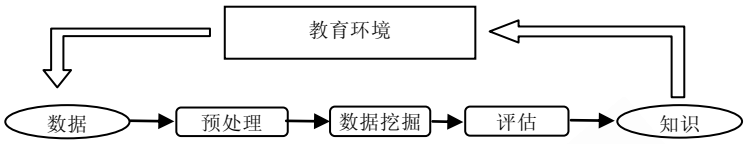

Fig.2 Flow chart of EDM[23]

图 2 EDM 流程图[23]

数据是 EDM 的研究素材.从教育环境中获取的数据通常具有多学科性、多情景和多语义等复杂特性,并且不同教育环境产生的数据也具有多样化的形态.例如,传统的教学方式产生的数据可能来自于手写的纸质文档,教务管理系统中的数据通常存储于结构化的关系数据库中,在线学习系统的数据可能记录在半结构化的日志文件中,而一些新型教育环境则涉及微博、音频和视频等非结构化数据.

知识则是 EDM 的研究结果.根据用途的不同,本文将 EDM 的知识分为以下 3 类:

- 原理类知识,其目的是验证或修正现有的教育理论,如发现新的学习规律;
- 实践类知识,其目的是帮助教师开展教学实践,如预测学生的期末成绩;
- 优化类知识,其目的是改进学习系统的效果和性能,如提高系统的自适应能力.

从图 2 可以看出,EDM 的工作流程与一般的数据挖掘应用完全相同,均要经历预处理、数据挖掘和评估这 3 个阶段.教育环境在整个流程中既是起点也是终点,并且是 EDM 研究不同于普通数据挖掘应用的一个要素.下文首先对教育环境进行说明,然后分别介绍 3 个处理阶段.

2.1 教育环境

教育环境是开展教学和学习活动的场所或载体,它可以是传统的学校和教室、互联网上的远程教育系统,也可以是安装在智能手机上的教学软件,或者是微博、微信等社交媒体.教育环境不仅是 EDM 研究的数据来源,也是其最终要改进的目标对象,因而在 EDM 研究中占有重要的地位.本文将教育环境分为 4 类:

- 传统教学环境,是指未采用或较少采用信息系统的教学环境,如中小学的教学课堂.
- 封闭式教学环境,是指以封闭式信息系统为主的教学环境,如单机版的学习软件.
- 开放式教学环境,是指以开放式信息系统为主的教学环境,如基于互联网的远程教学系统.
- 新型教学环境,是指近年来在大数据时代的背景下涌现出的新型教学场所或载体的总称,如智能手机和 MOOCs 等.

本文的第 3 节将详细介绍不同教育环境下的 EDM 研究成果.

2.2 预处理

数据挖掘算法处理的数据通常是符合一定标准的、规范的数据.而原始数据可能有多个来源,或者包含有噪音、缺失的和不一致的数据,数据挖掘算法很难直接使用这些数据.预处理,即是将原始数据转换为符合数据挖掘要求的数据格式的过程.由于数据的质量直接影响着数据挖掘的结果,预处理在数据挖掘中占有重要的地位.预处理主要包括:数据清理(data cleaning),其目标是消除数据中的噪声;数据集成(data integration),其目标是对多个数据源的数据进行合并;数据归约(data reduction),其目标是减少原数据的大小,从而提高数据挖掘的效率;数据变换(data transformation),其目标是将数值转换成数据挖掘算法需要的形式[24].

由于教育数据的复杂性,预处理通常是 EDM 研究中工作量最繁重的阶段.一些资深专家的经验表明:在 EDM 项目中,数据搜集和预处理阶段需要的人力资源往往是最多的[9].

2.3 数据挖掘

数据挖掘的目的是从数据中建立模型,主要包括预测模型(predictive model)和描述模型(descriptive model)两类.其中,预测模型通过已知的数据去预测未知的数据[25],而描述模型则通过分析数据发现新的模式或结构[26].这两类模型常见的数据挖掘方法包括:

© 中国科学院软件研究所 http://www.jos.org.cn

- 分类,其目的在于为数据对象指定一个类别,例如判断学生的性格类型.常见的分类算法有决策树[27,28]、贝叶斯理论[29]和人工神经网络[30]等.
- 回归,其目的在于为数据对象赋予一个数值,例如预测学生的 GPA 成绩.常用的是线性回归[31]和逻辑回归[32]等.
- 聚类,其目的在于将相似的数据对象归为同一类别,例如将知识点相近的试题放入同一目录下.与分类不同的是,聚类要划分的类别是未知的.常见的聚类算法有 *k*-means[33]等.
- 关联规则挖掘,其目的在于发现数据对象之间的关联或关系,例如发现学生同时选修的课程.常见的关联规则挖掘算法有 Apriori 算法[34]、散列[35]、事务压缩[36]和 FP-树频集算法[37]等.

其他方法还包括文本挖掘、马尔可夫模型、序列模式挖掘和推荐算法等.

2.4 评 估

实验数据通常会被分为 3 个部分,即训练集,用以训练模型;验证集,用以选出最优模型;测试集,用以评价模型的性能.

常见的评估分类器性能的度量有:准确率(accuracy),即全部样本中被正确识别的百分比;召回率(recall),即真实的正样本中被正确识别的百分比;精度(precision),即识别为正样本中真实的正样本所占的百分比.除了这些评估指标之外,还有一些其他指标,如 *F*-score,Kappa,AUC 等.而多类别分类器、回归、聚类和关联规划一般采用其他评估指标,受篇幅所限,在此不做描述,感兴趣者可参考文献[24].

3 EDM 的最新研究进展

教学环境不仅是EDM研究的起点和终点,也决定了数据特征和教学形态.最原始的教学环境即师生间的面对面交流,它仍是当前最主要的教学环境之一.计算机技术和通信技术引发了教育变革,产生了基于计算机的教学模式.这是一种全新的教学环境,也为各种新型教学环境的出现奠定了基础.互联网和人工智能技术在教育中的应用则创造了更加开放和智能的教学环境,它不仅增强了学生间的交流互动,也产生了更丰富的教学数据.以上几种教学环境都是依次产生和逐渐发展的,但在过去几年中,一批新型的教学环境在短时间内集中出现并呈现爆发式的增长.这也成为大数据时代背景下的教学环境的一个显著特点.

本节将介绍不同教学环境下 EDM 研究的最新进展,对每一种教学环境,将讨论 1~2 个典型的研究案例,并列出多个有代表性的研究成果,包括其数据来源、研究方法和 EDM 应用类型等内容.其中,数据来源指产生数据的系统或包含数据的记录集,研究方法是指研究采用的数据挖掘技术,而应用类型则是对 EDM 应用场景的分类.主要的 EDM 应用类型如下:

- 可视化(visualization,简称 VS)将信息或知识作形象化地展示.在 EDM 中,可视化技术能够帮助人们更加直观地理解教育数据,如用户在线论坛数据[38]、在线评估过程中产生的数据[39]、教师和学生之间的互动[40]、考试成绩[41]或者学生团体活动的相关数据[42]等.
- 学生建模(student modeling,简称 SM)通过对学生的行为、动机和学习策略等方面建立模型来揭示其学习特征.在 EDM 中,采用了贝叶斯网[43−47]、序列模式挖掘[48−50]、关联规则[51,52]和逻辑回归[53]等方法对学生特点和学习行为进行自动建模[54].
- 学生表现预测(predicting student performance,简称 PSP)通过现有数据预测学生未来的学习表现,是EDM 最早也是最流行的应用之一[55],例如根据学习记录预测学生的最终分数[56]或者学术表现[57].
- 推荐系统(recommender system,简称 RS)可以根据学生的特点向其推荐课程、学习资料或学习方法,例如根据学生的学习情况推荐合适的学习材料[58].
- 自适应系统(adaptive system,简称 AS)可以根据学生建模的结果做自适应变化的学习系统.

3.1 传统教学环境

传统教学环境,主要指师生之间面对面交流的课堂教学环境.EDM 对传统教学环境的研究在早期较为流

© 中国科学院软件研究所 http://www.jos.org.cn

行,通常采用机器学习或统计学技术对传统教育研究方法(如访谈、观察记录等)收集的数据进行分析.表 4 列出了几个属于传统教学环境的研究案例,研究数据分别来自学生成绩记录、课堂观察记录和调查问卷.由于这类数据通常以纸质文档的形式存放,在应用数据挖掘技术前,需要对原始数据进行编码、录入和格式化等预处理.

Table 4 Related researches about traditional educational environment

表 4 传统教学环境相关的研究

文献编号	数据来源	主要方法	论文要点	发表时间	学生类型	应用类型
文献[59]	对学生课堂行为的观察	决策树和回归树算法	发现教学形式与学生上课分心之间的关系	2013	小学生	SM
文献[60]	调查问卷	机器学习算法	发现视觉检测(visual inspection)的性别差异	2013	成年人	SM
文献[61]	学生 CET 4 和 CET 6 的成绩以及他们的 GPA 分数	可视化技术、关联规则、决策树算法和聚类算法	对学生 CET 4,CET6 和 GPA 成绩关系的可视化	2009	研究生	VS

尽管传统教学环境已有几千年的历史,但它仍然是校园教学的主流.因此,近年来不断有少量的研究成果出现.例如,Godwin 等人观察了 22 个班级小学生的课堂表现,并运用回归树算法对记录数据进行分析[59].研究结果表明,学生在课堂上“开小差”的原因分别是同学间的互相干扰(占 45%)、个人注意力分散(占 18%)和环境干扰(占 16%).该研究说明,EDM 技术可以使我们对一些经典的课堂现象有更深的理解.

3.2 封闭式教学环境

封闭式教学系统主要包括单机学习系统和基于 C/S 结构的信息管理系统.这类系统一般仅供内部学生和工作人员使用,且学生之间没有互动和交流.近年来,EDM 对封闭式教学环境的研究成果较少,表 5 列出了有代表性的几篇论文.这些论文的研究数据来自于教学管理信息系统和学习管理系统,其研究目的是利用数据挖掘技术,帮助学生更有效地学习或者为教育管理者提供决策支持.

Table 5 Related researches about closed educational environment

表 5 封闭式教学环境相关的研究

文献编号	数据来源	主要方法	论文要点	发表时间	学生类型	应用类型
文献[27]	某高等教育信息系统	决策树算法	发现影响学生课程成绩的要素	2014	大学生	PSP
文献[32]	土耳其中等教育过渡系统	分类,决策树,回归算法	预测学生的分班考试成绩	2012	大学生	PSP
文献[62]	学生在学习系统中的记录及其个人信息	决策树算法	根据学生特点向其推荐课程内容的学习顺序	2009	大学生	RS

封闭式教学系统经过多年运行后积累了大量数据,由于缺乏技术支持,这些数据未能得到及时的整理和分析,往往是凌乱和繁杂的.对于教育机构而言,这些数据就像未开发的“金矿”,经挖掘后可以产生较大的价值.例如,文献[27]对 106 名本科生的课程成绩进行了分析,以期通过学生的个人信息(如性别、年龄和是否全日制等)及其在各教学环节中的得分来预测该课程的最终等级.研究结果表明,采用决策树算法可以实现较高的预测准确率.该研究还发现,学生在教学活动中的表现及其笔试成绩是影响课程最终成绩的关键因素.文献[32]则是从土耳其中等教育过渡系统中抽取了 5 000 名 8 年级学生的数据(包括学生前一年的成绩和奖学金情况等),采用多种数据挖掘算法来预测学生的入学分班成绩.其中,逻辑回归模型的预测精度为 82%;人工神经网络和支持向量机分别为 89%和 91%;而 C5 决策树的预测精度最高,达到 95%.研究结果表明:借助数据挖掘技术,学校可以不开展大规模测试而对学生直接分班,从而节省教育资源.

3.3 开放式教学环境

20 世纪末,互联网的快速发展推动了网络技术在教育中的应用,远程教育课程开始流行并取得了较大的成功.本世纪初,一类新型网络教学环境开始兴起,它们一般基于 Web 技术,并采用了某种程度的人工智能技术.与

© 中国科学院软件研究所 http://www.jos.org.cn

封闭式的教学环境相比,它们的最大特点是开放性,允许学生之间互相交流和协作学习.我们把这类教学环境统称为开放式教学环境,其中,最典型的代表是智能导学系统(intelligent tutoring system,简称 ITS)和计算机支持的协作学习(computer-supported collaborative learning,简称 CSCL).

ITS 是一种智能的学习系统,提供学生交流的机会,并能提供给老师管理和记录学习情况等功能;同时, ITS 记录的数据十分丰富,包括学生的登录日志、论坛发言、作业和教学资源等,因此成为 EDM 研究最常见的数据来源之一.表 6 列出近年来基于 ITS 的一些研究成果.这些 ITS 系统中既包括时下流行的开源系统,如 MOODLE, ASSISTMent 等,也有一些仅在小范围使用的智能学习系统.

Table 6 Related EDM researches about ITS

表 6 ITS 相关的 EDM 研究

文献编号	数据来源	主要方法	论文要点	发表时间	学生类型	应用类型
文献[4]	MOODLE	神经网络和支持向量机	预测学生是否能完成在线课程	2014	大学生	PSP
文献[63]	学生在线课程记录及其 GPA	二元逻辑回归算法	预测学生能否完成在线课程	2014	大学生	PSP
文献[64]	一个智能导学系统	离散马尔可夫模型,*K*-means 聚类算法和逻辑回归分析	分析学生求助策略与学业成绩间的关系	2014	大学生	SM
文献[31]	ASSISTMent	逻辑回归和贝叶斯知识追踪	预测学生能否考上大学	2013	大学生	PSP
文献[58]	MOODLE	聚类算法和关联规则挖掘	向学生推荐课程	2013	–	RS
文献[65]	一个电子学习系统	关联规则挖掘算法	自动化地构建概念图(concept map)	2013	–	AS
文献[66]	学生的在线问答及对话记录	数据挖掘和文本挖掘	发现学生提问与成绩之间的关系	2013	大学生	SM
文献[67]	学生在线课程的参与情况	分类和聚类	根据学生使用论坛的情况预测学生的成绩	2013	大学生	PSP
文献[68]	一个智能导学系统	线性回归模型	预测学生是否出现沮丧情绪	2013	小学生	PSP
文献[69]	学习者对学习资源的评级	协同过滤和基因算法	向学生推荐学习资源	2013	–	RS
文献[30]	学生的英语在线课程学习记录及其个人信息	神经网络	根据学生的特点决定学习材料的难度	2011	大学生	AS

近几年对 ITS 的研究主要集中在对学生的表现和行为进行建模,如,文献[66]使用文本挖掘技术对 138 门在线课程中的问答和聊天记录进行分析,揭示了学生提问的次数与最终成绩之间的关系;Lara 等人则通过 MOODLE 上课程的历史学生数据建立了参考模型,利用该模型,可以预测某一个学生是否能够顺利完成课程[4]. ITS 系统同时也朝着自动化和自适应的方向发展,通过对学生的目标、偏好和知识等进行建模后,个性化地适应每个学生的学习方式.如,Wang 等人设计并实现了一个自适应的英语学习系统[30].该系统使用 5 名英语教学专家提供的样本对 BP 神经网络进行训练.正式运行时,系统可根据学生的性别、性格和学习焦虑程度向其推荐不同难度等级的词汇、语法和阅读材料.实验结果表明,采用自适应学习系统的学生成绩要明显优于对照组.Aher 等人对学生在 MOODLE 上的课程学习记录进行聚类和关联规划分析,然后向学生推荐合适的课程[58].例如,当学生完成《操作系统》课程后,向其推荐《分布式系统》课程.研究发现,结合 *K*-means 与 Apriori 算法推荐的课程与学生选课的历史数据最吻合.这些研究说明:数据挖掘技术使我们在辅导大量学生时,依然可以实现“因材施教”这一教学目标.

CSCL 是指团队成员在网络和软件的支持下,通过对话和联合行动共同完成学习任务的形式.表 7 列出了近年来对 CSCL 的部分研究成果.这些研究对不同在线学习平台的数据进行了分析,其主要研究目的是发现影响协作学习效果的因素和规律.

尽管各类学科对学生的团队协作能力都很重视,然而对该技能的教学和评估一直是个难题.Perera 等人对 7 组学生参加软件开发项目的团队表现进行了研究,数据来自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[70].该研究利用聚类技术获得了 3 类小组和 4 种成员角色在团队协作中的特征,通过序列模式挖掘,发现了优异和平庸的小组在使用 3 种工具时的差异.研究结果表明,数据挖掘技术可以帮助高校开展团队协作技能的教学与实践.它不仅能够发现学生使用团队协作工具的规律,为团队协作中的抽象概念提供案例与数据,也

© 中国科学院软件研究所 http://www.jos.org.cn

能自动识别各小组在项目协作中的问题,帮助学生监控并改进个人在小组合作中的表现.Ding 等人研究了不同性别组合的学生在求解问题时的合作模式[71],96 名中学生被随机分配到 48 个小组中,两个小组成员利用计算机进行远程通信,合作解答物理问题.对通信内容做可视化处理和多层回归分析后发现,女生与同性别同学合作的学习效果要优于与异性合作的效果,而男生则不存在这一现象.

Table 7 Related EDM researches about CSCL

表 7 CSCL 相关的 EDM 研究

文献编号	数据来源	主要方法	论文要点	发表时间	学生类型	应用类型
文献[23]	MOODLE	协同过滤和关联规则挖掘	向教师推荐用于改善教学的信息	2011	–	RS
文献[70]	一个在线协作学习工具	聚类和序列模式挖掘	发现软件开发小组成员的协作模式及其对软件质量的影响	2009	–	SM
文献[71]	一个在线学习系统	可视化和回归模型	学生分组中性别组合对学习效果的影响	2011	中学生	VS
文献[72]	在线课程论坛	统计分析	发现教师的反馈与学生成绩的关系	2014	研究生	SM

3.4 大数据时代下的新型教学环境

大数据时代见证了众多新型教学环境的诞生和飞速发展,包括基于游戏、社交网络、智能移动设备和增强现实技术的教学环境和 MOOC 等教学形态.目前,EDM 对它们的研究还较少,然而借助日渐成熟的大数据分析技术,新型教学环境正在成为 EDM 的研究热点,并反过来推动 EDM 的发展.

基于游戏的学习系统(game-based learning system,简称 GBLS)是指融合了游戏元素的学习系统,它可以给学习者带来轻松愉悦的学习氛围,激发其内在的学习动机和激情,甚至提高协作学习的效果[73].EDM 可利用 GBLS 来分析学生的性格和特征(见表 8).例如,文献[29]搜集了 47 名计算机专业的大学生在某个策略类小游戏上的尝试次数、持续时间和最终等级等数据,采用 Naïve Bayes 分类器对学生的感知类型(感觉性或直觉性)进行判断,其准确率超过 85%.与传统的方法相比,该方法的成本更小,且学生的接受度更高.

Table 8 Related researches about GBLS

表 8 GBLS 相关的研究

文献编号	数据来源	主要方法	论文要点	发表时间	学生类型	应用类型
文献[29]	一个益智类游戏的记录数据	朴素贝叶斯分类器	根据游戏记录判断学生的学习风格(learning style)	2014	大学生	PSP
文献[74]	一个大型的多人在线数学游戏	聚类	发现学生团队协作中的规律	2014	中小学生	SM

社交网络(social network,简称 SN)已成为当代学生日常生活的一部分.EDM 研究结果表明,社交网络可以帮助我们更好地了解学生(见表 9).例如,文献[75]利用社交分析技术和随机图模型对 39 名学生相互之间发送的 617 封电子邮件进行了分析,使用图(graph)来表示学生收发邮件的社交关系.研究结果表明,随着学习负担的增加,邮件的个数相应增加,图却变得更稀疏.研究中还发现,在学习负担最重的阶段,图包含的典型结构与其他阶段不同.该研究既可以向教师显示学生的学习状态,也可以让学生了解自己与同学们的交流情况.Chen 等人对 Twitter 上发表的微博进行了研究,旨在帮助大学的管理层以及相关政策的制定者了解工程专业的大学生学习和生活的真实体验[76].该研究获取了在 Twitter 上发表的标签为#EngineeringProblem 的 2 万多条微博,首先采用社会学研究中的质性分析方法对随机选取的近 3 000 个微博进行处理,将微博反映的学生体验分成 6 个类别;然后,利用文本处理技术和 Naïve Bayes 多标签分类器建立预测模型.实验结果表明,该预测模型能够达到较高的准确率;最后,研究者使用该模型对在美国普渡大学附近发表的 3 万多条微博进行了分析.研究中发现:工程专业的学生通常面临着睡眠不足、学习负担过重、缺乏社交和不适应社会多样化等问题;而普渡大学由于采取了相应措施,学生对社会多样化问题较能适应.该项研究结果表明:与传统的社会学调查方法相比,数据挖掘技术可用较小的成本完成对大规模样本的分析.

© 中国科学院软件研究所 http://www.jos.org.cn

Table 9 Related researches about social network

表 9 社交网络相关的研究

文献编号	数据来源	主要方法	论文要点	发表时间	学生类型	应用类型
文献[75]	学生的电子邮件通信	社会网络分析和指数随机图模型	发现电子邮件通信与学业负担间的关系	2014	大学生	SM
文献[76]	工科学生的 Twitter 微博	文本处理与分类算法	通过挖掘微博发现工科学生面临的主要问题	2014	大学生	SM
文献[77]	调查问卷与访谈	回归模型	发现 Facebook 的使用与学习投入程度之间的关系	2012	大学生	SM

智能移动设备凭借其优越的物理特性(可触摸、便携性、自带无线上网和多种传感器功能)和丰富的应用为学生带来了新的学习体验,已有研究结果表明:利用移动设备可以提升学生的学习兴趣[78],提高注意力[79],或者帮助学生更好地理解植物[80]和动物[81]方面的知识.

增强现实技术允许使用者在真实的物理空间上叠加虚拟对象,在教育上使用时,可以增加学生的学习动机[82],提供给学生一个更好的学习体验.其有效性已经被众多研究所证实,如在结构工程[83]、电磁学[84]和少儿阅读[85]方面都有很好的效果.还有其他诸如虚拟实验室 LabViEW[86]、虚拟学习环境[87,88]等应用.

MOOCs 是一种可在互联网上同时教授大量学生的远程教育形式,MOOCs 不对学生设限,只要通过网络申请即可学习.自 2012 年以来,MOOCs 在全球范围内取得了巨大的成功.截至 2014 年,仅 Coursera,edX 和 Udacity 这三大 MOOCs 网站的用户数就超过 1 500 万.由清华大学发布的中文 MOOC 平台“学堂在线”也受到广泛欢迎[89].MOOC 课程可以为 EDM 提供大量的研究资料.文献[90]对美国 SJSU 大学与 Udacity 联合开发的 3 门 MOOCs 课程进行了研究,每门课程均有 50 名正式学生(matriculated student)和 50 名非正式学生(主要来自合作高中和网络用户).利用逻辑回归分析建模,研究者发现:学生能否及格主要与个人的努力程度(如登陆次数、观看视频的时间以及完成的作业数量)相关,而与学生的基本特征(如性别、年龄和家庭收入)无关.研究者还发现:那些使用在线支持较多的非正式学生(尤其是高中生)不及格的概率更大,可能与他们不习惯在线学习有关.并基于以上发现提出了几个提升 MOOCs 教学质量的建议.该研究结果表明,数据挖掘技术不仅可以发现 MOOCs 课程的一些新现象,也能帮助 MOOCs 课程的创建者和实施者改善教学效果.

4 EDM 研究的分析与评价

4.1 典型案例的对比与分析

我们对表 4~表 9 列出的 26 个文献中的案例进行了对比和分析,在选择参考文献时,我们主要考虑 3 个原则:

- 及时性:所选文献均在 2009 年以后发表,其中 69%以上发表于 2013 和 2014 年.
- 重要性:所选文献主要来自 EDM 领域的重要期刊或会议.
- 创新性:所选案例在研究内容或研究方法上具有明显创新.

因此,这些案例基本能反映近年来 EDM 研究的概况.以下从多个方面对这些案例做概要性的对比和分析:

- 从学生类型来看,小学案例共有 2 个(约占 8%),中学 2 个(约占 8%),高校 16 个(约占 61%),其他类型 6 个(约占 23%).当前的 EDM 研究以高校为主,可能在于 3 个原因:一是高校有充足的资金,信息化建设相对完善;二是高校学生对信息技术的熟练程度较高;三是高校的教学体制更加灵活.随着技术的发展和普及,这些因素都在发生改变.预计未来,面向中小学生和职场人士的 EDM 研究将大幅度增加.
- 从教育环境来看,传统和封闭式教育环境的案例共有 6 个(约占 23%),开放式教育环境 15 个(约占 58%),新型教育环境 5 个(约占 19%).目前,开放式教育环境仍然是 EDM 研究的主流,因为这类环境广泛存在,可以方便地获取数据.新型教育环境刚出现不久,目前所占比例较小,未来将成为 EDM 的研究重点.
- 从应用类型来看,SM 共有 10 个(约占 38%),PSP 有 8 个(约占 31%),RS 有 4 个(约占 15%),VS 和 AS 各有 2 个.SM 和 PSP 成为 EDM 的研究热点体现了一种现代教育理念,即,有效的教学和学生培养应建立在对学生的了解与理解的基础上.然而,与传统的教育研究相比,EDM 很少涉及对教师的研究.尽管教

© 中国科学院软件研究所 http://www.jos.org.cn

师在教学中的作用也很重要,但是采集教师的数据要比学生困难得多.

- 从采用的数据挖掘技术来看(如图 3 所示),分类、聚类和回归是 EDM 研究中最常用的技术,它们同时也是数据挖掘最基本、最成熟的技术,包含在常见的数据挖掘工具箱中.关联规则、协同过滤和可视化技术也是 EDM 中的常用技术.其他技术(如文本挖掘、马尔可夫模型、序列模式挖掘等)分属不同的类别,但每一类技术出现的频率都很低,类似于“长尾分布”.

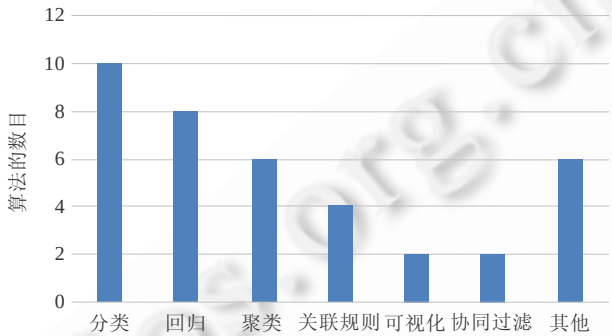

Fig.3 Distribution of data mining techniques (note that some cases employ more than one technique)

图 3 数据挖掘技术分布(注:某些案例使用了多种技术)

我们对 26 篇文献全体作者的学科背景也进行了统计(如图 4 所示).从统计结果来看,研究人员的构成具有多样性.其中,从事教育学、心理学和管理学等社会科学的研究人员比例较高.与理工研究人员相比,社会科学研究者在研究问题的提出、原始数据的理解以及研究结果的解释等方面更有优势.而来自计算机科学领域的研究者相对较少,这也解释了当前 EDM 研究主要采用成熟的数据挖掘技术这一现象.随着越来越多的计算机技术专家开展 EDM 的研究,未来很可能出现许多教育领域的专用数据挖掘技术.

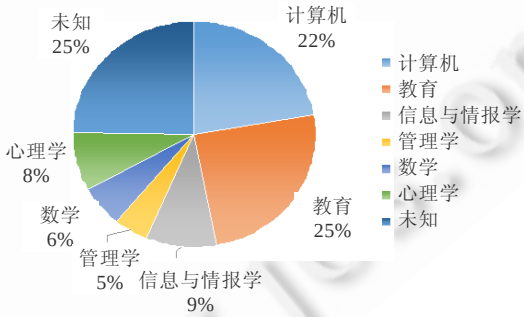

Fig.4 Distribution of researchers’ discipline background

图 4 研究者的学科背景分布

此外,我们还对各个案例涉及的学生人数进行了估计.其中,学生人数为 500 人以下的案例共有 17 个(约占 65%),人数为 500~5 000 的案例有 6 个(约占 23%),人数在 5 000 人以上的案例有 3 个(约占 12%).这说明当前的 EDM 研究涉及的海量数据还不多.然而,借助大数据处理技术,EDM 可以在短时间内对数万学生的数据完成建模、预测和可视化等复杂的操作,这是其相对于传统教育研究的巨大优势.

以上案例表明,EDM 的研究成果遍及各个教育环境、学生类型和应用类型,体现出其“从教育中来,回到教育中去”的研究理念以及“以学生为中心”的教育理念.另一方面,EDM 研究在很大程度上仍然受到数据来源和研究者技术能力的限制.

© 中国科学院软件研究所 http://www.jos.org.cn

4.2 现有研究的不足及发展趋势

EDM 研究目前仍存在许多不足.这些不足不仅有助于我们理解其研究现状,同时也为我们指出了未来的发展趋势:

- 首先是研究选题的不足.在 EDM 众多的研究类型中,PSP 和 AS 对教育的影响最大,它们有两个共同点:一是改变了我们对教育的理念与实践,二是实现了传统技术难以达到的教学效果.然而在过去 10 年中,EDM 暂未出现具有同样影响力的新的研究方向.近年来,教育和信息领域连续出现多项技术变革,极有可能孕育一批重要的 EDM 研究课题.在思考选题时,研究者应把握好教育与数据挖掘技术的关系.在 EDM 研究中,教育是其目的,而数据挖掘技术为其方法.因此,我们建议信息技术专家与教育专家深度合作,首先思考教育的本质问题,然后利用先进技术解决这些问题或发现新的规律.
- 其次是研究方法的不足,表现在两个方面:
  - 对数据预处理技术的研究较少.现有的 EDM 文献中处理的数据一般是意义清晰的最终数据集,很少对数据预处理工作进行详细描述.然而,EDM 具有多情景、多语义、存在大量噪声和数据缺失等特征,而将教育学、心理学和社会学概念与数据进行准确对应也是一项挑战.事实上,数据预处理方法对于EDM研究的重要性不亚于数据挖掘算法,在有的情况下甚至超过后者.因此,研究者应特别重视数据预处理方法的研究和论述,特别是那些具有推广价值的预处理技术.
  - 采用的数据挖掘算法相对简单.当前的 EDM 研究文献主要采用成熟的数据挖掘算法,许多研究直接采用封装好的数据挖掘工具处理数据,只有少数文献针对具体应用和场景来改进数据挖掘算法.究其原因,许多研究课题为首次提出,对算法性能的提高并非其优先考虑的问题.此外,许多研究者缺乏信息技术背景,不具备算法设计和改进的能力.因此,信息技术专家积极参与该领域的研究将有利于 EDM 的快速发展.
- 第三是数据来源的不足,表现在 3 个方面.
  - 缺少公开数据集.大多数 EDM 文献目前未将研究数据集发布在互联网上或附在论文中,研究者不愿公开数据集主要有两个原因:一是数据集涉及研究对象的隐私,按照学术道德和法律规定不适合公布;二是数据集的获取耗费大量了时间、人力和经济成本,是研究者的宝贵财富.然而对研究者而言,不公开数据集可能会降低研究成果的可信度和影响力;对 EDM 研究社区而言,公开数据集的匮乏会阻碍 EDM 研究的发展.我们建议 EDM 研究者在综合考虑隐私保护、经济投入和学术意义的基础上,共享更多的教育数据集.
  - 对新型教育环境的研究较少.现有的 EDM 研究成果对智能手机、增强现实和 MOOC 等新型教育环境的研究较少,由于这些新技术可能对教育产生深远的影响,同时又能方便地搜集大量数据,对该类型的 EDM 研究将成为未来的研究趋势.
  - 研究涉及的数据量较小.目前的 EDM 研究涉及的人数一般从几十人到几百人,少数研究涉及几千名学生,数据集大小则从几 KB 到几十 MB 不等.这些研究还称不上大数据研究.事实上,在数据搜集方面,我国高校比国外更有优势:一是中国许多高校的学生都在万人以上;二是我国高校对许多数据都进行了集中式处理,如校园卡和网络计费系统.我们期待在“教育大数据”领域,中国的研究者能走在世界前列.

5 总结与展望

本文首先描述和总结了 EDM 研究的相关背景知识,然后介绍了不同教育环境下的 EDM 研究进展,涉及研究的数据来源、研究方法、研究结果及意义和应用效果等方面.此外,对近年来的 EDM 研究成果做了对比与分析,并指出现有研究的不足及未来的发展趋势.

在过去两年中,大数据技术在舆论界、学术界和工业界均获得了前所未有的关注,这一背景为 EDM 的发展同时带来机遇与挑战.EDM 面临的机遇包括政策、资源和技术等多个方面:

© 中国科学院软件研究所 http://www.jos.org.cn

- 政策机遇:EDM 体现了“教育大数据”的理念.随着大数据技术上升为国家战略,EDM 将逐渐受到各政府部门和教育机构的重视,教师与管理人员对 EDM 的接受度也会越来越高.
- 资源支持:由于政府的重视和教育机构意识的转变,EDM 将得到更多政策、人力、资金和基础设施的支持,从而为 EDM 的发展提供必要的教育资源和研究资源.
- 技术支持:大数据技术的研究成果为 EDM 中海量数据的存储、处理和知识发现提供了方法、标准和工具,可以帮助 EDM 解决许多技术难题.

另一方面,EDM 在研究和实践中也面临着诸多挑战:

- 伦理方面的挑战:EDM 的研究过程通常涉及学生的隐私数据,其研究结果也可能对学生和教师产生不良影响.既要遵从伦理限制、保护学生隐私,又要最大化研究的学术价值,这对 EDM 的研究者是一个挑战.
- 技术方面的挑战:大数据技术有利于数据的后期处理和知识发现,然而 EDM 的工作量和难点主要集中于数据的采集、理解和预处理.为了理解数据,研究者通常需要采集一些线下的数据,这要求其掌握教育学、心理学和统计学方面的知识和技术;同时,研究者还应精通数据处理算法和工具,以提高数据预处理的效率.
- 管理方面的挑战:EDM 研究需要学生、教师和管理人员同研究者紧密配合.由于涉及的角色众多,且不同的人参与研究的动力、对项目的期望和对技术的理解有很大的差异,EDM 研究通常比普通项目更复杂,需要从整个教育机构的层面来协调人员与活动.

经过 30 多年的发展,EDM 受到越来越多研究者的关注.近年来,众多新型教学环境为 EDM 的研究提供了丰富的应用和海量的数据来源,研究成果不断涌现.在大数据时代背景下,EDM 面临着政策、资源和技术等多方面的机遇,即将迎来重大的转折.EDM 的研究有益于教育乃至整个社会的发展,我们期待它更加成熟和繁荣.

致谢 郑友杰和孟瑶为本文的完成提供了帮助,陈自郁、葛亮、赵素芬和朱郑州仔细阅读原稿并提出了建议,本文编辑和审稿专家在审阅原稿时给出了许多宝贵意见,提高了论文的质量和可读性,在此表示感谢.

References:

[1] Witten IH, Frank E. Data mining: Practical Machine Learning Tools and Techniques. 2nd ed., Morgan Kaufmann Publishers, 2005.

[2] Anjewierden A, Kolloffel B, Hulshof C. Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. In: Proc. of the Int’l Workshop on Applying Data Mining in e-Learning (ADML 2007). 2007.

[3] Cole J, Foster H. Using Moodle: Teaching with the Popular Open Source Course Management System. 2nd ed., O’Reilly Media, Inc., 2007.

[4] Lara JA, Lizcano D, Martínez MA, Pazos J, Riera T. 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. Computers & Education, 2014,72:23–36. [doi: 10.1016/j.compedu.2013.10.009]

[5] Worldwide smartphone user base hits 1 billion. 2012. http://www.cnet.com/news/worldwide-smartphone-user-base-hits-1-billion/

[6] Facebook users reach 2.2 billion, one third of the global population. 2014 (in Chinese). http://tech.qq.com/a/20140725/000288.htm

[7] Coursera. https://www.coursera.org/

[8] Romero C, Ventura S. Data mining in education. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 2013, 3(1):12–27. [doi: 10.1002/widm.1075]

[9] Bienkowski M, Feng M, Means B. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Technical Report, Washington: Office of Educational Technology, U.S. Department of Education, 2012. 1–57.

[10] Li T, Fu GS. An overall view of the educational data mining domain. Modern Educational Technology, 2010,20(10):21–25 (in Chinese with English abstract). [doi: 10.3969/j.issn.1009-8097.2010.10.004]

© 中国科学院软件研究所 http://www.jos.org.cn

[11] Wang YG, Zhang Q. MOOC: Characteristics and learning mechanism. Education Research, 2014,(9):112−120, 133 (in Chinese with English abstract).

[12] Meng WJ. Essence of network-based education: individualized and self-regulated learning supported by interactive systems with emotional communication. Education Research, 2002,(4):52−57 (in Chinese).

[13] Chang TS. Developing an institutional intelligence system: A new trend of institutional reaserach. Journal of Higher Education, 2009,30(10):49−54 (in Chinese with English abstract).

[14] Wu YW, Li S, Tian QH. Research and Implementation of mashup intelligent question-answering system. Computer Engineering, 2013,39(7):233−236, 241 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-3428.2013.07.052]

[15] Jiang YR, Han JH, Wu WM. Adaptive approach to personlized learning sequence generation. Computer Science, 2013,40(8): 204−209 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2013.08.043]

[16] Peña-Ayala A. Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 2014,41:1432−1462. [doi: 10.1016/j.eswa.2013.08.042]

[17] Mohamad SK, Tasir Z. Educational data mining: A review. Procedia — Social and Behavioral Sciences, 2013,97:320−324. [doi:10.1016/j.sbspro.2013.10.240]

[18] Baker RS, Yacef K. The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 2009,1(1):3−17.

[19] Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 2007,33(1): 135−146. [doi: 10.1016/j.eswa.2006.04.005]

[20] Borrego M, Foster MJ, Froyd JE. Systematic literature reviews in engineering education and other developing interdisciplinary fields. Journal of Engineering Education, 2014,103(1):45−76. [doi: 10.1002/jee.20038]

[21] Romero C, Ventura S, Pechenizkiy M, Baker RS. Handbook of Educational Data Mining. CRC Press, 2011.

[22] Romero C, Ventura S, De Bra P. Knowledge discovery with genetic programming for providing feedback to courseware authors. User Modeling and User-Adapted Interaction, 2004,14(5):425−464. [doi: 10.1007/s11257-004-7961-2]

[23] García E, Romero C, Ventura S, de Castro C. A collaborative educational association rule mining tool. The Internet and Higher Education, 2011,14(2):77−88. [doi: 10.1016/j.iheduc.2010.07.006]

[24] Han J, Kamber M. Data Mining: Concepts and Techniques. 3rd ed., Morgan Kaufmann Publishers, 2011.

[25] Hand DJ, Mannila H, Smyth P. Principles of Data Mining. The MIT Press, 2001.

[26] Peng Y, Kou G, Shi Y, Chen Z. A descriptive framework for the field of data mining and knowledge discovery. Int'l Journal of Information Technology & Decision Making, 2008,7(4):639−682. [doi:10.1142/S0219622008003204]

[27] Natek S, Zwilling M. Student data mining solution–knowledge management system related to higher education institutions. Expert Systems with Applications, 2014,41(14):6400−6407. [doi: 10.1016/j.eswa.2014.04.024]

[28] Quinlan JR. Simplifying decision trees. Int'l Journal of Man-Machine Studies, 1999,51:497−510. [doi: 10.1016/S0020-7373(87)80053-6]

[29] Feldman J, Monteserin A, Amandi A. Detecting students' perception style by using games. Computers & Education, 2014,71:14−22. [doi: 10.1016/j.compedu.2013.09.007]

[30] Wang YH, Liao HC. Data mining for adaptive learning in a TESL-based e-learning system. Expert Systems with Applications, 2011,38(6):6480−6485. [doi: 10.1016/j.eswa.2010.11.098]

[31] San Pedro MOZ, Baker RS, Bowers AJ, Heffernan NT. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In: Proc. of the 6th Int'l Conf. on Educational Data Mining. 2013. 177−184.

[32] Şen B, Uçar E, Delen D. Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications, 2012,39(10):9468−9476. [doi: 10.1016/j.eswa.2012.02.112]

[33] Hartigan JA, Wong MA. Algorithm AS 136: A *k*-means clustering algorithm. Journal of the Royal Statistical Society, Series C (Applied Statistics), 1979,28(1):100−108. [doi: 10.2307/2346830]

[34] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Bocca BJ, Jarke M, Zaniolo C, eds. Proc. of the 20th Int'l Conf. on Very Large Data Bases (VLDB'94). San Francisco: Morgan Kaufmann Publishers, 1994. 487−499.

[35] Park JS, Chen MS, Yu PS. Efficient parallel data mining for association rules. In: Pissinou N, Silberschatz A, Park EK, Makki K, eds. Proc. of the 4th Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 1995. 31−36. [doi: 10.1145/221270.221320]

© 中国科学院软件研究所 http://www.jos.org.cn

[36] Han J, Fu Y. Discovery of multiple-level association rules from large databases. In: Dayal U, Gray PMD, Nishio S, eds. Proc. of the 21st Int'l Conf. of Very Large Databases (VLDB'95). San Francisco: Morgan Kaufmann Publishers, 1995. 420−431.

[37] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: Dunham M, Naughton JF, Chen WD, Koudas N, eds. Proc. of the 2000 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2000). New York: ACM Press, 2000. 1−12. [doi: 10.1145/342009.335372]

[38] Burr L, Spennemann DH. Patterns of user behaviour in university online forums. Int'l Journal of Instructional Technology and Distance Learning, 2004,1(10):11−28.

[39] Pechenizkiy M, Trcka N, Vasilyeva E, van der Aalst W, De Bra P. Process mining online assessment data. In: Proc. of the Int'l Working Group on Educational Data Mining. 2009. 279−288.

[40] Mostow J, Beck J, Cen H, Cuneo A, Gouvea E, Heiner C. An educational data mining tool to browse tutor-student interactions: Time will tell. In: Proc. of the Workshop on Educational Data Mining, National Conf. on Artificial Intelligence. 2005. 15−22.

[41] Shen R, Yang F, Han P. Data analysis center based on e-learning platform. In: Hommel G, Huanye S, eds. Proc. of the Internet Challenge: Technology and Applications. Springer-Verlag, 2002. 19−28. [doi: 10.1007/978-94-010-0494-7_3]

[42] Juan AA, Daradoumis T, Faulin J, Xhafa F. SAMOS: A model for monitoring students' and groups' activities in collaborative e-learning. Int'l Journal of Learning Technology, 2009,4(1):53−72. [doi: 10.1504/IJLT.2009.024716]

[43] Baker RS, Corbett AT, Aleven V. Improving contextual models of guessing and slipping with a truncated training set. In: Proc. of the Educational Data Mining 2008. 2008. 67−76.

[44] García P, Amandi A, Schiaffino S, Campo M. Evaluating Bayesian networks' precision for detecting students' learning styles. Computers & Education, 2007,49(3):794−808. [doi: 10.1016/j.compedu.2005.11.017]

[45] Jonsson A, Johns J, Mehranian H, Arroyo I, Woolf B, Barto A, Fisher D, Mahadevan S. Evaluating the feasibility of learning student models from data. In: Proc. of the Educational Data Mining: Papers from the AAAI Workshop. 2005. 1−6.

[46] Chang KM, Beck J, Mostow J, Corbett A. A Bayes net toolkit for student modeling in intelligent tutoring systems. In: Ikeda M, Ashley KD, Chan TW, eds. Proc. of the 8th Intelligent Tutoring Systems. Springer-Verlag, 2006. 104−113. [doi: 10.1007/11774303_11]

[47] Arroyo I, Murray T, Woolf BP, Beal C. Inferring unobservable learning variables from students' help seeking behavior. In: Lester JC, Vicari RM, Paraguacu F, eds. Proc. of the Intelligent Tutoring Systems. Springer-Verlag, 2004. 782−784. [doi: 10.1007/978-3-540-30139-4_74]

[48] Antunes C. Acquiring background knowledge for intelligent tutoring systems. In: Proc. of the EDM. 2008. 18−27.

[49] Andrejko A, Barla M, Bieliková M, Tvarozek M. User characteristics acquisition from logs with semantics. In: Proc. of the Int'l Conf. on Information System Implementation and Modeling. 2007. 103−110.

[50] Robinet V, Bisson G, Gordon M, Lemaire B. Searching for student intermediate mental steps. In: Proc. of the 11th Int'l Conf. on User Modeling. 2007. 35−39.

[51] Huang J, Zhu A, Luo Q. Personality mining method in Web based education system using data mining. In: Proc. of the IEEE Int'l Conf. on Grey Systems and Intelligent Services 2007 (GSIS 2007). IEEE, 2007. 155−158. [doi: 10.1109/GSIS.2007.4443256]

[52] Matsuda N, Cohen WW, Sewall J, Lacerda G, Koedinger KR. Predicting students' performance with simstudent: learning cognitive skills from observation. In: Luckin R, Koedinger KR, Greer J, eds. Proc. of the 2007 Conf. on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work. Amsterdam: IOS Press, 2007. 467−476.

[53] Feng M, Beck J. Back to the future: A non-automated method of constructing transfer models. In: Barnes T, Desmarais M, Romero C, Ventura S, eds. Proc. of the Int'l Working Group on Educational Data Mining, Spain, 2009. 240−248.

[54] Frias-Martinez E, Chen SY, Liu X. Survey of data mining approaches to user modeling for adaptive hypermedia. IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2006,36(6):734−749. [doi: 10.1109/TSMCC.2006.879391]

[55] Romero C, Ventura S. Educational data mining: A review of the state of the art. IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2010,40(6):601−618. [doi: 10.1109/TSMCC.2010.2053532]

[56] Romero C, Ventura S, Espejo PG, Hervás C. Data mining algorithms to classify students. In: Proc. of the EDM. 2008. 8−17.

[57] Minaei-Bidgoli B, Kashy DA, Kortemeyer G, Punch WF. Predicting student performance: An application of data mining methods with an educational Web-based system. In: Proc. of the 33rd Annual Frontiers in Education 2003 (FIE 2003). IEEE, 2003. T2A-13. [doi: 10.1109/FIE.2003.1263284]

© 中国科学院软件研究所 http://www.jos.org.cn

[58] Aher SB, Lobo LMRJ. Combination of machine learning algorithms for recommendation of courses in e-learning system based on historical data. Knowledge-Based Systems, 2013,51:1−14. [doi: 10.1016/j.knosys.2013.04.015]

[59] Godwin KE, Almeda MV, Petroccia M, Baker RS, Fisher AV. Classroom activities and off-task behavior in elementary school children. In: Proc. of the Cognitive Science Society. 2013. 2428−2433.

[60] Heidl W, Thumfart S, Lughofer E, Eitzinger C, Klement EP. Machine learning based analysis of gender differences in visual inspection decision making. Information Sciences, 2013,224:62−76. [doi: 10.1016/j.ins.2012.09.054]

[61] Jin H, Wu T, Liu Z, Yan J. Application of visual data mining in higher-education evaluation system. In: Proc. of the 2009 1st Int'l Workshop on Education Technology and Computer Science. 2009. 101−104. [doi: 10.1109/ETCS.2009.285]

[62] Wang YH, Tseng MH, Liao HC. Data mining for adaptive learning sequence in English language instruction. Expert Systems with Applications, 2009,36(4):7681−7686. [doi: 10.1016/j.eswa.2008.09.008]

[63] Hachey AC, Wladis CW, Conway KM. Do prior online course outcomes provide more information than G.P.A. alone in predicting subsequent online course grades and retention? An observational study at an urban community college. Computers & Education, 2014,72:59−67. [doi: 10.1016/j.compedu.2013.10.012]

[64] Vaessen BE, Prins FJ, Jeuring J. University students' achievement goals and help-seeking strategies in an intelligent tutoring system. Computers & Education, 2014,72:196−208. [doi: 10.1016/j.compedu.2013.11.001]

[65] Chen SM, Sue PJ. Constructing concept maps for adaptive learning systems based on data mining techniques. Expert Systems with Applications, 2013,40(7):2746−2755. [doi: 10.1016/j.eswa.2012.11.018]

[66] He W. Examining students' online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior, 2013,29(1):90−102. [doi: 10.1016/j.chb.2012.07.020]

[67] Romero C, Lopez MI, Luna JM, Ventura S. Predicting students' final performance from participation in on-line discussion forums. Computers & Education, 2013,68:458−472. [doi: 10.1016/j.compedu.2013.06.009]

[68] Rajendran R, Iyer S, Murthy S, Wilson C, Sheard J. A theory-driven approach to predict frustration in an ITS. IEEE Trans. on Learning Technologies, 2013,6(4):378−388. [doi: 10.1109/TLT.2013.31]

[69] Salehi M, Kamalabadi IN, Ghoushchi MBG. An effective recommendation framework for personal learning environments using a learner preference tree and a GA. IEEE Trans. on Learning Technologies, 2013,6(4):350−363. [doi: 10.1109/TLT.2013.28]

[70] Perera D, Kay J, Koprinska I, Yacef K, Zaiane OR. Clustering and sequential pattern mining of online collaborative learning data. IEEE Trans. on Knowledge and Data Engineering, 2009,21(6):759−772. [doi: 10.1109/TKDE.2008.138]

[71] Ding N, Bosker RJ, Harskamp EG. Exploring gender and gender pairing in the knowledge elaboration processes of students using computer-supported collaborative learning. Computers & Education, 2011,56(2):325−336. [doi: 10.1016/j.compedu.2010.06.004]

[72] Coll C, Rochera MJ, de Gispert I. Supporting online collaborative learning in small groups: Teacher feedback on learning content, academic task and social participation. Computers & Education, 2014,75:53−64. [doi: 10.1016/j.compedu.2014.01.015]

[73] Li Q, Lau RW, Shih TK, Li FW. Technology supports for distributed and collaborative learning over the internet. ACM Trans. on Internet Technology, 2008,8(2):1−24. [doi: 10.1145/1323651.1323656]

[74] Araya R, Jiménez A, Bahamondez M, Calfucura P, Dartnell P, Soto-Andrade J. Teaching modeling skills using a massively multiplayer online mathematics game. World Wide Web, 2014,17(2):213−227. [doi: 10.1007/s11280-012-0173-5]

[75] Uddin S, Thompson K, Schwendimann B, Piraveenan M. The impact of study load on the dynamics of longitudinal email communications among students. Computers & Education, 2014,72:209−219. [doi: 10.1016/j.compedu.2013.11.007]

[76] Chen X, Vorvoreanu M, Madhavan KPC. Mining social media data for understanding students' learning experiences. IEEE Trans. on Learning Technologies, 2014,7(3):246−259. [doi: 10.1109/TLT.2013.2296520]

[77] Junco R. The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. Computers & Education, 2012,58(1):162−171. [doi: 10.1016/j.compedu.2011.08.004]

[78] Hong JC, Hwang MY, Liu MC, Ho HY, Chen YL. Using a "prediction-observation-explanation" inquiry model to enhance student interest and intention to continue science learning predicted by their Internet cognitive failure. Computers & Education, 2014,72:110−120. [doi: 10.1016/j.compedu.2013.10.004]

[79] Sun JCY. Influence of polling technologies on student engagement: An analysis of student motivation, academic performance, and brainwave data. Computers & Education, 2014,72:80−89. [doi: 10.1016/j.compedu.2013.10.010]

[80] Liu TC, Lin YC, Paas F. Effects of prior knowledge on learning from different compositions of representations in a mobile learning environment. Computers & Education, 2014,72:328−338. [doi: 10.1016/j.compedu.2013.10.019]

© 中国科学院软件研究所 http://www.jos.org.cn

[81] Song Y. “Bring your own device (BYOD)” for seamless science inquiry in a primary school. Computers & Education, 2014,74:50−60. [doi: 10.1016/j.compedu.2014.01.005]

[82] Chang MM, Lin MC. The effect of reflective learning e-journals on reading comprehension and communication in language learning. Computers & Education, 2014,71:124−132. [doi: 10.1016/j.compedu.2013.09.023]

[83] Behzadan AH, Kamat VR. Enabling discovery-based learning in construction using telepresent augmented reality. Automation in Construction, 2013,33:3−10. [doi: 10.1016/j.autcon.2012.09.003]

[84] Ibáñez MB, Di Serio Á, Villarán D, Delgado Kloos C. Experimenting with electromagnetism using augmented reality: Impact on flow student experience and educational effectiveness. Computers & Education, 2014,71:1−13. [doi: 10.1016/j.compedu.2013.09.004]

[85] Cheng KH, Tsai CC. Children and parents’ reading of an augmented reality picture book: Analyses of behavioral patterns and cognitive attainment. Computers & Education, 2014,72:302−312. [doi: 10.1016/j.compedu.2013.12.003]

[86] Wei Z, Porter JR, Morgan JA. Experiential learning of digital communication using LabVIEW. IEEE Trans. on Education, 2014, 57(1):34−41. [doi: 10.1109/TE.2013.2264059]

[87] Pedersen S, Irby T. The VELscience project: Middle schoolers’ engagement in student-directed inquiry within a virtual environment for learning. Computers & Education, 2014,71:33−42. [doi: 10.1016/j.compedu.2013.09.006]

[88] Myneni LS, Narayanan NH, Rebello S, Rouinfar A, Pumtambekar S. An interactive and intelligent learning system for physics education. IEEE Trans. on Learning Technologies, 2013,6(3):228−239. [doi: 10.1109/TLT.2013.26]

[89] Kang YQ. An analysis on SPoC: Post—MooC era of online education. Tsinghua Journal of Education, 2014,35(1):85−93 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-4519.2014.01.010]

[90] Firmin R, Schiorring E, Whitmer J, Willett T, Collins ED, Sujitparapitaya S. Case study: Using MOOCs for conventional college coursework. Distance Education, 2014,35(2):178−201. [doi: 10.1080/01587919.2014.917707]

附中文参考文献:

[6] Facebook 用户总数达到 22 亿人,占全球总人口 1/3. http://tech.qq.com/a/20140725/000288.htm

[10] 李婷,傅钢善.国内外教育数据挖掘研究现状及趋势分析.现代教育技术,2010,20(10):21−25.

[11] 王永固,张庆.MOOC:特征与学习机制.教育研究,2014,(9):112−120,133.

[12] 孟万金.网络教育的真谛:人文交互环境下的个性化自主学习.教育研究,2002,(4):52−57.

[13] 常桐善.构建院校智能体系:院校研究发展的新趋势.高等教育研究,2009,30(10):49−54.

[14] 吴彦文,李诗,田庆恒.Mashup 智能答疑系统的研究与实现.计算机工程,2013,39(7):233−236+241. [doi: 10.3969/j.issn.1000-3428.2013.07.052]

[15] 蒋艳荣,韩坚华,吴伟民.一种自适应的个性化学习序列生成研究.计算机科学,2013,40(8):204−209. [doi: 10.3969/j.issn.1002-137X.2013.08.043]

[89] 康叶钦.在线教育的“后 MOOC 时代”——SPOC 解析.清华大学教育研究,2014,35(1):85−93. [doi: 10.3969/j.issn.1001-4519.2014.01.010]

周庆(1979－),男,重庆人,博士,教授,博士生导师,CCF 会员,主要研究领域为人工智能,数据挖掘技术.

杨丹(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为科学与工程计算,软件工程及应用.

牟超(1989－),男,博士生,CCF 学生会员,主要研究领域为数据挖掘.

© 中国科学院软件研究所 http://www.jos.org.cn