DOI:10.13811/j.cnki.eer.2017.10.005

面向 MOOC 课程评论的学习者话题挖掘研究

刘三女牙，彭 晛，刘 智，孙建文，刘 海

（华中师范大学 国家数字化学习工程技术研究中心，湖北 武汉 430079）

[摘 要] 研究以果壳网 MOOC 学院的“财务分析与决策”课程为实验对象，通过分析课程评论帖进行学习者话题的挖掘。文章不仅采用了高频词汇分析的定量方法，实现对学习者课程评论内容的整体认识，并且，根据参与评论学习者的课程完成情况，分别对已完成和未完成两种类型的学习者展开定性的学习分析研究，应用非监督学习方法 LDA 模型自动挖掘和解析文本评论信息的特征结构和语义内容，并探究和追踪学习者关注的热点话题演化趋势。实验结果表明，学习者认可和赞赏了该门课程，并且尤为关注课程内容以及教师授课形式话题；相比课程完成者，未完成者更倾向于解释其未完成课程的主要原因，表达出更为消极的话题内容，并较少涉及课程本身相关的专业理论知识。

[关键词] MOOC；文本评论；话题挖掘；LDA

[中图分类号] G434 [文献标志码] A

[作者简介] 刘三女牙（1973—），男，安徽桐城人。教授，主要从事计算机应用、人工智能、教育信息技术等方面的研究。E-mail：lsy5918@mail.ccnu.edu.cn。

一、引 言

“互联网+教育”的盛行已引起了国内外教育学府、科研机构、商业公司等的关注，作为其最重要的应用创新产物之一——MOOC（Massive Open Online Courses，大规模网络在线课程）学习平台，正成为变革教学思想、教学设计过程和学习模式的一把利刃。为适应数据驱动教学的实际需求，学习分析技术应运而生。当前，该技术较侧重于外显行为的分析，如在线学习平台中学习者的登录/退出次数、发布/回复评论次数、提交/未提交作业次数等数据[1]。通过对这些结构化行为数据的分析，学习分析应用于 MOOC 平台的研究主要包括学习者在平台中的参与度[2-3]、学习行为方式分类及其与学习成效之间的关系[4-6]、学习效果预测[7-10]等。

随着 MOOC 互动学习场景的多样化发展，学习者在学习过程中已生成了越来越多的非结构化的交互式文本数据，其主要源于课程点评区、讨论区、实时答疑室、同伴互评等多个学习场景，这为理解和优化学习者学习过程与学习情境提供了大量线索。文本作为在线教育中重要的互动载体，可真实地反映出学习者的兴趣话题、情感态度、学习体验等特征。通过对交互式文本数据的挖掘，有助于提取学习者在互动学习中隐藏的内在含义，并实时评估和追踪学习者状态和认知心理。相比定量地探究学习者行为方式及课程成效的研究，定性地挖掘和解析文本语义内容也尤为重要。目前部分研究者已经展开了对 MOOC 平台中文本话语行为和内容的分析，开放大学 Ferguson 等人[11]构建了一套基于标注特征的训练模型来自动化探究话语的类型：探究性对话和非探究性对话，完成对论坛中话语的二元分类；北卡罗莱纳州立大学的 Ezen-Can 研究团队[12-13]运用聚类方法自动化识别系统平台中发表的文本数据结构，理解学习者话语的交互内容和行为方式；马里兰大学 Ramesh 等人[14]提出了一种基于种子词的话题模型方法来挖掘 MOOC 平台中学习者的话语内容，旨在帮助预测其课程通过率；卡耐基梅隆大学的 Wen 等人[15]通过分析 Coursera 平台中的讨论帖，采

基金项目：2016 年教育部—中国移动科研基金项目“国家教育大数据相关问题研究”（项目编号：MCM20160401）；2016 年度教育部人文社会科学研究青年基金项目“高校慕课环境下的互动话语行为及其对学习效果的影响机理研究”（项目编号：16YJC880052）

用情感分析技术来监测学习集体在课程中的情感演化趋势,并发现其情感比率与退课率有显著的关联。

为快速、深入地分析和理解在线课程评论区中学习共同体交互的文本表达内容,并探究不同学习者类型之间(课程已完成/未完成)的话题及细粒度单词的分布,本文提出采用 LDA(Latent Dirichlet Allocation)话题模型来实现评论文本的自动建模,并追踪热点话题演化的趋势。LDA 是机器学习、自然语言处理领域中的一种非监督学习方法,作为在线教育中支撑数据驱动的一种新的论证方法,突破了教学互动话语分析研究的经验式判断和过度人为主观干预的局限[16-17],能够为学习分析研究提供新的思路,以期为教学实践指导和学习体验优化提供数据支撑服务。

二、研究方法

(一)问题描述

针对学习者话语内容的多样性特点,本文定义学习者评论的文本集合 $L=\{r_1, r_2, \cdots\cdots, r_m\}(1\leqslant m\leqslant M)$以及相应的话题集合为 $L=\{z_1, z_2, \cdots\cdots, z_k\}(1\leqslant k\leqslant K)$(以话题概率大小进行降序排列),就某个具体的话题而言,定义其单词集合 $z_k=\{w_1, w_2, \cdots\cdots, w_n\}(z_k \in L)$(以单词概率大小进行降序排列)。本文假设每个单词都属于某一个具体的话题,则所有的课程评论集合可由高维单词向量空间构成 $R=\{w_{1k}, w_{2k}, \cdots\cdots, w_{nk}\}(1\leqslant k\leqslant K)$,$w_{nk}$ 表示属于第 k 个话题的某个单词,当 k 等于某一固定值时,对课程评论集中的所有单词进行多次迭代采样直到稳定,挖掘学习者关注的话题以及这些话题的具体细粒度单词信息。

(二)LDA 模型

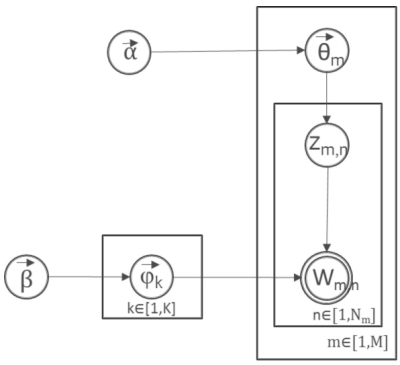

图 1　LDA 模型的结构图

LDA 作为机器学习领域的一种非监督话题建模方法[18],是基于 PLSI (Probabilistic Latent Semantic Indexing)模型的延伸,最早由 Blei 在 2003 年提出,旨在发现大规模文档集中隐藏的话题结构和内容的算法,如今已被广泛应用到多个领域,尤其在商务智能方面,如新闻智能定制服务[19]、电影个性化推荐[20]、社交网络互动[21]等。LDA 是一个由文档、话题、单词构成的三层贝叶斯概率图模型,如图 1 所示。

它的主要思想可概括为:一篇文档通常由多个话题构成,而每一个话题由服从多项分布的单词组成。模型的生成过程可理解为包括观测随机变量(文档中的单词)和隐藏随机变量(话题结构)的联合概率分布,在已知可观察变量的前提下来计算隐藏变量的后验分布,其大致描述如图 2 所示。

输入:学习者的文本评论集合 $L=\{r_1, r_2, \cdots, r_m\}(1\leq m\leq M)$;
LDA 模型迭代的次数:500;先验参数α 和 β分别为 0.2 和 0.1。
输出:学习者评论文档-话题矩阵θ_m和话题-单词矩阵φ_k。
步骤 1:初始化模型的相关参数 K、α和 β等。
步骤 2:对于所有文档中的每一个话题 z_k (k=1, … , K),采样该话题下服从多项分布的单词:$\varphi_k \sim Dir(\beta)$。
步骤 3:其次,当每次生成一篇新文档时,重复如下过程生成该文档的每一个单词 $w_{m,n}(n=1,\ldots,N)$
a. 随机的从文档-话题分布中选择一个话题 z_k,$z_{d,n} \sim Multi(\theta_d)$;
b. 随机的从话题-单词中选择一个单词 $w_{m,n}$,$w_{m,n} \sim Multi(\varphi_{z_{m,n}})$。
步骤 4:在每次迭代过程中,采用 Gibbs Sampling 算法赋予每个单词新的主题标签,并更新模型的相关参数,然后直到模型达到最优化输出最终结果θ_m和φ_k。

图 2　LDA 模型的生成求解过程

图 1 表示第 m 篇文档中的第 n 个单词所对应的话题,α 和 β 是模型的两个先验参数,K 表示文档中所有的主题数。根据 LDA 概率图中变量的依赖关系,构建联合概率分布,其形式化描述如下所示:

$$p(w_{m,n}, z_{m,n}, \theta_m, \varphi_k \mid \alpha, \beta) = \prod_{m=1}^{N_m} p(w_{m,n} \mid \varphi_{z_{m,n}}) p(z_{m,n} \mid \theta_m) p(\theta_m \mid \alpha) p(\varphi_k \mid \beta)$$

$$= \prod_{z=1}^{K} \frac{\Delta(n_z + \beta)}{\Delta\beta_l} \cdot \prod_{m=1}^{M} \frac{\Delta(n_m + \alpha)}{\Delta\alpha}$$

根据以上公式即可计算出最终求解的两个参数,分别为文档—话题概率分布矩阵 θ_m,即学习者发表的课程评论可抽象表示为多个话题直方图(根据概率大小排序),和话题—单词概率分布矩阵 φ_k,即学习者评论的话题可表征为多维细粒度的单词内容(根据概率大小排序)。这将有助于理解学习者评论的大量课程话题内容并为其提供自动化支持[22]。

三、实验设计

(一)实验设计框架图

本模块旨在系统地描述 MOOC 环境下学习者隐含评论话题的挖掘过程,构建一幅服务于管理者、教学者、学习者对象的设计框架图,形成自适应性的闭环,从而支持管理者决策的制定,指导教学者实践教学的开展,提升学习者的学习体验。如图 3 所示,首先,通过收集 MOOC 平台中的课程评论文本集,对其进行数据清洗和筛选;然后,结合定量与定性的学习分析方法,实现学习者评论文本的话题挖掘;最后,为管理者、教学者、学习者直观呈现分析的结果,并完成适应性反馈与干预。

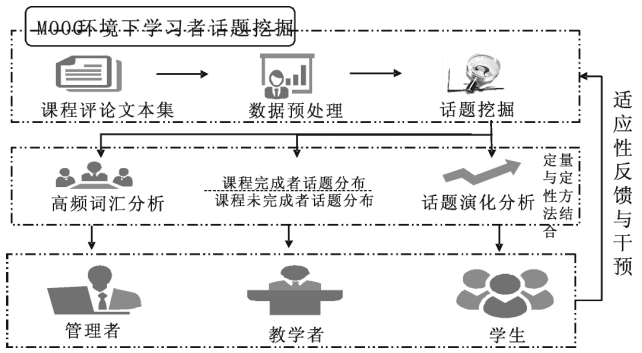

图 3　MOOC 环境下学习者话题建构框架图

(二)实验数据及预处理

本研究的数据来源于果壳网 MOOC 学院“财务分析与决策”课程的评论区(mooc.guokr.com)。果壳网 MOOC 学院是当前融合我国教学特色的在线开放课程平台的应用示范之一,“财务分析与决策”无疑是 MOOC 学院的在线课程代表,在线课程学习人数已达到 2565 人,有着丰富的交互数据,从 2013 年 10 月到 2016 年 3 月之间,767 名在线学习者发表了评论文本,通过网络爬虫方法一共获取到 767 条文本点评数据,每个参与者平均发表 1 条评论。所有参与发表课程评论的学习者中,课程完成者共计 573 人,课程未完成者共计 12 人(由于课程未完成者的评论数较少,对实验的效果可能造成一定的影响),正在上课者共计 182 人(处于上课状态的学习者,无法判断课程完成情况)。如表 1 所示,已完成课程者的评论字符平均数明显高于未完成课程者,表明两者在课程评论参与方面存在一定的差异性。

表 1　　课程评论数据集基本信息

评论课程的学习者类型	评论发表数	课程评论字符平均数	比率	课程总参与人数/课程评论人数
已完成	573	76.77	99.80%	2565/767
未完成	12	57.17	0.02%	
在上课	182			

在收集原始数据的基础之上,对这些数据进行预处理,首先,利用中科院中文分词系统 ICTCLAS 对其进行分词[23];然后,针对课程评论中的专业术语和特殊网络词汇,建立用户词典进行强制约束,以获得更为完整的语义信息词汇,例如:“财务分析”“财务报表”“价值创造”“学神”“果壳”等;最后,剔除停用词、噪音词、低频词和特殊符号等,仅保留形容词、动词和名词三种便于理解话题语义的关键词性。

四、实验分析和结果

为更深入地挖掘和解析学习者文本评论数据中隐含的话题信息,本部分首先采用定量分析的方法统计所有课程评论内容中的高频词汇,实现对其整体的观测和认识。然后对文本评论内容展开定性分析,运用 LDA 话题模型自动挖掘不同学习者类型之间话题分布的特征结构和语义内容,探究其相似性和差异性,试图构建学习者类型和话题空间分布的映射关系,并直观呈现热点关注话题的演化态势,为进一步调整教学方法、改善在线学习体验服务提供参考依据。

(一)高频词汇分析

在面向所有参与发表课程评论的学习者对象的基础上,共产生 767 条评论,本模块采用词频分析法[24-25]来捕捉和描述学习者的热点评论内容,共获得 2376 个关键词汇,经过筛选提取出 20 个高频关键词,见表 2。

表 2　　词频统计

编号	词汇	词频	编号	词汇	词频
1	肖老师	258	11	财务分析	88
2	会计	252	12	作业	72
3	好课	236	13	企业	64
4	财务	227	14	考试	61
5	基础	181	15	财务报表	60
6	内容	149	16	财务知识	56
7	证书	139	17	生动	50
8	理解	135	18	案例	45
9	专业	114	19	公司	40
10	简单	114	20	决策	40

从表 2 左侧可发现,频次最高的关键词汇是“肖老师”,肖老师是该门在线课程的讲师,其次是“会计”“好课”“财务”“内容”等。这些词汇一定程度上能够表明,学习者不仅表现出对教师本人的认可,而且对课程的内容和教师讲解方式给予了肯定和支持。值得一提的是,“证书”成为排名第七的高频词汇,较强地反应出学习者对在线课程认证的讨论热度和追求。此外,从表 2 右侧可知,大部分词汇都是与本课程紧密相关的基础知识和专业术语。

(二)不同学习者类型话题挖掘

在依据参与评论学习者的课程完成状态(已完成/未完成)基础上,本模块旨在应用 LDA 话题模型来挖掘学习者文本表述中隐含的话题结构和语义内容,观察和对比不同学习者类型之间评论文本的话题分布,为进一步实行针对性的干预和反馈提供数据支撑。本研究中实验效果的衡量主要采用话题间的分离度和话题内部信息的一致性两个指标[26-27]。经过多次反复试验,当模型中先验参数 α 和 β 分别为 0.2 和 0.1,话题数 K 等于 10 时,表 3a 实验效果达到最佳;

表 3a　　课程已完成者的话题—单词矩阵

话题 1(0.043)	话题 2(0.384)	话题 4 (0.039)	话题 6(0.366)	话题 9(0.034)
单词　概率值	单词　概率值	单词　概率值	单词　概率值	单词　概率值
证书 (0.059)	肖老师(0.073)	财务 (0.039)	好课 (0.073)	公司 (0.022)
考试 (0.029)	易懂 (0.040)	基础 (0.038)	肖老师 (0.051)	决策 (0.020)
时间 (0.026)	通俗 (0.026)	知识 (0.022)	简单 (0.032)	表　(0.019)
知识 (0.025)	讲解 (0.018)	推荐 (0.021)	内容 (0.024)	资产 (0.016)
作业 (0.021)	赞　(0.018)	报表 (0.019)	难度 (0.024)	利润 (0.010)
完成 (0.018)	不错 (0.016)	专业 (0.017)	棒 (0.020)	银行 (0.009)
内容 (0.017)	生动 (0.015)	会计 (0.016)	案例 (0.018)	指标 (0.008)
MOOC(0.014)	会计 (0.015)	财务分析 (0.013)	有意思(0.014)	亏损 (0.008)
晒　(0.012)	有趣 (0.015)	学生 (0.013)	容易 (0.012)	负债表(0.008)
认真 (0.011)	清楚 (0.015)	深入 (0.012)	设计 (0.012)	联系 (0.008)

话题数 K 等于 5 时,表 3b 实验效果达到最佳。表 3a 为课程已完成者的话题—单词矩阵,首先根据学习者关注话题的概率分布，选取出 5 个概率值显著的话题，然后按照概率值大小列举出每个话题下的 10 个单词,即某个评论话题下具体的细粒度语义内容。表 3b 为课程未完成者的话题—单词矩阵。

表 3b　　课程未完成者的话题—单词矩阵

话题 1　(0.203)	话题 3　(0.274)	话题 5　(0.261)
单词　概率值	单词　概率值	单词　概率值
错过　(0.061)	同学　(0.139)	视频　(0.048)
证书　(0.038)	推荐　(0.093)	注册　(0.033)
好课　(0.037)	大学　(0.089)	用户　(0.033)
肖老师 (0.037)	选修　(0.089)	报名　(0.033)
晚　(0.037)	范围　(0.089)	上课　(0.033)
考试　(0.037)	困难　(0.089)	打开　(0.033)
互评　(0.037)	清晰　(0.023)	网页　(0.033)
遗憾　(0.037)	讲解　(0.023)	请问　(0.033)
达到　(0.037)	兴趣　(0.023)	指教　(0.033)
考核　(0.037)	易懂　(0.023)	制作　(0.030)

由表 3a 可见，课程已完成者评论聚焦程度最高的是话题 2(0.384)和话题 6(0.366),占据整个话题比重的 70%以上,可以反映出这部分学习者整体上最热衷的评论话题内容。从话题 2 中的单词概率分布情况,可推测出此话题是有关授课教师的讲课方式。学习者不仅对肖老师的讲课风格表示认同和赞赏,并且认为肖老师讲课通俗易懂,非常生动、有趣、清晰。从话题 6 可知,此话题与课程内容相关,学习者不仅对该门课程表达出青睐之感,认为其是好课、棒,并从课程内容的难度、案例、设计等方面进行了评论。而其他三个话题概率值明显低于话题 2 和话题 6,话题 1 主要与课程考核相关,关注知识内容储备、作业完成情况、考试时间安排和晒证书,这比较符合学习者常规的在线课程学习路线,只有顺利完成作业和达到考试要求,才能获得证书。话题 4 更多地与课程推荐相关,他们指出“财务分析与决策”是一门值得推荐的课程,尤其适合会计和财务相关专业的学生。话题 9 涉及财务信息架构体系的基础知识和常用术语,借助财务信息(表、利润、指标、亏损等),理解影响价值创造的各种因素,帮助公司进行商业决策。其他话题(3、5、7、8、10),主要从理论知识的实际应用、平台资源建设和制作、财务管理和决策等方面展开描述。

由表 3b 可见，课程未完成者关注度最高的为话题 3(0.274)和话题 5(0.261)。从话题 3 可知,他们认为课程值得推荐和肯定了教师的讲课方式。话题 5 与在线课程内容制作相关。话题 1 主要关注课程考核方面。值得注意的是,通过观测这 3 个话题围绕的具体内容,我们发现这部分学习者在试图解释未完成课程的可能原因。例如:也许是因为出于感兴趣而能力有限或者超出大学选修范围,也许由于新的注册用户难以熟练操作在线学习平台或者在线视频资源播放的问题,也许错过了考试、互评、练习而没能达到课程考核要求等。因课程未完成者的人数较少,我们对他们发表的课程评论内容进行了观察验证,发现这些评论内容与 LDA 话题挖掘的结果具有较高的一致性。例如:“这是一门同学极力推荐的课程,虽然并不在我的大学选修范围内,但相信如果有机会的话,我一定会把这门课上完的。”“还在学习之中,学习到了后面就感觉有点困难了。”

通过观测和分析表 3a、3b,参与评论的学习者整体上对该门课程表现出认同和赞赏,都一致认为是一门值得推荐的好课。虽然课程完成者和未完成者关注的话题存在相似之处,但也有一定的差异,尤其在具

体话题的内容方面。相比课程完成者，课程未完成者在肯定课程的同时，经常会含蓄地表达出未完成课程的可能原因，较少谈及财务相关的基础理论知识和专业术语，并且传递出更多的消极情感内容，如“错过”“缺乏”“晚”“遗憾”“困难”等，表明未完成者在学习上遇到了更多的困难，需要教师进行针对性的学习干预和情感指导。

(三)话题演化分析

话题演化分析可以快速追踪不同话题的变化态势，直观认识和定位当前的热点话题及概率大小。首先将时间离散化，划分文本集合到不同的时间窗口，然后运用 LDA 模型计算相应的话题分布，最后描绘出学习者关注话题的演化趋势图。本文基于以上学习者话题挖掘的结果，选取了“课程内容”“讲课方式”“课程考核”“在线课程制作”“财务基础知识”5 个热点话题，从 2013 年 10 月起到 2016 年 3 月止，时间粒度以 3 个月为单元，进而展开学习者话题演化分析，如图 4 所示。

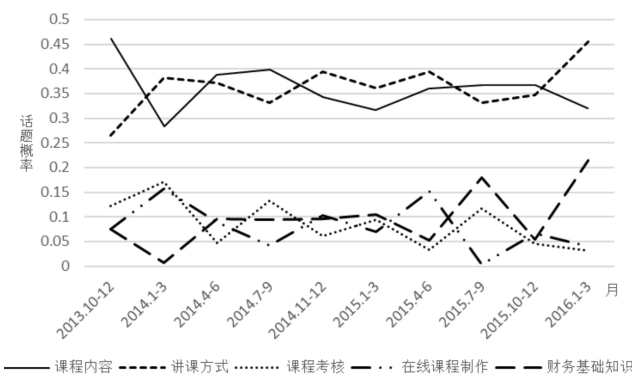

图 4　课程评论者的话题演化图

从图 4 可知每个话题在不同时间段的分布情况以及在整个时间区间的变化趋势。最上面两条曲线是关于课程内容评价和教师授课方式的话题，其概率平均值显著高于其他 3 个，分别达到了 0.37，占据了整个话题分布的 70%以上，并且在各个时间段一直处于平稳状态，换言之，学习者的评论内容始终在围绕课程内容和授课方式话题展开，这也正好吻合了不同学习者类型话题挖掘的结果。而课程考核、在线课程制作和财务基础知识 3 个话题的概率变化基本处于一个水平波动趋势，仅占据了整个话题比重的小部分，尚未发现显著的演化规律。

五、研究结论和建议

MOOC 平台的广泛式发展正促使在线互动学习中产生的文本评论集不断激增，合适的数据挖掘方法对学习分析研究至关重要[28]。本文旨在探究在线学习者评论文本中的话题分布情况，结合定量和定性研究方法，首先利用高频词汇分析法实现对所有课程评论内容的整体认识，然后提出应用 LDA 模型自动挖掘不同学习者类型发表的课程评论中的隐含话题，并可视化呈现学习者热点聚焦话题的演化趋势。实验结果表明：该模型具有自动分析和深度解读在线文本内容和结构的能力，可为基于文本挖掘的学习分析研究提供一种新的思路；学习者高度认可和赞赏此课程，并且尤为关注课程内容和教师讲课形式的话题；相比课程完成者，课程未完成者更倾向于解释他们未完成课程的主要原因，表达出更多的消极情感内容，并较少涉及课程相关的专业知识。这些结果将有助于管理者改善在线学习平台的基础建设和服务体验，有助于教学者快速定位话题分布来评估学习过程和优化教学设计，有助于为学习者直观呈现其话语贡献度、活跃度、与课程本身的契合度以及与学习同伴间评论内容的相似性和差异性，引发自我反思，并为进一步实现个性化干预和自适应反馈奠定坚实的基础。在当前研究的基础上，本文提出以下建议。

1. 推送个性化学习资源

基于学习者课程评论的话题分布，向学习者群体和个体推荐不同粗细粒度的学习资源，不仅包括视频和文件资料，而且可推送兴趣课程、同伴、社区等各种虚拟与实体的学习资源，满足学习者的内在需求，并根据学习者对资源的反馈信息，自适应动态调整并优化推荐策略。

2. 加强课程交互

Kiemer 明确指出[29]，具备交互性对话的课程可以显著提高学习者的内在学习动机和行为参与度，促进课程活力，有助于其达到优秀的学业成效。本研究发现 MOOC 平台中学习者课程交互程度不高，参与课程评论的学习者仅占课程注册学习者的 1/3 左右，并且课程未完成者的评论平均字符数明显低于课程已完成者。那么，加强课程交互并创设有效的互动情境对改善学习者学习状态至关重要。

3. 构建话题—情感语义空间

情感是教学活动中的一种非智力因素，它对激发学习者的学习兴趣、促进优良学习情境、提高学习成效至关重要[15]。将话题挖掘和情感识别相结合，建立话题—情感映射视图，能够识别不同话题内容的情感类别[30]，并检测出隐藏负面情绪的学习话题，尤其应当加强对未完成学习者关注话题—情感内容分析的重视，试图解释他们没能完成课程的可能原因，为教

师及时掌握学习者的思想状态及实施心理干预提供科学的决策支持。

4. 可视化呈现文本分析结果

可视化技术是提高识别结果可理解性的一种可行方法,能够以直观、形象的图形来展示文档的话题特征结构和单词内容,结合有效的话题标记,方便浏览者轻松辨认课程评论的话题信息,并且呈现不同情绪类别的话题演化趋势,以增强抽象信息的接收和认知[31]。

以数据驱动的教学理念为洞察学习者的学习需求、认知风格和情感状态提供了充足的支撑,必将改变传统教育数据应用的范式[32]。除了本文提出的非监督学习方法 LDA 模型,分类、聚类、社交网络分析、关联规则等数据挖掘方法在学习分析领域的应用均有远大前景,从而为创新教学决策和变革教学模式提供实践指导。

[参考文献]

[1] 吴永和,曹盼,那万里,马晓玲. 学习分析技术的发展和挑战——第四届学习分析与知识国际会议评析[J]. 开放教育研究,2014,20(6):72-80.

[2] REICH J. Rebooting MOOC Research[J]. Science, 2015, 347(6217): 34-35.

[3] KIZILCEC R F,PIECH C,SCHNEIDER E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses [C]//Proceedings of the Third International Conference on Learning Analytics and Knowledge. New York:ACM Press,2013:170-179.

[4] ANDERSON A, HUTTENLOCHER D, KLEINBERG J, et al. Engaging with massive online courses [C]//Proceedings of the 23rd International Conference on World Wide Web. New York: ACM Press, 2014: 687-698.

[5] WEN M, ROSE C P. Identifying latent study habits by mining learner behavior patterns in massive open online courses [C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York: ACM Press, 2014: 1983-1986.

[6] 傅钢善,王改花. 基于数据挖掘的网络学习行为与学习效果研究[J]. 电化教育研究,2014(9):53-57.

[7] 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(3):614-628.

[8] HALAWA S, GREENE D, MITCHELL J. Dropout prediction in MOOCs using learner activity features [J]. Experiences and best practices in and around MOOCs, 2014(7):3-12.

[9] RAMESH A, GOLDWASSER D, HUANG B, et al. Learning latent engagement patterns of students in online courses[C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Toronto: AAAI Press, 2014:1272-1278.

[10] RAMESH A, GOLDWASSER D, HUANG B, et al. Modeling learner engagement in MOOCs using probabilistic soft logic [C]// Workshop on Data Driven Education, Advances in Neural Information Processing Systems 2013. United States: Curran Associates Press, 2013:1-7.

[11] FERGUSON R, WEI Z, HE Y, et al. An evaluation of learning analytics to identify exploratory dialogue in online discussions[C]// Proceedings of the Third International Conference on Learning Analytics and Knowledge. New York: ACM Press, 2013: 85-93.

[12] EZEN-CAN A, BOYER K E, KELLPGG S, et al. Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach [C]//Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. New York: ACM Press,2015b:146-150.

[13] EZEN-CAN A, GRAFSGAARD J F, LESTER J C, et al. Classifying student dialogue acts with multimodal learning analytics[C]// Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. New York: ACM Press,2015: 280-289.

[14] RAMESH A, GOLDWASSER D, HUANG B, et al. Understanding MOOC Discussion Forums Using Seeded LDA[C]//Proceedings of the 9th ACL Workshop on Innovative Use of NLP for Building Educational Applications. New York: ACM Press, 2014:28~33.

[15] WEN M, YANG D, ROSE C P. Sentiment analysis in MOOC discussion forums: what does it tell us [J]. Proceedings of educational data mining, 2014, 45(4): 1-8.

[16] KELLY S. Classroom discourse and the distribution of student engagement [J]. Social psychology of education,2007,10(3):331-352.

[17] WANG Z, PAN X, MILLER K F, CORTINA K S. Automatic classification of activities in classroom discourse [J]. Computers & education, 2014, 78: 115-123.

[18] BLEI D M, NG A Y and JORDAN M I. Latent dirichlet allocation[J]. The journal of machine learning research,2004(3):993-1022.

[19] LI L, ZHENG L, YANG F, et al. Modeling and broadening temporal user interest in personalized news recommendation [J]. Expert systems with applications, 2014, 41(7): 3168-3177.

[20] MAO Q, FENG B, PAN S. Modeling user interests using topic model[J]. Journal of theoretical and applied Information technology, 2013, 48(1): 601-606.

[21] JIANG B, SHA Y. Modeling temporal dynamics of user interests in online social networks[J]. Procedia computer science, 2015, 51(1): 503-512.

[22] SHATNAWI S, GABER M M, COCEA M. Automatic content related feedback for MOOCs based on course domain ontology[M]. Switzerland: Springer, 2014.

[23] ZHANG H P, LIU Q, CHENG X Q, et al. Chinese lexical analysis using hierarchical hidden markov model [C]//Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. New York: ACM Press, 2003:63-70.

[24] GOLDER S A, MACY M W. Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures[J]. Science, 2011, 333(6051): 1878-1881.

[25] 刘清堂,武鹏,张思,黄景修,吴林静. 教师工作坊中的用户参与行为研究[J]. 中国电化教育,2016(1):103-108.

[26] CELIKYILMAZ A, HAKKANI-TUR D, TUR Q. LDA based similarity modeling for question,answering [C]//Proceedings of the NAACL HLT 2010 Workshop on Semantic Search. New York: ACM Press, 2010: 1-9.

[27] GENG L, WANG H, WANG X, et al. Adapting LDA model to discover author-topic relations for email analysis [J]. Lecture notes in computer science,2008, 5182: 337-346.

[28] 张琪,武法提. 学习分析中的生物数据表征——眼动与多模态技术应用前瞻[J]. 电化教育研究,2016(9):76-81,109.

[29] KIEMER K, GRÖSCHNER A, PEHMER A K, et al. Effects of a classroom discourse intervention on teachers´ practice and students´ motivation to learn mathematics and science [J]. Learning and instruction, 2015, 35: 94-103.

[30] FU X, LIU G, GUO Y, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon [J]. Knowledge-based systems, 2013, 37: 186-195.

[31] 马秀麟,赵国庆,朱艳涛. 知识可视化与学习进度可视化在 LMS 中的技术实现[J]. 中国电化教育, 2013(1):121-125.

[32] BIENKOWSKI M, FENG M, & MEANS B. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief[R]. Washington: U.S. Department of Education, 2012:1-57.

Study on Learners´ Topics Mining of MOOC-oriented Course Review

LIU Sanya, PENG Xian, LIU Zhi, SUN Jianwen, LIU Hai

(National Engineering Research Center for e-Learning, Central China Normal University, Wuhan Hubei 430079)

[Abstract] This study takes the course Financial Analysis and Decision-making of MOOC college in Guokr as the experimental subject to mine the learners´ topics through analysis of course review posts. Firstly, the study adopts the quantitative method of high frequency words analysis to realize the overall understanding of the content of learners´ course review. Then, the learning analysis is used to study the learners who have completed the course as well as learners who haven´t completed it respectively. The unsupervised learning method LDA model is employed to automatically excavate and resolve the feature structure and semantic content of text review information, explores and tracks the trend of hot topics that learners are concerned about. The study results show that learners highly recognize and appreciate this course, and pay special attention to the course content as well as teachers´ teaching forms. Compared to the completer, the learners who haven´t completed the course tend to explain the main reasons for the unfinished course, express more negative topics and less professional theoretical knowledge of the course.

[Keywords] MOOC; Text Review; Topic Mining; LDA