**基金项目：**国家自然科学基金项目“网络学习资源深度聚合及个性化服务机制研究”（71704062）；华中师范大学基本科研业务费专项资金项目“大数据驱动的学习者认知过程建模研究” （CCNU20ZN003）。

引用：刘清堂,贺黎鸣,吴林静,杨炜钦,李晶(2020).智能时代的教育文本挖掘模型与应用[J].现代远程教育研究,32(5):95-103.

**摘要：**教育文本挖掘是指通过数据采集和处理，利用数据挖掘算法或工具，从非结构化文本文档中提取有意义的模式或知识的过程。教育文本蕴含着丰富的学习者认知、行为和情感等信息，对其进行深度挖掘和分析，有助于深入探索教育教学的基本规律，解释教育中存在的问题和现象。大数据支持下的教育文本挖掘模型包括数据生产和使用的主体（利益相关者）、教学环境、数据和挖掘工具等核心要素，涉及数据产生、数据采集、数据处理、知识发现、评估解释、教学应用等过程和方法。其常用的数据来源包含问卷调查、在线互动、学习反馈、在线评论、社交媒体和教学文件，主要用于学习者成绩预测、学习者建模、学习者水平评价、教学材料结构分析、学习者反馈和内容可视化等。当前教育文本挖掘在海量数据处理、数据降维保真、结果评估与解释等方面还面临挑战，研究者需深度融合教育学、认知心理学、语言学等多学科研究方法，结合教育教学的基本理论和具体的教育情境，注重多模态分析和验证，保证将其应用于教育研究的科学性。随着相关技术的突破和应用发展，教育文本数据将成为教育现代化发展的推动力，在深度学习、精准教学等领域中发挥更大作用。

**关键词：**教育大数据；数据挖据；学习分析；教育文本挖掘

**一、引言**

智能时代，作为人工智能发展基础的大数据愈发受到关注。《自然》和《科学》分别在2008年和2011年设立了专刊对大数据的特征及应用前景进行研讨，探索利用其破解不同领域难题的途径和方法（孟小峰等，2013）。国务院也于2015年发布《促进大数据发展行动纲要》，指出要全面推进我国大数据发展和应用，加快建设数据强国（国务院，2015）。大数据蕴含着海量有价值的信息，给各行各业带来了历史性的机遇。大数据技术也强烈影响着教育系统，正成为推动教育系统创新、变革的颠覆性力量（杨现民等，2016）。教育大数据具有复杂性、多样性、差异性和内隐性等特征，对其进行挖掘、聚合、组织和应用等一直是智能教育研究的热点问题。

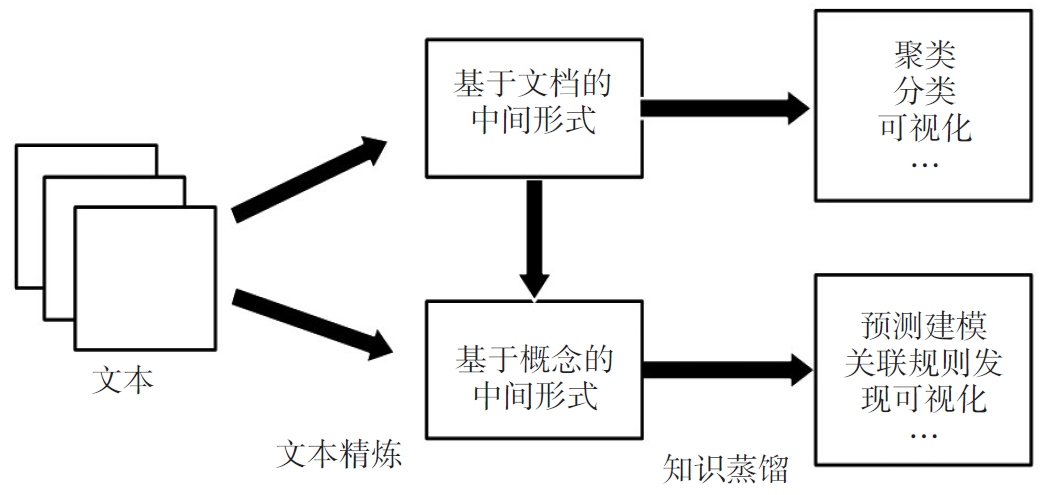
在传统的教育数据挖掘中，研究者往往对结构化数据关注较多，例如课程管理系统中的学生学习活动日志，包括登录次数、浏览时长、提交作业次数、发言次数等（Chen et al.，2014），以及学生的考评信息、学业成绩等（Baker et al.，2009 ）。而据IBM统计，一个组织中大约80%的数据是开放式和非结构化的，这些数据实际上很少被使用（IBM Corporation，2019），而文本数据又是非结构化数据中最主要的组成部分（Grimes，2008）。教育领域中的文本数据挖掘和分析，是一个价值巨大且有待进一步发展的新兴研究领域。与结构化数据相比，文本数据以言语数据为主，可以更加真实地反映学习者的学习动机、认知发展、情感态度、学习体验等（Witten et al.，2016）。利用文本挖掘技术获取教育文本中蕴藏的有用信息并发现复杂教育系统的规律，给教育研究者带来了新的研究视角。

本文以教育文本为研究对象，探索了教育文本挖掘的模型和应用框架，并从数据来源及清洗、分析算法与工具，以及典型应用等方面分析了教育文本数据的采集、分析、挖掘方法和挑战，并对该领域的未来发展趋势进行了展望，以期能为相关研究者提供参考。

**二、文本挖掘基本原理**

文本挖掘（Text Mining），又称文本数据库中的知识发现、文本数据挖掘，一般是指从非结构化的文本文档中提取有趣或者有意义的模式或知识的过程（Tan，1999）。1995年Feldman将数据挖掘与文本分类结合，首次提出文本挖掘的概念（Feldman et al.，1995）。由于文本挖掘处理的是非结构化的数据，因此它涉及额外的处理步骤，以便在知识发现步骤前从文本中定位、提取和构造相关信息（Ananiadou et al.，2010）。

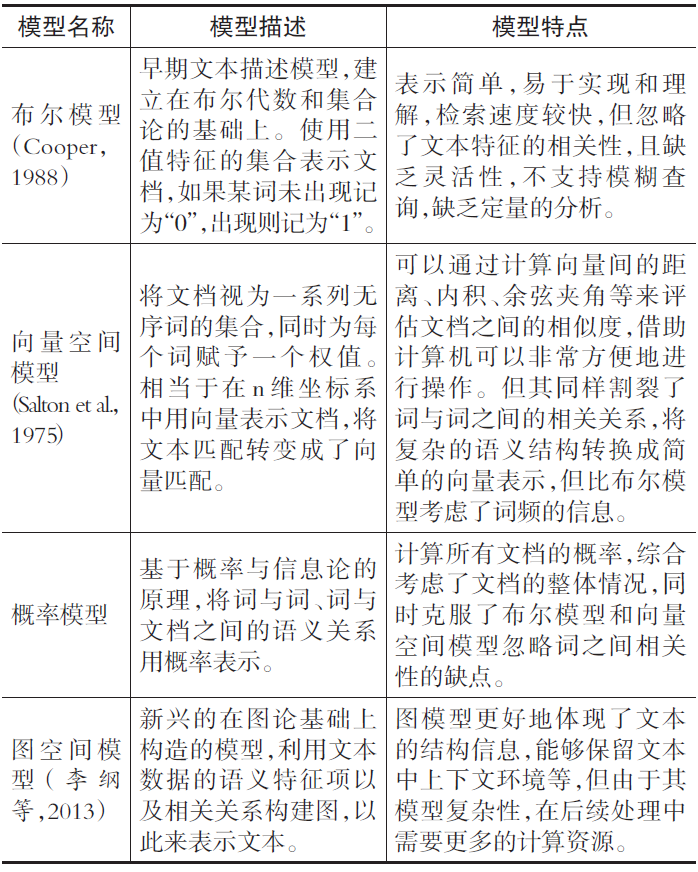
文本挖掘一般包括：数据采集、文本预处理（数据选择与清洗、文档表示、特征选择等）、文本挖掘（分类、聚类、关联规则挖掘等）、文本后处理（模型评估与反馈、知识的解释与可视化等）等步骤。已有许多研究者提出了通用的文本挖掘模型，其中较有代表性的是Tan（1999）提出的两阶段模型（见图1）。该模型把文本挖掘分为文本精炼（将文本文档转换成计算机可以理解的中间形式）和知识蒸馏（从中间形式中推导出模式或知识）两个阶段。中间形式可以是基于文档的，也可以是基于概念的。基于文档的每个实体表示一个文档，通常与领域无关；而基于概念的每个实体表示特定领域中的对象或者概念，与领域相关。



**图1　文本挖掘模型**

常用的中间形式模型有：布尔模型、向量空间模型、概率模型等，各模型描述及特点如表1所示。

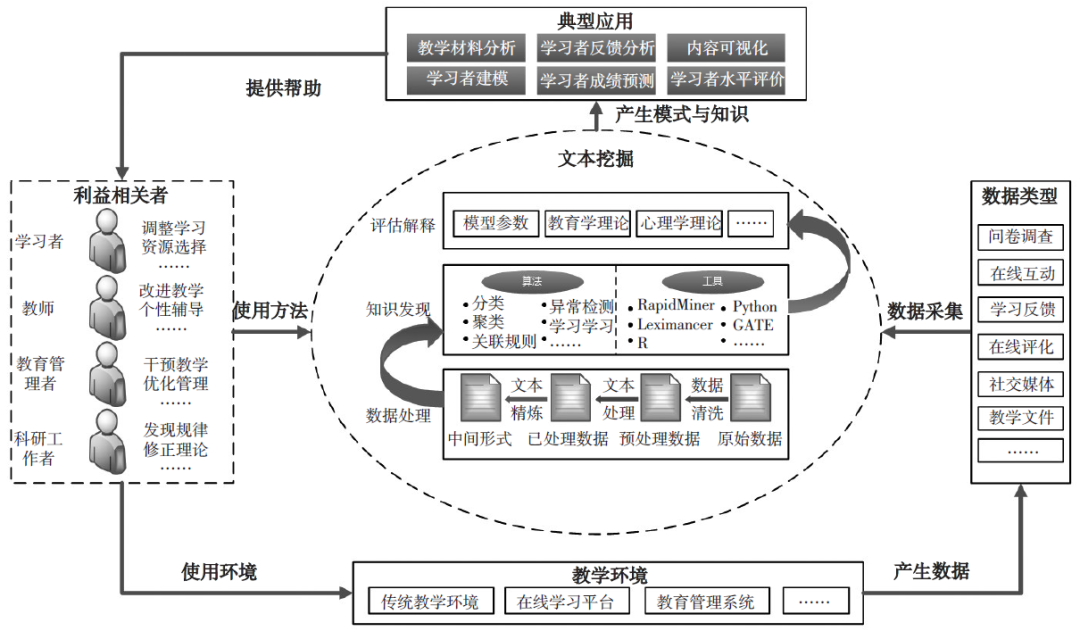
**表1    文本挖掘常用的中间形式模型**



文本挖掘已广泛应用于多个领域，如生物和生物医学领域文本的挖掘（Cohen et al.，2005）、电子邮件的分类与过滤（Kiritchenko et al.，2011）、商业领域的运作和营销改进（Sullivan，2001）、专利的自动化分析（Tseng et al.，2007）等，其在教育中的应用也逐渐成为研究者关注的热点。

**三、大数据支持的教育文本挖掘模型及方法**

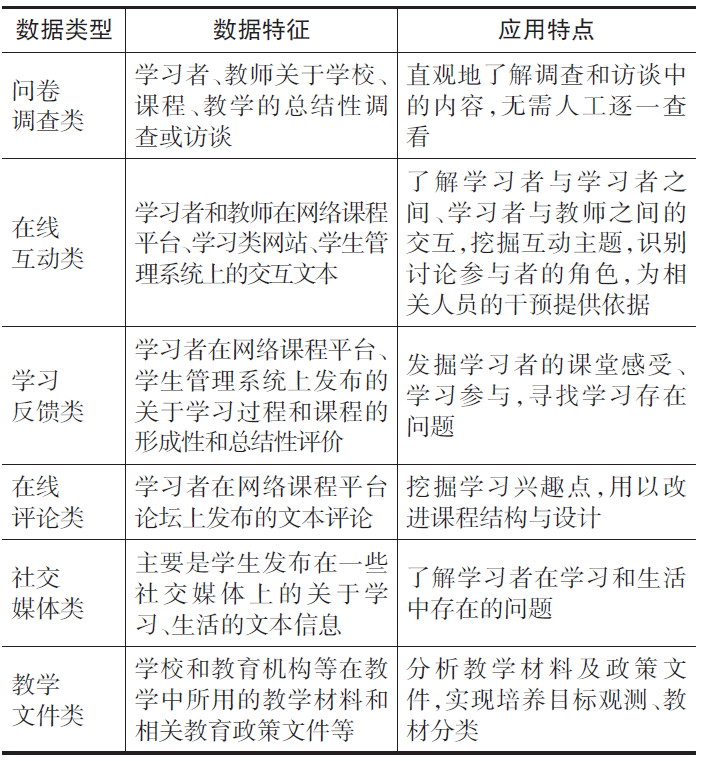
大数据支持下的教育文本挖掘模型如图2所示。该模型包括数据生产和使用的主体（利益相关者）、教学环境、数据和挖掘工具等核心要素，涉及到数据产生、数据采集、数据处理、知识发现、评估解释、教学应用等过程和方法。学习者、教师、教育管理者和科学研究者等利益相关者，既是教育文本数据的使用者，又是生产者。其在教育环境中产生的数据，通过数据采集、数据处理、知识发现、评估解释等文本挖掘过程，产生模式和知识，为教学、管理和科研提供帮助。具体而言，学习者可以借助文本挖掘，获得合适的资源，并通过反馈调整学习。教师可以借助文本挖掘的结果，掌握学习者学习动态，预测教学效果，通过干预改进教学，实现个别化指导等。教育管理者借助文本挖掘，可以评估教学效果，进行教学监测，进而做出科学决策。科研工作者可以借助文本挖掘，发现和解决教育问题，改善教育环境，增进对教育现象和规律的认识。大数据支持下的教育文本挖掘包括数据采集、文本挖掘以及模式应用三个重要环节。

**图2    大数据支持下的教育文本挖掘模型**

1.教育文本的数据类型及采集

在教育环境中，文本数据有着十分广泛的来源，既有教育系统中已经存在的资源类文本数据，也有教与学过程中产生的过程性文本数据，包括传统的教学材料如课件、教材、讲义，各种问卷、试卷的主观回答，学生的作业等，以及随着在线学习平台、网络公开课等的流行出现的各种论坛讨论数据、评论数据、反思数据等。此外，随着社交媒体如Twitter、Facebook、新浪微博、知乎等的广泛应用，其也产生了大量有助于了解教师和学生情感、问题的真实数据。相较于传统文本数据，网络中的文本数据不仅在数量上呈指数级的增长，也更加容易获取和处理。目前教育文本挖掘常用的数据来源大致可以分为如表2 所示的6类：问卷调查、在线互动、学习反馈、在线评论、社交媒体和教学文件。

**表2    教育文本挖掘常用数据来源**

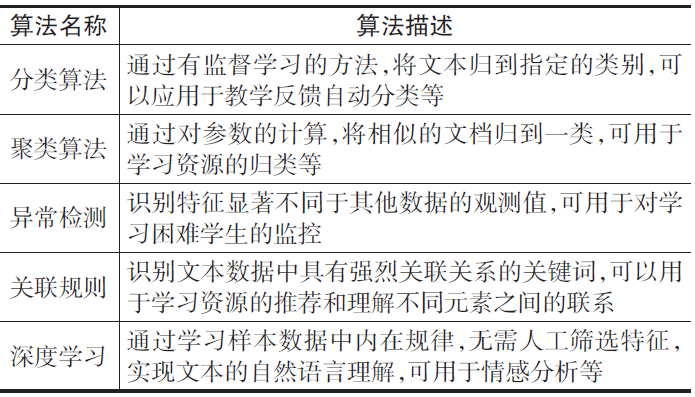


首先，采集的原始教育文本数据并不能直接用于文本数据挖掘，其主要原因在于数据存在冗余、缺失、歧义和冲突等，需要进行数据清洗。其次，根据文本数据的语言学特性，需要加入额外的文本处理环节，包括分词、词性标注、停用词去除、词干提取（英文数据）、词频统计等。文本预处理完成后，可以得到高维的文本特征。高维的文本特征存在大量冗余，需要通过特征选择和提取得到较低维度的、有代表性的特征。常用的特征选择方法包括信息增益法（Information Gain）、互信息法（Mutual Information）、卡方检验法（Chi-square）等。最后，通过对文本特征进行建模，将非结构化的文本数据转换成便于计算机可以处理的中间形式。

2.教育文本挖掘方法及工具

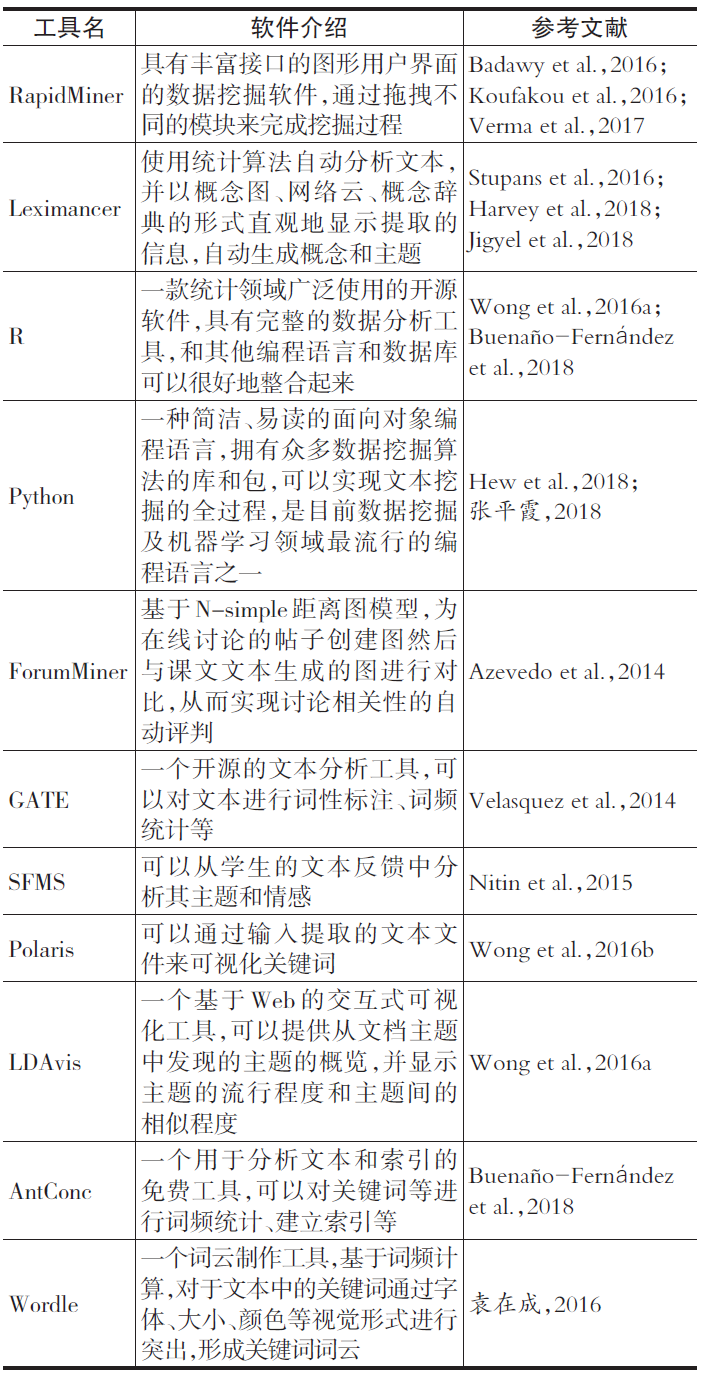
文本挖掘算法通常可以分为两大类：一类是预测性算法，这类算法通过已有的特征值来预测未知的特征值，如各种分类算法等；另一类是描述性算法，其目的是描述概括数据中已经存在的关系和模式，如聚类、关联规则挖掘、异常检测等（Tan，2018）。除了传统的数据挖掘算法外，还有深度学习的相关算法，包括CNN（卷积神经网络）、DBN（深度置信网络）、RNN（循环神经网络）等。此外，还有一些在文本挖掘中十分重要的方法，如Word2Vec（词向量）、LDA（隐含狄利克雷分布）模型、马尔可夫模型、深度学习等。常用方法的具体描述如表3所示。

**表3    教育文本挖掘常用算法**



此外，随着文本挖掘应用需求的不断增长，有许多研究人员和商业公司致力于文本挖掘工具的研发，目前已有许多成熟的工具和软件，表4是对所分析文献中出现的主要文本挖掘工具的概述。

**表4     教育文本挖掘主要工具**



文本挖掘就是使用数据挖掘方法和工具从数据中发现描述性知识或者预测性模型，并对其进行评估与反馈，最后利用相关模型的评价指标和教育学、心理学的相关理论对得到的信息进行评估和解释的过程。如效果不理想则回到之前的步骤，重新选取合适的中间形式或者算法与工具。

3.知识及模式的提取及应用场景

文本挖掘得到的知识或产生的模式可以给学习者、教师、教育管理人员、相关科研工作者提供帮助，用以解决教育问题、提升教学的效果。提取知识或产生模式的方法与具体应用场景紧密关联，如表5所示。

**表5    教育文本挖掘典型应用**



例如，学习成绩预测通常采用关联规则挖掘方法，通过发现学习行为、学习表现间的关系规则，建立成绩预测模型。学习者建模一方面可借助文本数据，通过关键词或特征，依据心理学、教育学基本原理等进行人工标注案例，利用机器学习方法构建学习者特征模型；另一方面，可对学习者进行特征归类，如活泼型、好动型、沉思型等，构建学习者分类预测模型，服务于个性化学习。学习水平评价方法则通常对学习者文本数据进行不同层次水平特征归类，如将主题研讨中的交互文本归纳为描述性、分析性和批判性认知层次，利用词袋模型进行学习交互水平模型的构建。教学材料结构分析、内容可视化通常采用矩阵关联计算方法，建立要素与要素之间的关系，并通过概念图、云签图等进行可视化表示。

此外，当获取到海量的文本数据后，还可以根据教学应用需求，对文本数据进行矢量化处理，利用深度学习算法构建学习者认知模型、学习评价模型等，其特点是不需要进行特征提取，便可在对文本数据标签化后构建精准的分析模型。

**四、教育文本挖掘的典型应用**

教育文本挖掘的典型应用主要包含6类：

1.学习者成绩预测

学习者成绩预测通过收集和挖掘学习者学习文本数据，建立学习者学习行为和学业表现间的关系模型，用以预测学习者最终的成绩。例如香港教育学院的研究者使用Polaris工具对24个本科学生的反思帖子和同伴评语进行建模，实现在通识教育中利用学生机会发现（Merton et al.，2011）的隐藏模式和联系来预测学习者的学习成绩（Wong et al.，2016b）。另一项研究中，研究者基于MOOC论坛和其他学习活动数据共同评估学习质量，发现论坛帖子和评论的语义特征对预测学习质量有较大的影响（Liu et al.，2018）。

2.学习者建模

学习者建模即通过学习者产生的文本数据，塑造不同学习者的特点，如情绪、认知、领域知识和学习偏好等。其主要目的是为了识别不同类型的学习者，从而对其施加相应的措施，以满足学习者的需求。例如，普渡大学的研究者开发了一个多标签的分类器，可以根据该校工科学生在社交媒体（Twitter）上的发言内容自动识别存在特定问题的学生（Chen et al.，2014）。印度尼西亚的研究者基于内容分析的分类模型，使用朴素贝叶斯算法构建意见分类器，通过学生在社交媒体（Twitter）上的情绪评价来评估学生的满意度（Candra Permana et al.，2017）。Doleck等人在一个医学领域的计算机学习环境BioWorld中，使用文本挖掘技术对学习者的书面案例摘要进行自动分析，以区分学习者的专业水平，从而向其提供相应的反馈（Doleck et al.，2015）。

3.学习者水平评价

对于文本内容的评价，传统的方式费时费力，特别是网络课程中由于参与人数较传统课堂巨幅增长，评价任务更加艰巨。而通过文本挖掘的方法可以实现自动评价。如Azevedo等人基于N-Simple距离图模型，开发了一种自动评估异步论坛讨论相关性的工具，该工具在多数情况下能够识别学生所发内容与讨论主题的相关程度，其结果与教师人工判别结果相似，可以很好地帮助教师评估学生的参与度（Azevedo et al.，2014）。马来西亚麦地那国际大学的研究者使用文本挖掘和自然语言工具包（Natural Language Tool Kit），采用基于本体的信息提取方法，通过对作文的标记化、单词标注、字符计数、频率分布计算以及文本语义匹配等操作实现作文的自动评分（Contreras et al.，2018）。Yu等人采用基于词向量的相似度计算实现了学习者的答题自动化评价（Yu et al.，2017）。张平霞基于布鲁姆的认知分类法构建评价框架，通过朴素贝叶斯分类器实现了对MOOC讨论区中帖子的认知水平自动评价（张平霞，2018）。王漪通过文本的“单向贴近度”和“语义相似度”特征，利用改进的KNN分类算法分别实现了简答题和论述题的自动评分（王漪，2014）。

4.教学材料结构分析

教学材料作为最基本的教育文本数据来源，以往常常只能通过内容分析等方法对其进行研究，但引入文本挖掘的方式后可以从一个更加客观的角度分析其内在的信息。如韩国延世大学的一项研究中，研究者使用狄利克雷多项式回归主题模型分析多个高校机器学习相关课程的教学材料，如课堂讲稿/讲义、相关的辅导论文、作业、答题纸和试卷等，用以发现其主要教学主题的变化趋势，总结教学内容的热点和重点。此外，他们还分析了各个高校的教学材料和出版物，用以探究高校教学情况与科学研究的关系，结果表明两者通常是相互独立的，教学和研究所关注的热点主题并不相同（Lee et al.，2015）。此外，Badawy等人分析了埃及开罗大学统计学院的一本教材，通过RapidMiner将每一章的教学目标与词库进行对比，将章节自动分为：知识与理解、智力技能和专业技能三类，从而为选修章节的学习者提供参考（Badawy et al.，2016）。澳大利亚塔斯马尼亚大学一项计算课程相似性的研究中，研究者分析了多所大学的计算机科学学位所教课程的学科大纲，通过N-Gram关键词抽取的方式，使用基于维基百科语料库的度量方法计算相似度，成功地实现了课程相似程度的自动分析，为学习者的课程选择以及教育部门对课程学分的认证提供了参考（Langan et al.，2016）。

5.学习者反馈

学习者反馈在教学过程中至关重要，不仅能为改善教学及教育系统提供建议，还能为其他学习者选择课程提供参考。如在新加坡管理大学信息系统学院，研究者开发了一个学生反馈挖掘系统，通过收集学生在学期中产生的对于教学和课程的文本反馈意见，进行主题抽取和情感分类，实现对定性反馈的量化分析（Koufakou et al.，2016）。实验结果显示，在情感分类方面，利用对数回归模型训练的分类器，精度可以达到80.1%。在印度尼西亚的AKPRIND科学技术研究所，研究者通过隐马尔可夫词性标记器对收集到的学生问卷中的评论文本进行分析，可以识别出其中包含有针对性意见的评论（Nitin et al.，2015），并通过基于规则的方法，可以确定该意见的类别和所持有的情感态度。其意见检测和意见分类精度均达到95%以上，而情感分类精度也达到80%以上。与此类似的一项针对远程教育课程评价的研究中，研究者提出了一种利用层次分类模型来自动识别学习者意见和情感观点的方法。他们以课程评价网站fernstudiumcheck.de上的评论作为数据来源，训练出了一个多标签分层文本分类器，用以判断学习者的评论所属类别以及持有的情感态度，从而给相关学习者选择课程提供参考，同时也为课程制作者改善课程提供意见（Hamzah et al.，2016）。此外在美国的佛罗里达海湾海岸大学的软件工程系，研究者通过对课程的评价文本数据进行关联规则挖掘和情感分析，可以提取调查中广受关注的关键内容以及学生对课程所持有的情感（Jaskolski et al.，2016）。

6.内容可视化

研究者通过文本数据的交互结构图形来形象地展示信息或者知识，可以帮助人们更加直观地理解相关教育数据。比如，在日本青森县的一项关于高中教师的调查问卷中，Takeyasu等人通过文本挖掘技术抽取关键词并形成关系图，可以让相关教育部门负责人直观地看到高中教师的工作负担情况（Takeyasu et al.，2016）。香港的一项分析学习者学术讨论的研究中，研究者以40个本科生发布的200多个帖子为数据源，使用Forum Graph、R程序集合、LDAvis等工具形成学生互动的社会网络图和主题分布图，可以让管理者了解学生和教师之间的互动，识别出频繁贡献者和被动观察者，从而让教师可以针对性地提供干预，还可以让教师了解现有的和正在增长的讨论主题（Wong et al.，2016a）。此外，Karma等人在研究不丹特殊教育需求儿童的家长与教师的沟通与合作经验时，使用文本挖掘技术分析了26个家长的访谈记录，以概念图、网络云、概念辞典的形式直观地生成了访谈记录的概念和主题（Jigyel et al.，2018）。

**五、教育文本挖掘的挑战与展望**

2017年我国首部国家级人工智能发展规划《新一代人工智能发展规划》正式出台，强调利用智能技术加快推动人才培养模式、教学方法改革，构建包含智能学习、交互式学习的新型教育体系。作为智能教育技术重要组成部分的教育文本挖据技术也日益受到广泛的关注。但受技术、分析方法等制约，教育文本挖据的发展和应用仍然面临诸多挑战。

1.海量教育文本数据的采集与清洗

首先，正如前文所述，教育文本数据来源广泛，涉及到问卷调查、师生互动、学习反馈、在线评论、教学材料、社交内容等。这些文本数据多属于非结构化的模糊教育信息，面临数量化处理难题。其次，教育文本数据是一个高维度的特征数据，如何在进行数据降维的同时保留教育教学的基本特征是教育文本数据预处理的难点。第三，教育文本数据蕴含丰富的师生交流信息，涉及认知、情感、行为等维度，也涉及到教育教学的主体特性，如学习动机、态度、价值观等，如何建立文本数据与教育特征之间的联系一直是其深度应用的难点。

2.多学科研究方法应用和交叉研究

教育文本挖掘中，数据挖掘结果的评估与解释往往是重点和难点。文本挖掘是自然科学领域的方法，而教育领域具有丰富的人文特性和社会属性，单纯的定量分析难以发现文本数据深层次的隐性知识。冰冷的文本数据需要结合教育教学的基本理论，为使用者提供解读的依据，体现出温度。发掘出的知识和提取出的模式需要与教育系统中教师、学习者的情感、状态和认知规律等联系起来，为解读教育教学的真实状态、现象和问题提供可能。因此，研究者需深度融合教育学、认知心理学、语言学等学科研究方法，通过多学科研究方法应用和交叉研究助推教育文本挖掘方法的应用和推广。

3.基于教育情境的多模态分析与验证

由于教育领域的特殊性，教育文本数据来源于不同的应用场景，具有不同的认知、行为和情感特征及其规律。教学场景不同，其教育教学的理论依据也具有差异，反映出的教学规律和特征亦有所不同。教育文本挖掘需要对文本数据进行预处理，以便为后续挖掘提供支撑。文本数据的预处理需要结合教育理论进行编码（标签处理），进而构建合理的分析模型，从而得到可视化的分析结果，为教育现象和问题的解读和阐释提供支持。因此，教育文本挖掘不能陷入唯技术论的误区，而应在教育教学理论指导和教育场景关照基础上利用好这一工具。深度的教育文本挖掘需要对特定的教学情境进行多模态分析和验证，保证教育研究的科学性。

作为一种新兴的研究方法，得益于大数据和自然语言处理技术的飞速发展，教育文本挖掘近年来发展迅速，虽然还没有形成成熟的研究体系，但其给出了教育大数据利用的新视角。文本挖掘技术是数据挖掘的重要组成部分，并随着机器学习方法研究进展而发展，尤其是深度学习方法的突破给文本挖掘技术提供了新的思路。此外，教育文本挖掘方法及其应用尚需要借助自然语言处理的突破而进入真正的应用阶段。随着相关技术的突破和应用发展，教育文本数据必将成为教育现代化发展的推动力，在深度学习、精准教学等领域发挥更大作用。

**参考文献：**

[1]国务院(2015).国务院关于印发促进大数据发展行动纲要的通知[EB/OL]. [2019-04-28].http://www.gov.cn/zhengce/content/2015-09/05/content\_10137.htm.

[2]李纲,毛进(2013).文本图表示模型及其在文本挖掘中的应用[J].情报学报,32(12):1257-1264.

[3]孟小峰,慈祥(2013). 大数据管理:概念、技术与挑战[J]. 计算机研究与发展, 50(1): 146-169.

[4]王漪(2014).文本挖掘技术的研究及其在教学平台中的应用[D]. 北京:北京交通大学.

[5]杨现民,唐斯斯,李冀红(2016). 发展教育大数据:内涵、价值和挑战[J]. 现代远程教育研究, (1): 50-61.

[6]袁在成(2016).可视化工具“词云”在英语教学中的应用研究[J].中国教育学刊, (S1): 102-103.

[7]张平霞(2018). 基于文本挖掘的MOOC讨论区学习评价研究[D]. 重庆:重庆师范大学.

[8]Ananiadou, S., Thompson, P., & Thomas, J. et al. (2010). Supporting the Education Evidence Portal via Text Mining[J]. Philosophical Transactions of the Royal Society A, 368: 3829-3844.

[9]Azevedo, B., Reategui, E., & Behar, A. (2014). Analysis of the Relevance of Posts in Asynchronous Discussions[J]. Interdisciplinary Journal of e-Learning and Learning Objects, 10: 107-121.

[10]Badawy, M., El-Aziz, A., & Hefny, H. (2016). Analysis of Learning Objectives for Higher Education Textbooks Using Text Mining[C]// 2016 12th International Computer Engineering Conference (ICENCO). IEEE: 202-207.

[11]Baker, D., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions[J]. Journal of Educational Data Mining, 1(1): 3-17.

[12]Buenaño-Fernández, D., Villegas-Ch, W., & Luján-Mora, S. (2018). Using Text Mining to Evaluate Student Interaction in Virtual Learning Environments[C]// 2018 IEEE World Engineering Education Conference (EDUNINE). IEEE: 1-6.

[13]Candra Permana, F., Rosmansyah, Y., & Abdullah, A. S.(2017). Naive Bayes as Opinion Classifier to Evaluate Students Satisfaction Based on Student Sentiment in Twitter Social Media[J]. Journal of Physics: Conference Series, 893: 012051.

[14]Chen, X., Vorvoreanu, M., & Madhavan, K. (2014) . Mining Social Media Data for Understanding Students’Learning Experiences[J]. IEEE Transactions on Learning Technologies, 7(3): 246-259.

[15]Cohen, M., & Hersh, R. (2005). A Survey of Current Work in Biomedical Text Mining[J]. Briefings in Bioinformatics, 6(1): 57-71.

[16]Contreras, O., Hilles, S., & Abubakar, B. (2018). Automated Essay Scoring with Ontology Based on Text Mining and NLTK Tools[C]// 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE). IEEE: 1-6.

[17]Cooper, W. (1988). Getting Beyond Boole[J]. Information Processing & Management, 24(3): 243-248.

[18]Doleck, T., Basnet, R., & Poitras, E. et al. (2015). Mining Learner-System Interaction Data: Implications for Modeling Learner Behaviors and Improving Overlay Models[J]. Journal of Computers in Education, 2(4): 421-447.

[19]Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT)[C]// Proceedings of the First International Conference on Knowledge Discovery and Data Mining（KDD’95): 112-117.

[20]Grimes, S.(2008). Unstructured Data and the 80 Percent Rule[EB/OL]. [2020-02-01]. http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/.

[21]Hamzah, A., & Widyastuti, N. (2016). Opinion Extracting and Classification From Questionnaire Comments Using HMM-POS Tagger and Machine Learning Techniques[C]//2016 International Conference on Data and Software Engineering (ICoDSE). IEEE:1-6.

[22]Harvey, S., & Hyndman, B. (2018). An Investigation Into the Reasons Physical Education Professionals Use Twitter[J]. Journal of Teaching in Physical Education , 37(4): 383-396.

[23]Hew, K., Qiao, C., & Tang, Y. (2018). Understanding Student Engagement in Large-Scale Open Online Courses: A Machine Learning Facilitated Analysis of Student’s Reflections in 18 Highly Rated MOOCs[J]. International Review of Research in Open and Distributed Learning, 19(3): 69-93.

[24]IBM Corporation(2019). About IBM SPSS Modeler Text Analytics[EB/OL]. [2019-05-15]. https://www.ibm.com/support/knowledgecenter/en/SS3RA7\_15.0.0/com.ibm.spss.ta.help/tmfc\_intro.htm.

[25]Jaskolski, J., Siegberg, F., & Tibroni, T. et al. (2016). Opinion Mining in Online Reviews About Distance Education Programs[EB/OL].  [2020-02-01]. https://arxiv.org/abs/1607.06299.

[26]Jigyel, K., Miller, J., & Mavropoulou, S. et al. (2018). Parental Communication and Collaboration in Schools With Special Educational Needs (SEN) Programmes in Bhutan[J]. International Journal of Inclusive Education, 22(12): 1288-1305.

[27]Kiritchenko, S., & Matwin, S. (2011). Email Classification With Co-Training[C]// Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research. IBM Corp: 301-312.

[28]Koufakou, A., Gosselin, J., & Guo, D. (2016). Using Data Mining to Extract Knowledge From Student Evaluation Comments in Undergraduate Courses[C]// 2016 International Joint Conference on Neural Networks (IJCNN). IEEE:3138-3142.

[29]Langan, G., Montgomery, J., & Garg, S. (2016). Similarity Matching of Computer Science Unit Outlines in Higher Education[C]// Australasian Joint Conference on Artificial Intelligence. Springer, Cham: 150-162.

[30]Lee, H., Kwak, J., & Song, M. et al. (2015). Coherence Analysis of Research and Education Using Topic Modeling[J]. Scientometrics, 102(2): 1119-1137.

[31]Liu, Z., & Zhang, Y. (2018). A Semantic Role Mining and Learning Performance Prediction Method in MOOCs[C]//Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. Springer, Cham:259-269.

[32]Merton, R., & Barber, E. (2011). The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology Of Science[M]. Princeton University Press.

[33]Nitin, G., Swapna, G., & Shankararaman, V. (2015). Analyzing Educational Comments for Topics and Sentiments: A Text Analytics Approach[C]// 2015 IEEE Frontiers in Education Conference (FIE). IEEE: 1-9.

[34]Salton, G., Wong, A., & Yang, C. S.(1975). A Vector Space Model for Automatic Indexing[J]. Communications of the ACM, 18(11): 613-620.

[35]Stupans, I., Mcguren, T., & Babey, A. (2016). Student Evaluation of Teaching: A Study Exploring Student Rating Instrument Free-form Text Comments[J]. Innovative Higher Education, 41(1): 33-42.

[36]Sullivan, D. (2001). Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales[M]. John Wiley & Sons, Inc.

[37]Takeyasu, K., Oyanagi, T., & Ishii, Y. et al. (2016). Text Mining Analysis on the Questionnaire Investigation for High School Teachers’Work Load[C]// 2016 International Conference on Industrial Engineering, Management Science and Application (ICIMSA). IEEE:1-5.

[38]Tan, A. (1999). Text Mining: The State of the Art and the Challenges[C]// Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases: 65-70.

[39]Tan, P. (2018). Introduction to Data Mining[M]. Pearson Education India.

[40]Tseng, Y., Lin, C., & Lin, Y. (2007). Text Mining Techniques for Patent Analysis[J]. Information Processing & Management, 43(5): 1216-1247.

[41]Velasquez, N., Fields, D., & Olsen, D. et al. (2014). Novice Programmers Talking About Projects: What Automated Text Analysis Reveals About Online Scratch Users’Comments[C]// 2014 47th Hawaii International Conference on System Sciences. IEEE: 1635-1644.

[42]Verma, A., Rathore, S., & Vishwakarma, S. et al. (2017). Multilevel Analysis of Student’s Feedbackusing Moodle Logs in Virtual Cloud Environment[J]. International Journal of Computer Science & Information Technology, 9(5): 15-28.

[43]Witten, I., Frank, E., & Hall, M. et al. (2016). Data Mining: Practical Machine Learning Tools and Techniques[M]. Morgan Kaufmann.

[44]Wong, G., Li, S., & Wong, E. (2016a). Analyzing Academic Discussion Forum Data With Topic Detection and Data Visualization[C]// 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE). IEEE: 109-115.

[45]Wong, G., & Li, S. (2016b). Academic Performance Prediction Using Chance Discovery From Online Discussion Forums[C]// 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC):706-711.

[46]Yu, F., & Zheng, D. (2017). Education Data Mining: How to Mine Interactive Text in Moocs Using Natural Language Process[C]// 2017 12th International Conference on Computer Science and Education (ICCSE). IEEE:694-699.

收稿日期　2020-05-17　责任编辑　汪燕

