

大数据支持的慕课论坛教师干预预测及应用

吴林静, 马鑫倩, 刘清堂, 王瑾洁, 高喻

(华中师范大学 人工智能教育学部 教育信息技术学院, 湖北 武汉 430079)

[摘要] 针对慕课论坛中讨论主题数量巨大,教师难以及时反馈的现象,文章提出了一种基于大数据技术的慕课论坛教师干预预测方法。该方法根据学习者的干预需求将慕课论坛中的教师干预分为三种类型:内容相关需干预、管理相关需干预和无需干预。在该分类的基础上,提出了基于词类进行语义特征提取和基于课程知识图谱的内容特征提取方法,对讨论帖的文本内容进行表征,并通过机器学习的方法对教师干预类型进行预测。以中国大学慕课网中“数据库系统概论”课程的教师答疑区主题讨论为数据源,对上述方法的有效性进行验证,发现:(1)基于词类和知识图谱的语义表征方法能够对论坛主题的教师干预需求进行预测,准确率可达到75.86%;(2)不同类型的讨论帖具有不同的语义特征,反映出慕课学习中学习者不同的学习需求,需要教师给予及时、个性化的干预与指导。将慕课论坛教师干预的预测结果推送给慕课教师和课程管理人员,可以大大提升教学管理效率和学习者学习体验。

[关键词] 慕课; 大数据; 教师干预; 文本分类; 语义挖掘

[中图分类号] G434

[文献标志码] A

[作者简介] 吴林静(1987—),女,湖北松滋人。副教授,博士,主要从事数据挖掘、人工智能与教育应用研究。E-mail: wlj_sz@126.com。

一、引言

慕课课程由于其开放、优质、免费等特征,受到了教育者和学习者的广泛认可,并给教育领域带来了巨大的变化。根据第三方在线教育机构 Class Central 统计,截至2018年底,全球慕课课程数量为1.14万,用户注册数量达到1.01亿人^[1]。与此同时,我国的慕课规模更是飞速发展:教育部于2019年4月表示,我国的上线慕课数量和应用规模已跃居世界第一^[2]。慕课的广泛应用为数字化学习提供了极大的便利,使得最优质的教育资源得到了良好的利用。但慕课也存在一个先天的不足之处,即:由于过高的生师比,导致慕课缺乏师生交互,教师难以针对学习者的需求提供及时、个性化的指导。针对这一现象,各大慕课平台均提供在线论坛和讨论区等方式以促进师生之间和学习

者之间的沟通和交流,建立在线学习社区,从而缓解教学过程中缺乏交互的问题。但是由于慕课大规模的特征,慕课论坛中的讨论区往往会迅速累积大量的讨论帖,导致慕课教师没有足够的精力去处理这些讨论,为学习者答疑解惑。大量研究表明,主题讨论中如果缺乏教师反馈,会大大影响学习者的学习成效和社会存在感,并进一步导致辍学等现象发生^[3]。在慕课学习中,学习者的大规模与学习者的个性化学习需求之间已经成为一种尖锐的矛盾,极大地限制了慕课的发展和应用^[4]。针对这一矛盾,本文拟通过人工智能的相关技术对论坛中的主题讨论文本进行分析,以预测教师是否需要当前的讨论主题进行干预,以及进行何种类型的干预。这种预测可以帮助教师快速筛选出那些需要进行干预的主题讨论以及需要干预的类型,从而为慕课教师和助教提供参考,提升慕课管理的效率

基金项目:华中师范大学湖北省教育信息化研究中心开放基金项目“面向MOOC资源组织的课程知识图谱构建与应用研究”(项目编号:HRCEI2020F0105);华中师范大学校级教学研究项目“网络学习环境中教师接近行为对学生学习投入的影响研究”(项目编号:201958)

和学习者的学习体验。

二、相关研究

(一) 文本的语义特征

在慕课论坛中,师生主要通过文字进行交流,偶尔辅以图片等其他媒体形式。因此,对论坛数据的分析离不开对文本的分析和理解。在传统的自然语言处理技术中,对文本的语义表达主要使用词汇作为语义特征,并由此形成了经典的“词袋模型”(Bag of Words)。但由于论坛中主题帖大部分属于短文本,在使用词袋模型进行表征时会产生严重的数据稀疏问题,导致文本处理的准确率大大下降。为了缓解这一问题,很多学者提出了以词类作为文本表征的方法。韩普等人指出词性是文本聚类中的一种重要特征,使用词类作为文本特征进行文本分析时能够大大降低文本的维度并提升聚类效果^[5]。除了词性特征外,PenneBaker等人进一步指出话语中的词汇能够反映人们内部的心理特征与情感,因此,其从心理和认知的角度进一步提出了一系列的词类特征,以探索文本中所反映出的认知、情感、人格、社会心理等内部特征^[6]。在这些特征的基础上,Pennebaker等人进一步开发了基于词类的语义分析工具LIWC(Linguistic Inquiry Word Count)^[7],用于对文本进行分析。LIWC的有效性经过了大量的验证,并被广泛应用于各类文本的特征分析中^[8-10]。早期的LIWC工具缺少中文词典的支持,不适合于进行中文的分析与处理,因此,中国科学院心理研究所参照LIWC的词分类体系和词库,研发了“文心中文心理分析系统”,用于对简体中文进行语言心理分析^[11]。该系统在学习相关的文本分析中也被证明有效^[12]。在本研究中,对文本的表征将混合使用词袋模型和词类模型,以缓解论坛数据的短文本特征所导致的数据稀疏问题。

(二) 领域知识图谱

知识图谱最早起源于谷歌公司2012年所提出的“大规模知识图谱”,指的是一种新型的、结构化的语义知识网络,能够描述现实世界中各种概念及其语义关系,是大数据和人工智能中的前沿研究方向^[13]。在通用领域,公开的知识图谱主要有:Google Knowledge Graph、Microsoft Concept Graph、Dbpedia、Freebase、知立方、知心等。教育领域的知识图谱代表性项目主要有:美国Knewton公司利用知识图谱构建跨学科知识体系^[14];面向智能知识检索的知识库引擎Wolfram Alpha;微软与清华大学联合发布的“开放学术图谱”、北京师范大学的育人知识图谱^[15]、华中师范大学的学科知识图谱^[16]等。无论是通用领域的知识图谱,还是

教育领域的知识图谱,其本质作用均是为计算机提供领域内的基础知识,以支持各类智能应用中的分析与推理。典型研究如知识图谱支持的深度学习^[17]、教育资源推荐系统^[18]、适应性学习系统^[19-20]等;将领域知识图谱应用于MOOC课程教学中,以实现课程结构重构^[21]、对学生成绩进行预测^[22]、对教学过程进行优化^[23]等;领域知识图谱也可以作为学习者建模过程中的知识基础^[24]。在这些研究中,知识图谱均承担着为各类应用提供领域基础知识的主要作用。在本研究中,学习者在论坛中提出的各类疑问有些与学习内容直接相关,有些则可能不相关(如关于课程管理方面的疑问等)。因此,引入领域知识图谱对论坛中的讨论帖进行特征提取可以在快速降低特征维度的同时,有效区分内容相关类与无关类讨论。

(三) 慕课论坛中的教师干预

论坛是慕课中教师和学生沟通交流的主要工具和方式^[3]。已有研究指出慕课论坛中教师的干预能够提升学生的学习^[25],且学生能够从中受益^[26]。如果教师不参与论坛讨论,则学生可能会感觉到被忽视或孤立^[27]。然而,慕课本身大规模的特性导致教师难以遍历所有的讨论并给出有针对性的指导。如何快速从海量的讨论主题中识别出有价值的、需要教师干预的话题成为慕课管理和教学中的一大难题。为提升慕课管理效率,研究者尝试利用人工智能和机器学习的方法对论坛内容进行甄别,以辅助教师筛选需要干预的讨论主题。典型研究如:Chanaa等人通过深度学习模型和学科领域本体对慕课论坛中学习者可能存在疑惑的主题进行自动识别^[28];也有学者通过自然语言处理的方法预测慕课论坛中学习者的沮丧情绪,并进行干预^[29];Arguello等人对慕课论坛中每一条主题帖的语义角色进行自动识别,并将结果提供给教师以方便教师进行干预^[30]。总体而言,慕课论坛中教师干预的识别研究已经开始得到重视,且以英语为主要语言的慕课平台中已出现了部分有参考价值的研究,但中文慕课平台的相关研究较少,且由于中文理解的复杂性而准确率不高。本文将对中国大学慕课平台上的课程为研究对象,结合自然语言处理与知识图谱等方法,对慕课论坛主题中的教师干预进行自动识别,以提高教师管理慕课论坛的效率,提升学习效果。

三、大数据下的慕课论坛教师干预预测模型

(一) 慕课论坛教师干预预测方法

慕课论坛是促进教师和学习者沟通交流的工具。在论坛中学习者、教师、助教等通过文字进行讨论、答

疑、求助等。与传统基于机器学习的文本处理相比,慕课中的讨论文本具有如下的明显特征:

1. 慕课论坛中的评论数量巨大,但单条评论较短且字数较少。由于慕课学习者众多,论坛发帖数量庞大,给教师带来管理上的困难。同时,与其他类型文本相比,慕课单条评论较短,一般在3~100字之间,属于典型的短文本。由于文本太短,在将转换为向量空间模型以进一步进行分析时,会出现数据稀疏的问题,使得文本分析的准确率大大下降。

2. 慕课论坛中既包含与课程内容高度相关的评论,如关于课程内容本身的疑问和主题讨论等;也包含与课程内容无关、但与学习过程有关的评论,如关于考试、作业、资源等的讨论。论坛内容的复杂性增加了信息筛选的难度,为不同职责的管理人员分工管理论坛带来了不便,降低了工作效率。

3. 并非所有的慕课评论都需要教师进行干预,因为有些讨论主题并非学习者的学习疑问或学习者可以通过交流自行解决。Chandrasekaran 等人对慕课课程中的33665个主题讨论进行统计,发现其中10035个主题讨论有教师或者助教进行干预,其余讨论则没有教师干预^[20]。这一现象说明教师干预在主题讨论中并非必须,但筛选出那些需要干预的讨论对于慕课的教学至关重要。

鉴于以上特征,本文提出了一种基于知识图谱和机器学习技术的论坛教师干预预测方法,如图1所示。

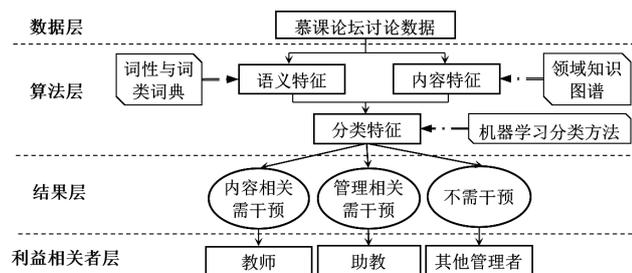


图1 慕课论坛教师干预预测方法

该方法包含四个层次:数据层、算法层、结果层和利益相关者层。数据层主要包括在慕课论坛中所形成的讨论数据,参与者主要包括学习者、教师和助教。在算法层,首先在语义分析工具的词性与词类词典的帮助下,提取出讨论文本的语义特征,包括各类词性特征、核心词类特征和基本句法特征。然后通过领域知识图谱,提取出讨论中的核心内容特征,尤其是与课程内容相关的概念特征。再次将语义特征和内容特征合并,形成特征集合,以作为讨论数据分析的分类特征。最后,通过机器学习中的相关分类方法对讨论进行文本分类,以将讨论数据中值得关注的部分筛选出

来。结果层主要包含讨论主题的分类结果,所有讨论被分为三个类别:内容相关需干预、管理相关需干预和不需干预。利益相关者层包含各类慕课教学的利益相关者,如教师、助教、学习者、其他管理人员等。这些利益相关者可以根据分类结果,有针对性地对讨论进行干预,如针对内容相关需干预的讨论,可以由教师根据教学内容提供相应的干预以促进学习;针对管理相关需干预的讨论,可以由助教针对相关问题向学生提供干预;针对不需干预的讨论,可以利用词云、主题识别等方法生成相关的大数据信息,为相关管理人员的管理和决策提供参考。

(二)慕课评论中的教师干预分类

慕课中的教师干预分类方法较多,最常见是二分类,即将评论分为需要干预和不需要干预两类。该分类已经可以大大提高慕课管理的效率,帮助教师迅速将需要关注的讨论筛选出来。本文在上述二分类体系的基础上进行了优化,将需要干预的讨论进一步划分为内容相关需干预和管理相关需干预。这样可以有效辅助慕课教学管理分工,如教师可以专门针对内容相关讨论进行关注和引导,而助教团队则可以针对管理相关讨论进行指导,使得慕课教学团队可以更加高效地分工合作。教师干预分类的具体定义及示例见表1。

表1 教师干预的分类及示例

分类	定义	示例
内容相关需干预	讨论主题与教学内容密切相关,包括对教学内容的疑问、评价、扩展等	如何让查询结果的列显示顺序与数据库中的列顺序不一样
管理相关需干预	讨论主题与教学内容本身无关,但与考试、作业、资源、日程安排、网络等教学管理工作相关	实验系统注册不了,进入不了实验系统。做不了实验
不需干预	不属于上述两类的其他讨论,包括社交信息、情绪表达、其他无关信息等	来报到啦,大家多多关照,哈哈

(三)基于词类和知识图谱的分类特征

由于慕课讨论文本的典型特征,本研究中用于文本分类的分类特征并非使用传统的词汇特征,而是由文本的语义特征与内容特征组合而成。下面对其原理进行详细说明。

1. 文本的语义特征

文本的语义体现为词汇,但由于中文词汇数量巨大,短文本如果直接使用词汇作为特征会导致严重的数据稀疏问题。使用词类特征则较好地避免这一问题。词类特征的基本思路是:将属于同一类型的词汇作

表 2

文本的语义特征列表

词类语义特征		部分实例
词性类特征(23个)	代名词、动词、助动词、副词、介系词、连接词、否定词、概数词、脏话	代名词:彼此、在下、对方、别人 后置词:之中、以上、为止
时态类特征(5个)	时态标定词、过去词、现在词、未来词、延续词	时态标定词:已经、之前、日后 延续词:至今、近期
情感类特征(6个)	情感历程词、正向情绪词、负向情绪词、焦虑词、生气词、悲伤词	情感历程词:气愤、感恩、失望 正向情绪词:信心、满足、棒
历程类特征(22个)	社会历程词、家族词、朋友词、人类词、认知历程词、洞察词、因果词	社会历程词:打招呼、报到 认知历程词:理解、选择、质疑
功能类特征(15个)	功能词、相对词、移动词、空间词、时间词、工作词、休闲词、金钱词	功能词:或许、许多、那些 相对词:以前、相比、达到
文本基本语法特征(17个)	句号、逗号、冒号、分号、问号、惊叹号、破折号、引号、缩写符号.....	每句平均次数=总词数/总句数

为一个类别,统计其数量,并计算其占总词数的百分比作为权重。这样将属于同一类别的词进行合并,从而减少特征数量,提高特征权重,改善数据稀疏问题,提升识别效果。除此词类特征之外,文本的句法特征也对语义的理解具有重要作用,如句子长度、标点符号等。Pennebaker 等人对文本语义表达具有重要作用的词类和句法特征进行了总结,并建立了相应的词典和工具 LIWC。由于本研究中所处理的全部数据均为中文,因此,本文选择使用与 LIWC 词类兼容的文本分析工具“文心”(TextMind)对讨论数据进行分析,以实现语义特征的提取。“文心”中所包含的词类特征见表 2。

2. 文本的内容特征

文本的内容特征体现为文本中与教学内容相关的核心词汇。为了提取出这些特征,需要预先定义领域知识图谱,对学科中的核心知识进行描述。然后根据知识图谱中的核心概念和术语,对文本进行特征筛选,并计算其 TF·IDF 值作为特征的权重。一个主题讨论 j 中所包含的特征词 t_i 的特征权重 w_{ij} 的计算方法如下:

$$w_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \times \lg \frac{|D|}{|\{j:t_i \in d_j\}| + 1}$$

其中 n_{ij} 表示特征词 t_i 在主题讨论 j 中出现的次数, k 为主题讨论集合中所有包含特征词 t_i 的讨论。|D| 为整个集合中主题讨论的数量, $|\{j:t_i \in d_j\}|$ 为集合中包含特征词 t_i 的主题讨论的数量。此外,特征词个数占

主题讨论总词数的百分比也被作为内容特征之一,用于描述讨论的内容相关性。

以如下主题讨论为例,首先将主题讨论中的所有帖子合并在一起,形成一段文本;然后通过知识图谱进行特征选择,实现特征提取;最后通过 TF·IDF 方法为特征计算权重,形成内容特征集合,如图 2 所示。

四、中国大学慕课案例分析与应用

(一)数据来源

为了验证提出方法的有效性,课题组从中国大学慕课(<https://www.icourse163.org/>)上选取了课程“数据库系统概论(基础篇)”从第 2 至 8 次开课共 7 轮课程(其中第 1 次开课时,课程尚处在完善阶段;第 9 次课目前尚未结束)的所有教师答疑区的全部讨论主题作为数据源,并通过本文所提出的方法进行了分析。7 轮课程共采集到讨论主题的数量为 1015 个,包含讨论帖的数量为 2877 个。由于教师的干预发生在学生发表讨论帖之后,是否需要进行干预主要通过学生主题帖的内容进行判断。因此,对每一个讨论主题中的所有讨论帖,按照教师是否进行了回复进行截断,仅保留教师回复前的所有讨论帖,并将同一主题下的讨论帖合并为一个文本,作为是否需要进行教师干预的原始数据。两名经过培训的教育技术学研究生对所有讨论主题按照表 1 中的分类进行了数据标注。若两名

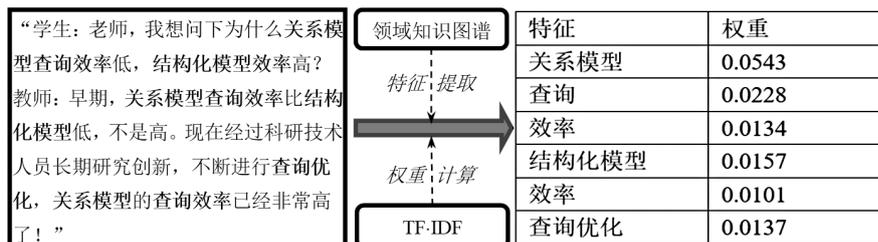


图 2 基于知识图谱的内容特征提取示例

标注员标注结果一致,则将该结果作为最终标注结果;若两名标注员标注结果不一致,则相互商量直到获得一致结果。最终的数据分布情况见表3。

表3 讨论主题的分布情况

教师干预类型	内容相关需干预	管理相关需干预	不需干预	合计
讨论主题数量	576	316	123	1015

(二)领域知识图谱构建

为了实现讨论主题内容特征的抽取,本文构建了面向“数据库系统概论”的课程知识图谱。参照谷歌知识图谱,本文所构建的知识图谱主要包括两个组成部分:第一部分是领域中的核心概念,主要包括概念及其属性;第二部分是该领域的基本知识结构,主要包括概念之间的各种语义联系。课程的核心概念是以课程教学大纲和教材为依据,由两位该课程的任课教师依据教学过程进行独立标注。标注后将两人的标注结果进行比较,就不一致的部分进行讨论,并最终取得一致结果。该课程知识图谱共包含核心概念432个。除了得到领域中的所有核心概念,知识图谱还对概念之间的关系进行标注,以形成领域的整体知识结构。概念之间的语义关系有上下位、整体一部分多种类型关系。图3展示了“数据库原理与技术”课程中的最顶层的概念和概念“关系数据库标准语言SQL”及其下位概念之间的语义层次结构。

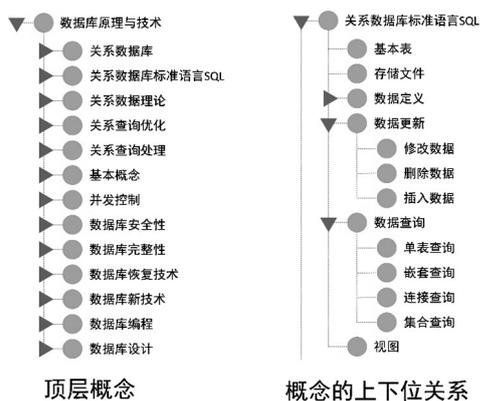


图3 概念之间的上下位关系

(三)分类预测结果

在词性与词类词典和课程知识图谱的帮助下,本文对数据集中的1015个讨论主题进行特征提取,并按照7:3的比例将数据集分为训练集和测试集。运用机器学习中的不同分类算法,通过训练集进行模型训练后,通过测试集验证方法的有效性。表4展示了通过不同的分类算法进行测试所得到的干预类型分类的总体精度、按类别加权准确率、按类别加权召回率

和kappa值。

表4 不同分类模型对教师干预类型的预测结果

分类模型	总体精度	加权准确率	加权召回率	kappa
KNN	64.04%	58.15%	64.17%	0.306
梯度提升树	75.86%	75.32%	75.96%	0.553
朴素贝叶斯	42.86%	60.73%	42.75%	0.190
J48	69.46%	69.88%	69.53%	0.465
随机森林	71.92%	62.81%	72.07%	0.436
神经网络	72.91%	72.94%	73.04%	0.487

从结果中可以看出,各分类模型的总体分类精度分布于42.86%至75.86%之间,其中三种模型能够达到70%以上的准确率,说明本文提出的特征提取方法能够有效地对论坛讨论中的教师干预类型进行分类预测。在所有模型中,梯度提升树模型获得了最好的准确率,达到75.86%,能够高效地识别出需要教师干预的讨论主题,并对干预类型进行判定,帮助教师及其团队更为高效地进行慕课教学和管理。

(四)各类别的语义特征

为了更好地了解需要教师干预的不同类别讨论主题的语义特征,从而为教师提供参考,本文对三类不同讨论主题的文本内容进行了词频统计,并分别选取了各类别中词频排序前10的名词短语,其结果见表5。

表5 不同类型教师干预讨论主题的语义特征

类别	语义特征
内容相关需干预类主题	数据库、数据、关系、老师、属性、用户、查询、理解、SQL、实体……
管理相关需干预类主题	实验、老师、注册、课程、平台、数据库、提交、问题、用户名、考试……
其他	老师、感觉、讲课、课程、视频、自己、希望、数据库、内容、讨论区……

从各类别的语义特征上可以看出,三类不同的讨论主题分别具有各自不同的语义特征。

1. 在内容相关需干预讨论主题中,学习者重点围绕课程内容提出疑问,如“属性”“SQL”“实体”等词汇展示了学习者重点关注的学习内容,也给慕课教师提示了课程重点内容。

2. 在管理相关需干预类讨论主题中,高频词汇包括“实验”“注册”“成绩”等,体现了在慕课课程管理方面,学习者遇到的问题主要包括以下几个方面:实验课程开设问题、平台注册与密码问题、成绩判定问题、资源更新问题。这些问题的及时处理与解决与学习者在慕课平台的学习体验密切相关。将这些疑问推送给助教团队和慕课管理人员以保证学习者得到及时回

复,将能够极大地提升慕课学习的满意度。

3. 在其他类讨论主题中,学习者主要讨论了自己的感受、期望、对课程和教师的看法等这些讨论内容,虽然不需要教师的直接干预,但其内容对于课程改进和提升课程体验依然有着重要作用。可以将这类主题汇总后通过词云等方式推送给课程组织者,以优化课程设计。

五、结 语

慕课以其优质、免费、开放等特征得到了学习者的广泛认可,促进了优质教育资源的共享与利用。本文针对慕课讨论区中讨论主题数量巨大、教师难以及时反馈的现象,提出了一种基于文本的语义特征和内容特征的慕课评论教师干预预测方法。在该方法中,学习者在讨论区所发布的讨论主题按照是否需要教

师干预分为三种类型:内容相关需干预、管理相关需干预和无需干预。在此三种类型的基础上,本文针对慕课讨论主题字数较少、且与课程内容相关的特征,提出了一种基于知识图谱的文本分类特征提取方法,融合文本的语义特征和内容特征,以缓解主题讨论短文本所导致的特征稀疏问题,提高教师干预类型的预测准确率。以中国大学慕课中“数据库系统概论”课程7轮开课中教师答疑区的全部讨论数据作为案例对本文的方法进行验证,结果表明该方法对教师干预类型的预测准确率可以达到75.86%,能够极大地提升慕课中主题讨论的管理效率。下一步,拟进一步扩大方法的应用范围,囊括多门课程,以验证模型和方法的泛化能力,提升其可推广性。

致谢:本研究中使用了中国科学院心理研究所开发的“文心中文心理分析系统”,在此表示感谢!

[参考文献]

- [1] 21世纪经济报道.全球慕课(MOOC)用户破亿!教育部:中国有超过2亿人次学习者[EB/OL]. [2020-10-13].<http://www.cedumedia.com/i/23466.htm>.
- [2] 央广网.教育部:我国上线慕课数量和应用规模居世界第一[EB/OL]. [2020-10-13].http://china.cnr.cn/news/20190415/t20190415_524577387.shtml.
- [3] CHANDRASEKARAN M K, RAGUPATHI K, TAN B C Y, et al. Towards feasible instructor intervention in MOOC discussion forums[C]// International Conference on Information Systems, Fort Worth: 2015.
- [4] 朱美娜, 赵云建. 慕课和开放教育:角色,教育实践,个性化学习和可能的发展趋势——访印第安纳大学教育技术专家柯蒂斯·邦克教授及其著作团队[J]. 中国电化教育, 2017(5): 35-44.
- [5] 韩普, 王东波, 刘艳云, 苏新宁. 词性对中英文文本聚类的影响研究[J]. 中文信息学报, 2013(2): 65-73.
- [6] TAUSCZIK Y R, PENNEBAKER J W. The psychological meaning of words: LIWC and computerized text analysis methods[J]. Journal of language and social psychology, 2010, 29(1): 24-54.
- [7] PENNEBAKER J W, FRANCIS M E, BOOTH R J. Linguistic inquiry and word count: LIWC 2001 [J]. Mahway: erlbaum, 2001, 71 (2001): 2001.
- [8] SELL J, FARRERAS I G. LIWC-ing at a century of introductory college textbooks: have the sentiments changed? [J]. Procedia computer science, 2017(118): 108-112.
- [9] YUAN Y, LI B, JIAO D, et al. The personality analysis of characters in vernacular novels by SC-LIWC [C]//International Conference on Human Centered Computing. Kazan: Springer, 2017: 400-409.
- [10] ZHAO N, JIAO D, BAI S, et al. Evaluating the validity of simplified Chinese version of LIWC in detecting psychological expressions in short texts on social network services[J]. PLoS one, 2016, 11(6): 1-15.
- [11] 文心中文心理分析系统[DB/OL]. [2020-10-15].<http://ccpl.psych.ac.cn/textmind/>.
- [12] 吴林静, 刘清堂, 毛刚, 黄焕, 黄景修. 大数据视角下的慕课评论语义分析模型及应用研究[J]. 电化教育研究, 2017(11): 43-48.
- [13] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(4): 589-606.
- [14] KNEWTON. Knewton adaptive learning building the world's most powerful education recommendation engine[DB/OL]. [2020-09-29]. <https://www.knewton.com/wp-content/uploads/knewton-adaptive-learning-whitepaper.pdf>.
- [15] 余胜泉, 彭燕, 卢宇. 基于人工智能的育人助理系统——“AI好老师”的体系结构与功能[J]. 开放教育研究, 2019, 25(1): 25-36.
- [16] WU L, LIU Q, ZHAO G, et al. Thesaurus dataset of educational technology in Chinese[J]. British journal of educational technology, 2015, 46(5): 1118-1122.

- [17] 姜强,药文静,赵蔚,李松.面向深度学习的动态知识图谱建构模型及评测[J].电化教育研究,2020,41(3):85-92.
- [18] 秦川,祝恒书,庄福振,等.基于知识图谱的推荐系统研究综述[J].中国科学:信息科学,2020(7):937-956.
- [19] 黄焕,元帅,何婷婷,吴林静.面向适应性学习系统的课程知识图谱构建研究——以“Java 程序设计基础”课程为例[J].现代教育技术,2019,29(12):89-95.
- [20] 李艳燕,张香玲,李新,杜静.面向智慧教育的学科知识图谱构建与创新应用[J].电化教育研究,2019,40(8):60-69.
- [21] 王亮.深度学习视角下基于多模态知识图谱的 MOOC 课程重构[J].现代教育技术,2018,28(10):101-107.
- [22] 陈曦,梅广,张金金,许维胜.融合知识图谱和协同过滤的学生成绩预测方法[J].计算机应用,2020,40(2):595-601.
- [23] 陆星儿,曾嘉灵,章梦瑶,郭幸君,张婧婧.知识图谱视角下的 MOOC 教学优化研究[J].中国远程教育,2016(7):5-9.
- [24] 黄涛,王一岩,张浩,杨华利.智能教育场域中的学习者建模研究趋向[J].远程教育杂志,2020,38(1):50-60.
- [25] 张敏,尹帅君,聂瑞,唐存周.基于体验感知的中外慕课学习平台持续使用态度对比分析——以 Coursera 和中国大学 MOOC 为例[J].电化教育研究,2016,37(5):44-49.
- [26] 徐恩芹.师生交互影响网络学习绩效的实证分析[J].电化教育研究,2016,37(9):61-68.
- [27] DOLAN V L B. Massive online obsessive compulsion: what are they saying out there about the latest phenomenon in higher education?[J]. International review of research in open and distance learning, 2014, 15(2):268-281.
- [28] CHANAA A, EL FADDOULI N E. BERT and prerequisite based ontology for predicting learner's confusion in MOOCs discussion forums[C]//International Conference on Artificial Intelligence in Education. Ifrane: Springer, 2020: 54-58.
- [29] ALGHAMDI N S, MAHMOUD H A H, ABRAHAM A, et al. Predicting depression symptoms in an Arabic psychological forum[J]. IEEE access, 2020(8): 57317-57334.
- [30] ARGUELLO J, SHAFFER K. Predicting speech acts in MOOC forum posts[C]// National conference on artificial intelligence, Oxford, 2015:2-11.

Big Data-supported Prediction of Teacher Intervention in MOOC Discussion Forum and Its Application

WU Linjing, MA Xinqian, LIU Qingtang, WANG Jinjie, GAO Yu

(School of Educational Information Technology, Faculty of Artificial Intelligence Education, Central China Normal University, Wuhan Hubei 430079)

[Abstract] In view of the large number of topics discussed in MOOC forums and the difficulty for teachers to provide timely feedback, this paper proposes a prediction method of teacher intervention in MOOC forums based on big data technology. This method divides teacher intervention in MOOC forums into three types according to learners' intervention needs: content-related interventions, management-related interventions and no interventions. On the basis of this classification, a semantic feature extraction method based on word classes and a content feature extraction method based on the curriculum knowledge graph are proposed to represent the text content of discussion posts and predict the type of teacher intervention by machine learning methods. Taking the topic discussions of "Introduction to Database System" course in MOOC of China as the data source, the validity of the above methods is verified, and it is found that: (1) the semantic representation method based on word classes and knowledge graphs can predict teachers' intervention needs for forum topics with an accuracy of 75.86%. (2) Different types of discussion posts have different semantic features, reflecting the different learning needs of MOOC learners, which requires timely and personalized intervention and guidance from teachers. Pushing the predicted results of teacher intervention in MOOC forums to MOOC teachers and course administrators can greatly improve the efficiency of teaching management and learners' learning experience.

[Keywords] MOOC; Big Data; Teacher Intervention; Text Classification; Semantic Mining