计算机科学
COMPUTER SCIENCE

基于Bloom分类法的CS1试题数据集的构建及其自动分类

董荣胜，卫晨雨，胡杰，乔宇澄，李凤英

引用本文

董荣胜，卫晨雨，胡杰，乔宇澄，李凤英基于Bloom分类法的CS1试题数据集的构建及其自动分类[J]．计算机科学, 2023, 50(6): 175-182.

DONG Rongsheng, WEI Chenyu, HU Jie, QIAO Yucheng, LI Fengying. Construction and Automatic Classification of CS1 Test Questions Dataset Based on Bloom's Taxonomy [J]. Computer Science, 2023, 50(6): 175-182.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200182

基于 Bloom 分类法的 CS1 试题数据集的构建及其自动分类

董荣胜　卫晨雨　胡　杰　乔宇澄　李凤英
桂林电子科技大学广西可信软件重点实验室　广西 桂林 541004
(ccrsdong@guet.edu.cn)

摘　要　课程评估是教学改革的一个关键环节,涉及教学案例、试题以及课堂教学等方面的内容。针对计算课程的试题评估,引入 Bloom 分类法,以普林斯顿大学和桂林电子科技大学"计算机科学导论"课程(CS1)的试题为语料库,给出针对 CS1 的 Bloom 分类法认知过程维度和知识维度的相应动词种子库和名词种子库,对试题所能达到的 Bloom 分类法二维矩阵的位置进行标注,构建 CS1 试题分类数据集。采用机器学习技术,给出 CS1 试题自动分类模型 TFERNIE-LR,该模型由 CSTFPOS-IDF 算法、ERNIE 模型和 LR 分类器 3 部分组成。CSTFPOS-IDF 算法是在 TFPOS-IDF 算法的基础上,通过计算课程关键词权重因子,来提高模型对计算课程关键词的关注程度,生成词权重。同时,基于实体知识增强预训练模型 ERNIE 进行试题词语级向量嵌入,组合词权重和词语级向量生成用于自动分类的试题文本向量。最后,采用 LR 分类器将试题自动分类到 Bloom 分类法二维矩阵。实验结果表明,TFERNIE-LR 模型具有良好的性能,在认知过程维度和知识维度上的加权精确率分别达到了 83.3%和 96.1%。

关键词:Bloom 分类法;课程评估;CS1 试题分类数据集;动词种子库;名词种子库;自动分类
中图法分类号　TP391

Construction and Automatic Classification of CS1 Test Questions Dataset Based on Bloom's Taxonomy

DONG Rongsheng, WEI Chenyu, HU Jie, QIAO Yucheng and LI Fengying
Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China

Abstract　Curriculum evaluation is a key link of teaching reform, which involves the evaluation of teaching cases, test questions and classroom teaching. In order to evaluate the test questions of computing courses, this paper introduces Bloom's taxonomy, and takes the test questions of "Introduction to Computer Science" course(CS1) of Princeton University and Guilin University of Electronic Science and Technology as corpus, and the corresponding verb seed bank and noun seed bank for the cognitive process dimension and knowledge dimension of Bloom's taxonomy for CS1 are given, the positions of the two-dimensional matrix of Bloom's taxonomy that could be reached by the test questions are manually labeled, classification dataset for CS1 test questions is constructed. Machine learning technology is used, the automatic classification model TFERNIE-LR of CS1 test questions is given, which is composed of CSTFPOS-IDF algorithm, ERNIE model and LR classifier. CSTFPOS-IDF algorithm is based on TFPOS-IDF algorithm, by the weight factor of the keywords in computing discipline, CSTFPOS-IDF algorithm pays more attention to the keywords improves and generates the weight of words. At the same time, the entity knowledge enhanced pre-training model ERNIE is used to embed the word level vector of test questions, and the combined word weight and word level vector are used to generate the text vector of test questions for automatic classification. Finally, the LR classifier is used to automatically classify test questions into Bloom's taxonomy two-dimensional matrix. Experimental results show that the proposed TFERNIE-LR model has good performance, and weighted-P in the cognitive process dimension and knowledge dimension reaches 83.3% and 96.1% respectively.

Keywords　Bloom's taxonomy, Curriculum evaluation, Classification dataset for CS1 test questions, Verb seed bank, Noun seed bank, Automatic classification

1　引言

课程评估是教学改革的一个关键环节,对课程直接进行评估难度较大,涉及课程教学案例、试题、课堂教学等方面的内容。本文关注的是课程试题评估,以普林斯顿大学和桂林电子科技大学"计算机科学导论"课程(简称 CS1)的试题为语料,采用 ACM 和 IEEE-CS 提交的 CC2020 报告[1]推荐的 Bloom 分类法[2-3]进行评估。该分类法将知识维度分为事实

到稿日期:2023-02-24　返修日期:2023-04-17
基金项目:国家自然科学基金(62062029)
This work was supported by the National Natural Science Foundation of China(62062029).
通信作者:李凤英(lfy@guet.edu.cn)

性知识、概念性知识、程序性知识、元认知知识4个维度，将认知过程维度分为记忆、理解、应用、分析、评估、创造6个维度。同时，本文基于知识增强的语义表示模型(Enhanced Language Representation with Informative Entities，ERNIE)[4]和改进的带有词性特征的TF-IDF加权算法(TF-IDF with Part-of-Speech for CS1，CSTFPOS-IDF)，结合逻辑回归分类器(Logistic Regression，LR)，提出CS1试题自动分类模型(Logistic Regression combining CSTFPOS-IDF and ERNIE，TFERNIE-LR)，通过多次对比实验验证了所提模型的有效性。

本文的主要工作如下：

(1)基于Bloom分类法，结合高质量的CS1试题数据语料，给出针对CS1的基于Bloom分类法的认知过程维度和知识维度的相应动词和名词种子库，构建CS1试题分类数据集。

(2)提出了CS1试题自动分类模型TFERNIE-LR，该模型基于ERNIE模型和CSTFPOS-IDF算法提取实体知识特征和词性特征并整合，生成试题文本向量，采用LR分类器对试题文本向量进行分类。

2 相关工作

为使机器更准确地赋予试题对应的Bloom分类法标签，辅助教育工作者评测试题，研究者们分析了试题数据的特征，采用的分类技术主要有两种：基于规则的方法和基于机器学习(数据驱动)的方法。

基于规则的方法是通过设定一组适用的分类规则，将待分类的数据与规则进行匹配，实现数据分类。Chang等[5]在其在线测试系统上使用了预定义的关键字列表，根据关键字列表检查问题中的动词，这种方法只能有效分类记忆层次的问题。Omar等[6]基于规则的方法和自然语言处理技术对认知过程维度的6个不同维度层次进行分类，通过词权重解决Bloom分类法中属于多个层次的关键动词重叠问题，然而该方法需要人工确定某一类别下特定动词的权重且主观性较强。Haris等[7]使用基于规则的*N*-gram方法以增强分类结果，该方法的优点是在问题未能按规则分类的情况下，可以利用统计分类器辅助分类。Jayakodi等[8]提出WordNet相似度算法，该算法由标记模式生成模块、语法生成模块、解析器生成模块和余弦相似度检测模块组成，可以有效分类没有动词的文本。尽管基于规则的方法不需要大量的训练数据且推理过程直观，但设计规则有一定难度，且无法推理未编码的条件和规则。

基于机器学习的方法可以使计算机模拟人类的判别行为，从历史经验(试题训练集)中总结规律(自动分类模型)，将规律应用到新的场景(试题测试集)。Fei等[9]探索可用于试题自动分类的电子学习系统，工作中构建了包含233个选择题的数据集，把题目分为难、中、易3个级别，使用人工神经网络训练的反向传播学习算法作为文本分类器。Yusof等[10]利用不同特征集(全特征、文档频率和类别频率-文档频率)的神经网络对问题进行分类，并利用比例共轭梯度学习算法对模型进行训练。为提高神经网络在高维输入空间上的可扩展性，文中采用了不同的特征降维方法。结果表明，文档降频方法最有效，其在保证分类精度的同时加快了收敛速度。Yahya等[11]根据预定义的标准(如Bloom分类法的认知过程维度)对问题进行注释，每个问题都通过删除标点和停用词、分词、词干提取、词加权和长度标准化来处理。结果表明，Support Vector Machine在分类准确率和精确率方面具有良好的效果。Abduljabbar等[12]提出了一种基于投票算法的组合策略，其将3种机器学习分类器(Support Vector Machine，Naïve Bayes，*K*-Nearest Neighbour)和3个特征选择方法(Chi-Square，Mutual Information，Odd Ratio)相组合对问题进行分类，结果表明，组合分类器的分类效果优于单独分类器。

近年来，在GPU算力和海量无标注文本数据的双重支持下，预训练模型打开了深度学习模型规模与性能齐飞的局面。其中预训练语言模型具有强大的语义表示能力，它学习深度神经网络的高效参数，在自然语言处理研究中被广泛用作特征提取器。Mohammed等[13]采用修订前Bloom分类法的认知过程维度进行自动分类研究，使用预训练模型Word2Vec和一个改进的TF-IDF算法TFPOS-IDF从试题中提取特征，提取的特征被输入到3种不同的分类器，即Support Vector Machine，*K*-Nearest Neighbour和Logistic Regression中。在实验中，他们使用了两组数据集，第一组数据集包含600个开放式英文问题，平均文本长度为11个单词，涉及生物、计算机、文学、数学等领域；第二组有141个英文问题，平均文本长度为10个单词。其提出的分类模型在短文本分类上取得了突出的效果，但对于长文本的分类缺乏实践；构建的数据集依赖关键词，模型存在迁移能力弱的问题；由于Word2Vec的词和向量是一对一的关系，因此无法解决多义词的词向量表示问题。

综上所述，为了推动缺乏高质量标注数据而受阻的试题分类研究，本文选用桂林电子科技大学国家一流课程“计算机科学导论”[14]，以及2020年由Gong等译、机械工业出版社出版的普林斯顿大学罗伯特·塞奇威克和凯文·韦恩撰写的教材[15]中的试题作为数据源，构建了由2 611道试题组成的CS1试题分类数据集。在试题自动分类研究中，采用动态的知识增强语义表示模型ERNIE进行词向量嵌入，通过CSTFPOS-IDF算法对CS1关键词赋予高优先级，并提取文本词性特征，提出了试题自动分类模型TFERNIE-LR。实验结果表明，TFERNIE-LR模型在认知过程维度上的加权精确率和加权F1值分别达到了83.3%和83.1%，在知识维度上的加权精确率和加权F1值分别达到了96.1%和96.0%。

3 CS1试题分类数据集构建

为构建高质量的CS1试题分类数据集，依据Bloom分类法认知过程维度6个层次和知识维度4个层次的细化准则，结合CS1试题数据调整二维表中组成要素，给出针对认知过程维度和知识维度的动词种子库和名词种子库，使试题标注更加准确。确定试题在Bloom分类法认知过程维度和知识维度二维矩阵的位置，构建了CS1试题分类数据集，数据集构建过程包括数据采集、数据标注和数据集分析3个阶段。

3.1 CS1试题数据采集

试题数据采集选自桂林电子科技大学国家一流课程“计算机科学导论”[14]的16套期末试卷、教材习题、MOOC题库，

以及普林斯顿大学罗伯特·塞奇威克和凯文·韦恩撰写的教材《计算机科学导论:跨学科方法》[15]中的习题,总计 2611 道题目数据。

3.2 CS1 试题数据标注

数据整理工作由作者所在研究室的 6 名经过专业训练的研究生和 2 名课程专家担任标注员,先进行一轮试标注与讨论,然后 6 名标注员依据标注规范独立地进行标注。对于标注不一致的情况,由课程专家进行仲裁。每个标注人员根据分类标准将试题文本的 6 个认知过程维度分别标注为 0,1,2,3,4,5,将 4 个知识维度分别标注为 0,1,2,3,以单个试题文本为单位进行标注。若认知过程维度或知识维度出现多个层级,只标注级别最高的层级。

试题的认知过程维度分类标准如表 1 所列,表中统计了标注过程中的关键动词,这些动词构成了 CS1 认知过程维度的动词种子库,种子库将随着数据集的不断扩展而扩展。

Bloom 分类法的认知过程维度细化准则如表 2 所列,各层次的子类多是动词描述的,把试题中的动词和各子类及其近义词进行比较有助于提高分类的准确性。

表 1 基于 Bloom 分类法的 CS1 认知过程维度动词种子库

Table 1 CS1 cognitive process dimension verb seed bank based on Bloom's taxonomy

级别	定义	动词
记忆	回忆数据、信息或特定事物,回忆某些术语的定义	简述、组成、定义、记忆、给出
理解	能用实例对定义和定义的内容进行说明、转译或补充	理解、了解、解释、说明、举例、描述
应用	能用学科的基础概念,如迭代和递归编写算法	采用、使用、利用、应用、根据
分析	将材料或概念分解成各个组成部分,以便理解其结构,目的在于理清信息	区别、分析、解析、比较、讨论
评估	基于明确的准则或标准,对想法、方法或材料的价值及正确性做出判断,特别是对算法的时间复杂度和空间复杂度进行评估	评估、评价、检查、证明、验证、时间复杂度、空间复杂度
创造	将要素组成内在一致的整体或功能性系统,将要素组织成新的模型或体系	设计、构建、创造、提出、创建

表 2 Bloom 分类法的认知过程维度细化准则

Table 2 Cognitive process dimension refinement criteria for Bloom's taxonomy

认知过程维度类别	认知过程维度子类别	近义词	定义
1.记忆(Remember)	1.1 识别(Recognizing)	辨认(Identifying)	在长期记忆中查到与呈现材料相吻合的知识
	1.2 回忆(Recalling)	提取(Retrieving)	
2.理解(Understand)	2.1 解释(Interpreting)	澄清(Clarifying) 释义(Paraphrasing) 描述(Representing)	理解知识,用各种方式描述呈现的信息或概括信息
	2.2 举例(Exemplifying)	示例(Illustrating) 实例化(Instantiating)	找到一个概念或一条原理的具体例子或例证
	2.3 推断(Inferring)	断定(Concluding) 预测(Predicting)	从提供的信息中得出合乎逻辑的结论
3.应用(Apply)	3.1 执行(Executing)	实行(Carrying out)	将一程序应用于熟悉的任务
	3.2 实施(Implementing)	使用,运用(Using)	将一程序运用于新的任务
4.分析(Analyze)	4.1 区别(Differentiating)	辨别(Discriminating) 区分(Distinguishing) 选择(Selecting)	区分呈现材料的相关与无关部分或重要与次要部分
	4.2 归因(Attributing)	解构(Deconstructing)	确定呈现材料背后的观点、倾向、价值或意图
5.评估(Evaluate)	5.1 检查(Checking)	检验(Testing) 证明(Demonstrating)	发现一个过程或产品内部的矛盾和谬误;确定一个过程或产品是否具有内部一致性;查明程序实施的有效性
	5.2 评论(Critiquing)	评判(Judging)	发现一个产品与外部准则之间的矛盾;确定一个产品是否具有外部一致性;查明程序对一个给定问题的恰当性
6.创造(Create)	6.1 计划(Planning)	设计(Designing) 构造(Constructing)	为完成某一任务设计的系统

针对 CS1 试题数据,在 Bloom 分类法的评估层次增加"计算复杂性"子类,计算复杂性包括算法的时间复杂性和空间复杂性。知识维度分类标准及名词种子库如表 3 所列。其中概念性知识指整体中基本要素的含义,如抽象的含义、递归的概念;程序性知识指解决问题的方法或算法的步骤,如求解汉诺塔的递归算法、求解 π 的蒙特卡洛方法。

表 3 基于 Bloom 分类法的 CS1 知识维度名词种子库

Table 3 CS1 knowledge dimension noun seed bank based on Bloom's taxonomy

知识维度	定义	名词
事实性知识	学生为了掌握特定学科知识或解决问题而需要了解的基本事实	符号、数字、字母、事件、地点、人物、时间
概念性知识	一个整体结构中基本要素之间的关系	概念、定义、含义
程序性知识	做事的方法,探究的方法,应用技能、算法、技术或方法的规范等	算法、程序、公式、代码、函数、树、图、表达式、队列、栈、链表
元认知知识	关于一般认知、个体特定认知的知识	形式模型、概念模型、E-R 图、图灵机、公理体系、充分关系、必要关系、证比求易算法(Verifying is easier than finding solutions)

3.3 CS1 试题分类数据集分析

本文最终标注了 2611 条数据，对 CS1 试题分类数据集中文本的长度进行统计。以 50 为步长，设置不同的阈值，分析 CS1 试题分类数据集样本文本长度小于阈值的样本数量在总数据集中的占比情况，结果如表 4 所列。数据集中样本文本长度在 0～100 范围内的数据约占总数据集的 44.87%，随着长度阈值的不断增大，小于阈值的数据占比提升愈加缓慢。当阈值设置为 500 时，样本文本长度小于 500 的数据占总数据集的 99.47%；再增大长度阈值，比例已经无法得到明显的提高。因此，通过上述的分析，本研究中规定文本的最大输入长度为 510，长度大于 510 的样本则保存前 510 个字符。

表 4 不同长度阈值下 CS1 试题分类数据集文本长度占比情况

Table 4 Proportion of text length in classification dataset for CS1 test questions with different length thresholds

长度	100	150	200	250	300	350	400	450	500
占比/%	44.87	63.42	75.43	83.55	89.62	92.43	94.65	96.75	99.47

CS1 试题在 Bloom 分类法二维矩阵的分布情况如表 5 所列，在知识维度，处于程序性知识的试题最多，占总数的 64.69%，其次是处于概念性知识的试题，占 25.74%。由于事实性知识中的符号、数字、字母、事件、地点、人物、时间及其他细节已蕴含在 CS1 试题其他知识之中，故将事实性知识这一行内容删除。在认知过程维度，处于理解层次的试题最多，占 30.95%；评估层次的试题最少，占 4.67%。

表 5 CS1 试题认知过程维度和知识维度分布情况

Table 5 Distribution of cognitive process dimension and knowledge dimension of CS1 test questions

知识维度	认知过程维度						总计	百分比/%
	记忆	理解	应用	分析	评估	创造		
概念性知识	160	343	8	154	7	0	672	25.74
程序性知识	1	360	391	450	102	385	1689	64.69
元认知知识	1	105	76	17	13	38	250	9.57
总计	162	808	475	621	122	423	2611	100.00
百分比/%	6.20	30.95	18.19	23.78	4.67	16.20	100.00	

4 文本数据预处理

为解决试题文本中存在的特征空间高维性、特征分布稀疏等问题，首先对试题数据进行预处理，主要过程包括文本标准化、去掉停用词、文本分词和词性标注，如图 1 所示。

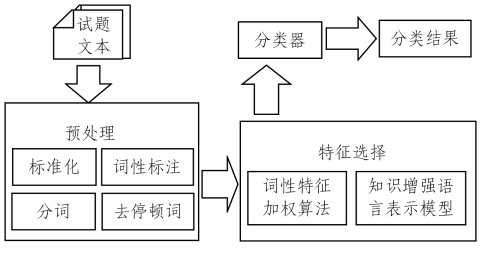

图 1 试题自动分类流程图

Fig. 1 Flow chart of automatic classification of test questions

标准化过程将消除不需要的数据，如标点符号、数字和英语字符。此外，使用哈工大停用词表去除停用词。在这一步中，并不是将所有的停用词都从试题中删除，因为一些重要的停用词对试题分类有较大影响。

对试题文本标准化后，利用 jieba 分词工具进行分词，以空格分隔表示。

词性标注即标注文本分词后每个词的词性，使用 jieba 分词工具进行词性标注。

5 TFERNIE-LR 试题自动分类模型

教师基于 Bloom 分类法评估计算课程教学目标时存在理解不一致，且对试题的教学目标评估耗时耗力等问题。针对此问题，引入机器学习技术为课程评估工作提供技术支撑，提出 TFERNIE-LR 试题自动分类模型，对试题进行基于 Bloom 分类法的自动分类。TFERNIE-LR 试题自动分类模型整体架构如图 2 所示，主要由 3 部分组成，分别是文本输入层、特征提取层和输出层。

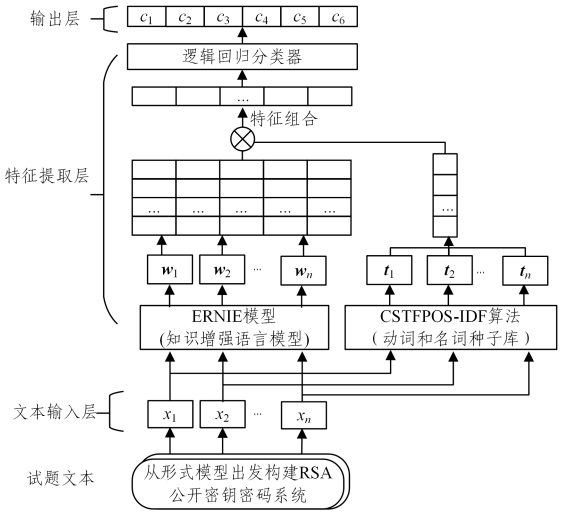

图 2 TFERNIE-LR 试题自动分类模型

Fig. 2 Automatic classification model TFERNIE-LR of CS1 test questions

5.1 CS1 试题文本特征提取

特征提取是本文模型对试题文本信息进行提取过滤的环节，具体过程如下：基于知识增强语义表示模型 ERNIE 进行词向量嵌入，利用实体的先验知识增强词向量的语义信息，得到词向量；结合 CS1 认知过程维度和知识维度的相应动词和名词种子库，对 TFPOS-IDF 算法进行改进，给出 CSTFPOS-IDF 算法，提高种子库中的关键动词和名词的权重因子，得到词权重；将试题文本词向量和词权重相结合，生成试题文本向量。

5.1.1 ERNIE 模型

ERNIE 模型[4]是一类知识增强语义理解模型，主要功能是与输入文本进行交互，生成文本的词向量表示。针对 BERT[16]在处理中文文本时难以获得语义完整表示的缺点，ERNIE 将知识模型中的实体表征整合到语义模型的底层中，结合大规模无监督语料库和知识图谱进行预训练，抽取和编码知识信息。ERNIE 改进了 BERT 中的 mask 机制，mask 机制通过掩盖词单元并利用文本上下文预测掩盖的词单元，以此获取掩盖的词单元的语义表示。主要增加了两种新的 mask 策略，一种是基于词组的 mask 策略，另一种是基于实体的 mask 策略。如图 3 所示，当输入的句子为“图灵机有什么

特点”时，ERNIE 会将“图灵机”3 个字当成一个单元进行 mask，通过这种方式，在训练过程中隐含地学习实体的先验知识，这种方式使模型学习到了更可靠的语言表达。

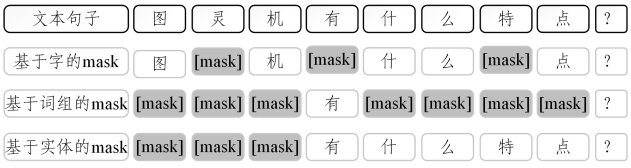

图 3 ERNIE 的 mask 策略

Fig. 3 mask strategy of ERNIE

ERNIE 模型采用 transformer 的编码部分作为其语义表示的骨架，整体结构如图 4 所示，由文本编码层和知识编码层组成。文本编码层由多头注意力机制层[17]和前向传播网络层构成，其中多头注意力机制层用来输入语料中各个字词之间的关系，能够捕捉长距离的依赖信息，更完整地表达出文本的实际含义。

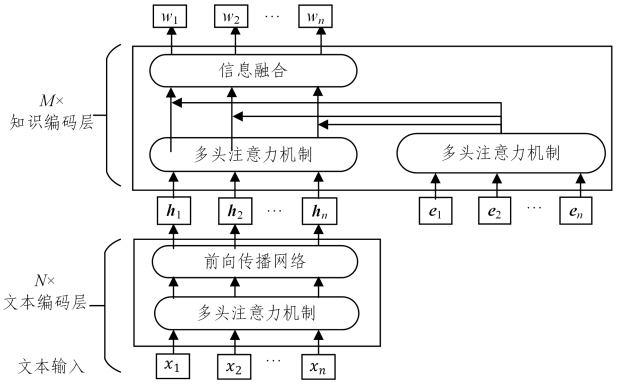

图 4 ERNIE 模型结构

Fig. 4 Structure of ERNIE model

为了更好地获取 CS1 试题中字词在多种特定语义场景下与其上下文构造的语义信息，引入了注意力机制，公式如下：

$$Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别表示查询向量、键向量和值向量，d_k 表示向量 $\boldsymbol{K}$ 的维度。

通过 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 点积计算试题字词相似度得到权重，同时，为降低向量 $\boldsymbol{K}$ 的敏感度，提高网络训练时的稳定性，使用 $\boldsymbol{K}$ 的维度 d_k 进行缩放。其次使用 softmax 函数对权重进行归一化得到概率分布。最后，将权重与相应的键值 $\boldsymbol{V}$ 进行加权求和得到目标的注意力向量。将 $Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$ 输入前向传播网络层，计算式如下：

$$\boldsymbol{H}_1=\boldsymbol{W}\cdot Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+b \tag{2}$$

其中，$\boldsymbol{H}_1$ 为文本编码层的输出，$\boldsymbol{W}$ 为权重矩阵，b 为偏置量。为了捕获试题的内部相关性，减少对外部信息的依赖，注意力机制中 $\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}$。

每道试题可以被表示为 $\boldsymbol{X}_i=\{x_1,x_2,\cdots,x_j,\cdots,x_n\}$，$x_j$ 表示为文本中第 j 个词。利用多头注意力机制并行地从输入中获取多组信息，得到多组 $\boldsymbol{H}_i(i=1,2,\cdots,t)$，$t$ 表示有 t 个查询、键值对。然后将多组 $\boldsymbol{H}_i$ 进行拼接，得到文本编码层的输出 $\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_j,\cdots,\boldsymbol{h}_n\}$，计算公式如下：

$$\boldsymbol{H}=Concat(\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_t)\boldsymbol{W}_0 \tag{3}$$

其中，$\boldsymbol{W}_0$ 是附加权重矩阵，作用是将拼接后的矩阵维度压缩成固定的文本长度大小。

经过文本上下文信息提取后，词 x_j 被表示为词向量 $\boldsymbol{h}_j$，词向量维度为 768。知识编码层融入实体的知识信息，$\boldsymbol{E}=\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_j,\cdots,\boldsymbol{e}_m\}$ 为知识实体输入，经多头注意力机制层和 $\boldsymbol{H}$ 进行知识融合，最后输出 $\boldsymbol{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_j,\cdots,\boldsymbol{w}_n\}$。知识融合后，词 x_j 被进一步表示为词向量 $\boldsymbol{w}_j$，词向量维度为 768。模型在知识编码层学习到更多语义联系，如实体类别、实体关系等，使得模型获得的向量与上下文信息更密切相关。$\boldsymbol{X}_i$ 和 $\boldsymbol{W}$ 在形式上相同，但是 ERNIE 模型已经将输入文本中的各个字向量转换为相同长度的增强语义向量。

5.1.2 CSTFPOS-IDF 算法

动词和名词对试题认知过程维度和知识维度自动分类起着重要作用，本文结合认知过程维度和知识维度的动词和名词种子库改进 TFPOS-IDF 加权算法，提出 CSTFPOS-IDF 算法，提高计算课程关键词关注程度，核心思想是字词的重要性随着它在文档中出现的次数与词性权重的积成正比增加，但同时也会随着它在语料库中出现的频率成反比下降。CSTFPOS-IDF 的计算式如下：

$$w_{\mathrm{pos}}(t)=\begin{cases}w_1, & \text{if } t \text{ is in the bank}\\ w_2, & \text{if } t \text{ is } VB, \text{not in the bank}\\ w_3, & \text{if } t \text{ is } NN, \text{not in the bank}\\ w_4, & \text{otherwise}\end{cases} \tag{4}$$

$$TFPOS(t,d)=\frac{c(t,d)*W_{\mathrm{pos}}(t)}{\sum_i(c(t_i,d)*W_{\mathrm{pos}}(t_i))} \tag{5}$$

$$IDF(t)=1+\log\left(\frac{D}{d_t}\right) \tag{6}$$

$$CSTFPOS-IDF(t,d)=TFPOS(t,d)*IDF(t) \tag{7}$$

其中，$c(t,d)$ 表示词 t 在文档 d 中出现次数，D 为数据集中出现的文档总数，d_t 为出现词 t 的文档数，w_1 表示种子库中词的权重因子，w_2，w_3，w_4 分别是词不同词性下的权重因子。

在认知过程维度试题自动分类中赋予关键词权重因子 7、动词权重因子 4、名词权重因子 2、其他词性权重因子 1，例如，试题“从形式模型出发构建 RSA 公开密钥密码系统”中“构建”为动词且存在于动词种子库中，赋予权重 7，经过 CSTFPOS-IDF 算法后获得 CSTFPOS-IDF（“构建”）= 0.597 073 129 316 630 3，相比其他词获得了模型更多的关注。同时，ERNIE 预训练模型得到字词的语义向量，通过语义向量和词性权重组合得到试题文本的向量表示，最终词向量计算式如下：

$$z_j=\boldsymbol{w}_j*CSTFPOS-IDF(x_j,d) \tag{8}$$

其中，x_j 表示试题文本分词后的词，$\boldsymbol{w}_j$ 表示 x_j 经 ERNIE 模型得到的词向量，d 表示试题文档。

5.2 CS1 试题自动分类

通过 ERNIE 语言模型和 CSTFPOS-IDF 算法得到试题文本向量，将文本向量输入多个不同的分类器 K-Nearest Neighbour，Logistic Regression 和 Support Vector Machine 中进行分类。

5.2.1 K 近邻分类

K 近邻分类(K-Nearest Neighbor, KNN)是模式识别领域中著名的分类模型,其基本思路是:

(1)将试题训练集中的文档转化为向量;

(2)计算需要分类的试题文档与训练集中每一个试题文档的近似程度大小,并降序排列;

(3)事先定义近邻个数 K,取出排列后排名前 K 的试题文档;

(4)需要分类的试题文档的最终标签便是排在前面 K 位的众数类别。

记待分类试题为$\hat{d}$,与其最近邻的 K 个试题集合为$\{d_i, i=1,2,\cdots,k\}$,试题d_i是否属于C_j用指示函数表示为:

$$I_j(d_i)=\begin{cases}1, & d_i\in C_j\\ 0, & d_i\notin C_j\end{cases}, j=1,2,\cdots,m, i=1,2,\cdots,k \tag{9}$$

使用 K 近邻分类方法做最终类别的判定的基本思路是以最近邻 K 个试题中的众数类别作为需要划分类别的文档的类,即多数表决法:

$$\hat{d}\in\{C_j, j=\arg\max_j \sum_{i=1}^{k} I(d_i\in C_j)\} \tag{10}$$

5.2.2 支持向量机

支持向量机(Support Vector Machine, SVM)理论是由 Corte 等[18]于 1995 年提出的。早期 SVM 是一个针对二分类问题的线性分类器,基本思想如下:

给定 CS1 试题分类数据集$\{(x_i, y_i), i=1,2,\cdots,m\}$,$x_i\in D$,D 为文档集合,$y_i\in\{-1,+1\}$表示$x_i$所属的类别。分类思想是找到一个超平面 $\boldsymbol{\omega}^{\mathrm{T}}x+b=0$,将试题训练集数据按不同类别进行分类,超平面的分类准则为:

$$f(x)=\mathrm{sign}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b)=\begin{cases}-1, & \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b>0\\ +1, & \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b<0\end{cases} \tag{11}$$

SVM 旨在追求找到一个超平面来将两类数据最大程度地分开,故 SVM 大多适用于解决二分类问题。要将 SVM 推广应用于解决试题认知过程维度六分类问题,就需结合多分类策略,常见的多分类策略有一对一、一对其余和多对多。文中利用一对一策略进行试题认知过程维度六分类。一对一策略的思路是在任意两类样本之间构造一个分类超平面,将一个六分类问题转化成 15 个二分类问题,构造出 15 个最优化问题,求解得到 15 个分类超平面。

5.2.3 逻辑回归分类器

逻辑回归分类器(Logistic Regression, LR)[19]基于 Logistic 函数,该函数被广泛应用于文本分类任务,并取得了显著的效果。其背后的思想是找出特征和特定输出之间的关系。Logistic 回归算法采用一对一的方法来处理多分类任务。例如,为了预测试题认知过程维度层级,需要考虑 6 个二元分类问题,即所有类是否为记忆、理解等。在此之后,使用极大似然估计对预测类进行赋值,其实现方法如下:

$$P(c|x)=\frac{\mathrm{e}^{\left(\sum_{i=1}^{N} f_i(c,x)w_i\right)}}{\sum_{c'\in C}\mathrm{e}^{\left(\sum_{i=1}^{N} f_i(c',x)w_i\right)}} \tag{12}$$

目标损失函数为极大似然估计,如下:

$$J(\boldsymbol{w},b)=-\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\log h(x^{(i)})+(1-y^{(i)})\log(1-h(x^{(i)}))\right] \tag{13}$$

其中,m 表示样本总数,$x^{(i)}$表示第 i 个样本,$y^{(i)}$表示第 i 个样本的真实标签,$h(x^{(i)})$表示将第 i 个样本预测为正类的概率。

CS1 试题分类数据集在知识维度上存在数据不平衡问题,改进的思路是修改损失函数,增加类别权重,缓解其对模型训练的影响,损失函数公式如下:

$$J(\boldsymbol{w},b)=-\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\log h(x^{(i)})*w_1+(1-y^{(i)})\log(1-h(x^{(i)}))*w_2\right] \tag{14}$$

$$w_i=\frac{N}{c*n_i} \tag{15}$$

其中,N 表示试题总数,c 表示类别数,n_i表示类别 i 下的试题数目,w_i表示类别 i 的权重。

得到试题文本的向量表示之后,将其输入逻辑回归分类器中,输出每个类别的概率,得到最终分类结果。对于 Bloom 分类法的认知过程维度,逻辑回归分类器将试题分为记忆、理解、应用、分析、评估、创造 6 个类别。对于知识维度,逻辑回归分类器将试题分为事实性知识、概念性知识、程序性知识、元认知知识 4 个类别。

6 实验设计与结果分析

6.1 度量标准

试题认知过程维度和知识维度分类属于多分类问题,为了计算分类模型在不同类别上的精确率、召回率和 F1 值,本文选用加权精确率(*weighted-P*)、加权召回率(*weighted-R*)和加权 F1 值(*weighted-F1*)作为模型预测评价指标。

$$weighted\text{-}P=\sum_{i=1}^{K}\left(P_i\times\frac{n_i}{K}\right) \tag{16}$$

$$weighted\text{-}R=\sum_{i=1}^{K}\left(R_i\times\frac{n_i}{K}\right) \tag{17}$$

$$weighted\text{-}F1=\sum_{i=1}^{K}\left(\frac{2\times P_i\times R_i}{P_i+R_i}\times\frac{n_i}{K}\right) \tag{18}$$

其中,K 为类别数,P_i为精确率,R_i为召回率,n_i为第 i 个类别的样本数。

6.2 数据与设计

为验证所提模型的有效性,本文用 4 种特征提取方法和 3 个分类器进行了多次实验。4 种特征提取方法包括:利用传统的 ERNIE 预训练语言模型提取实体知识特征;将传统的 TF-IDF 和 ERNIE 相组合,提取词的词频和实体知识特征;将 TFPOS-IDF 和 ERNIE 相组合,提取词的词性和实体知识特征;用 ERNIE 获取词向量,并采用 CSTFPOS-IDF 算法对种子库中关键词加权,提取关键动词、名词和实体知识特征。这些特征都被输入到 3 个分类器——KNN, LR, SVM 中。本实验数据集的训练集和测试集的划分比例为 7∶3,本文采用基于字符的方法对数据进行预处理。通过 2.3 小节的数据分析,实验中对试题文本进行短填长切,将每句话长度处理为 510。

6.3 实验结果分析

TFERNIE-LR 模型分类效果及其他几个特征提取组合得到的实验结果如表 6 所列。CS1 试题的认知过程维度自动分类中,CSTFPOS-IDF 赋予关键词权重因子 7、动词权重

因子 4、名词权重因子 3、其他词性权重因子 1。在使用 ERNIE 模型表示词向量时，利用 CSTFPOS-IDF 提取特征的效果优于 TFPOS-IDF 和 TF-IDF 提取特征的效果。使用 ERNIE，CSTFPOS-IDF，LR 的 TFERNIE-LR 模型在加权召回率、加权精确率和加权 F1 值上都优于其他特征提取及分类器组合的模型。基于 ERNIE 模型和 LR 分类器，CSTFPOS-IDF 加权算法比 TFPOS-IDF 加权算法的加权精确率高出 0.9%，比 TF-IDF 加权算法的加权精确率高出 13.7%。基于 ERNIE 和 CSTFPOS-IDF 特征组合方法，LR 分类器比 SVM 分类器的加权精确率高出 15.1%。

表 6 试题认知过程维度分类实验结果

Table 6 Experimental results of cognitive process dimension classification of test questions

分类器	ERNIE			ERNIE+TF-IDF			ERNIE+TFPOS-IDF			ERNIE+CSTFPOS-IDF		
	加权精确率	加权召回率	加权F1值	加权精确率	加权召回率	加权F1值	加权精确率	加权召回率	加权F1值	加权精确率	加权召回率	加权F1值
K-Nearest Neighbour	0.660	0.654	0.643	0.659	0.668	0.658	0.683	0.679	0.681	0.689	0.690	0.689
Support Vector Machine	0.619	0.587	0.587	0.653	0.646	0.643	0.679	0.619	0.623	0.682	0.684	0.683
Logistic Regression	0.722	0.713	0.715	0.696	0.689	0.692	0.824	0.822	0.821	**0.833**	**0.832**	**0.831**

CS1 试题的知识维度自动分类中，CSTFPOS-IDF 算法赋予关键词权重因子 7、名词权重因子 4、动词权重因子 3、其他词性权重因子 1，结果如表 7 所列。结果表明，基于相同的分类器和 ERNIE 预训练语言模型，利用 CSTFPOS-IDF 算法提取特征的效果优于 TFPOS-IDF 和 TF-IDF 算法提取特征的效果。TFERNIE-LR 模型优于其他特征提取及分类器组合模型方法。基于 ERNIE 模型和 LR 分类器，CSTFPOS-IDF 加权算法比 TFPOS-IDF 加权算法的加权精确率高出 0.6%，比 TF-IDF 加权算法的加权精确率高出 2.2%。基于 ERNIE 和 CSTFPOS-IDF 特征组合方法，LR 分类器比 SVM 分类器的加权精确率高出 2.6%。

Johnson 等[20]通过关于教学中 Bloom 分类法认知过程维度应用的课程教师访谈发现，教学评估大量集中在应用层次，并添加了一个“高层次应用”，但在访谈时，教师们对 Bloom 分类法的新划分存在异议。CS2013[21]将 Bloom 分类法认知过程维度的 6 个维度简化为 3 个，即熟悉、使用和评估，并用软件开发中迭代的概念，介绍了在不同层次的掌握程度。比如，在“熟悉”层次，掌握软件开发中迭代的定义，了解为什么使用这种方法；在“使用”层次，能够用迭代的方式编写程序；在“评估”层次，要了解该程序更多的迭代求解方法，并能从中选择一种合适的方法。这种方法虽简单，但存在缺陷，最高层次的创造被忽视了，理解和记忆混在一起，以致于 CC2020 不得不纠正这种简化，采用了 Bloom 分类法的 6 个认知过程维度。

表 7 试题知识维度分类实验结果

Table 7 Experimental results of knowledge dimension classification of test questions

分类器	ERNIE			ERNIE+TF-IDF			ERNIE+TFPOS-IDF			ERNIE+CSTFPOS-IDF		
	加权精确率	加权召回率	加权F1值	加权精确率	加权召回率	加权F1值	加权精确率	加权召回率	加权F1值	加权精确率	加权召回率	加权F1值
K-Nearest Neighbour	0.884	0.887	0.883	0.899	0.901	0.898	0.922	0.922	0.921	0.923	0.924	0.922
Support Vector Machine	0.881	0.873	0.876	0.918	0.912	0.914	0.929	0.929	0.929	0.935	0.935	0.935
Logistic Regression	0.925	0.924	0.924	0.939	0.939	0.939	0.955	0.955	0.955	**0.961**	**0.961**	**0.960**

TFERNIE-LR 模型在认知过程维度 6 个层次上试题自动分类的结果如表 8 所列。机器分类在一定程度上忽视了试题上下文语境，同时，课程专家们对试题进行标注时，也掺杂了一定的主观意识。

表 8 测试集认知过程维度各级别分类结果错误统计

Table 8 Error statistics of classification results at all levels of cognitive process dimension in test set

级别	测试数目	准确率/%	错误数	→记忆	→理解	→应用	→分析	→评估	→创造
记忆	49	93.87	3	—	1	0	2	0	0
理解	243	86.42	33	6	—	2	18	2	5
应用	143	86.01	20	0	4	—	7	1	8
分析	187	81.28	35	6	9	7	—	5	8
评估	36	55.56	16	1	4	3	7	—	1
创造	127	80.31	25	0	3	17	4	1	—

如试题“什么是关系？等价关系要满足哪些条件？”机器分类结果是分析，根据上下文语境，学生要回忆等价关系的定义，试题属于记忆层次。再如试题“证比求易算法（Verifying is easier than finding solutions）”这个案例，专家人工主观标注是元认知知识，机器分类是程序性知识。对于人主观确定的知识，机器是很难确定的。

结束语 本文基于 Bloom 分类法评估计算课程试题，分析 CS 课程试题所能达到的认知过程维度和知识维度，利用具有强大语义表征能力的预训练语言模型和具有强大特征提取能力的加权算法，提出了一种基于 Bloom 分类法的试题自动分类方法。具体地，给出针对 CS1 的认知过程维度和知识维度的相应动词和名词种子库，构建了高质量的 CS1 试题分类数据集。使用预训练语言模型 ERNIE 捕获试题文本中的语义特征获得词向量，使用 CSTFPOS-IDF 加权算法对

不同词性加权，组合后获得具有文本特征表示的文本向量，采用机器学习分类器进行试题分类，提出了 TFERNIE-LR 自动分类模型。在数据集上的实验结果证明了 TFERNIE-LR 模型的有效性。

参 考 文 献

[1] CLEAR A, PARRISH A, IMPAGLIAZZO J, et al. Computing Curricula 2020(CC2020): Paradigms for Future Computing Curricula[R]. New York: Technical Report, 2020.

[2] BLOOM B S. Taxonomy of educational objectives: the classification of educational goals: Handbook 1: Cognitive domain[M]. New York: David McKay Co. Inc, 1956: 1-9.

[3] ANDERSON L W, KRATHWOHL D R, AIRASIAN P W, et al. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives[M]. London: Longman Publishing Group, 2001: 25-80.

[4] ZHANG Z Y, HAN X, LIU Z Y, et al. ERNIE: enhanced language representation with informative entities[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019: 1441-1451.

[5] CHANG W C, CHUNG M S. Automatic applying Bloom's taxonomy to classify and analysis the cognition level of English question items[C]// 2009 Joint Conferences on Pervasive Computing (JCPC). IEEE, 2009: 727-734.

[6] OMAR N, HARIS S S, HASSAN R, et al. Automated Analysis of Exam Questions According to Bloom's Taxonomy[J]. Procedia-Social and Behavioral Sciences, 2012, 59: 297-303.

[7] HARIS S S, OMAR N. Bloom's taxonomy question categorization using rules and N-gram approach[J]. Journal of Theoretical & Applied Information Technology, 2015, 76(3): 401-407.

[8] JAYAKODI K, BANDARA M, MEEDENIYA D. An automatic classifier for exam questions with WordNet and Cosine similarity [C]// 2016 Moratuwa Engineering Research Conference(MERCon). IEEE, 2016: 12-17.

[9] FEI T, HENG W J, TOH K C, et al. Question classification for e-learning by artificial neural network[C]// Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint. IEEE, 2003: 1757-1761.

[10] YUSOF N, HUI C J. Determination of Bloom's cognitive level of question items using artificial neural network[C]// 2010 10th International Conference on Intelligent Systems Design and Applications. IEEE, 2010: 866-870.

[11] YAHYA A A, OSMAN A. Automatic classification of questions into Bloom's cognitive levels using support vector machines [C]// Proceedings of the International Arab Conference on Information Technology. Riyadh, Saudi Arabia. 2011: 335-342.

[12] ABDULJABBAR D A, OMAR N. Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination[J]. Journal of Theoretical & Applied Information Technology, 2015, 78(3): 447-455.

[13] MOHAMMED M, OMAR N. Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec[J]. PLoS ONE, 2020, 15(3): e0230442.

[14] DONG R S. Introduction to Computer Science: Thinking and Methods(Third Edition)[M]. Beijing: Higher Education Press, 2015: 1-335.

[15] SEDGEWICK R. Computer Science: An Interdisciplinary Approach[M]. GONG X L, et al, translate. Beijing: China Machine Press, 2020: 1-636.

[16] DEVLIN J, CHANG M, LEE K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. MN, USA: Association for Computational Linguistics, 2019: 4171-4186.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[18] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[19] COX D R. The regression analysis of binarysequences[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1958, 20(2): 215-232.

[20] JOHNSON C G, FULLER U. Is Bloom's taxonomy appropriate for computer science? [C]// Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006. 2006: 120-123.

[21] SAHAMI M, ROACH S M. Computer Science curricula 2013 [J]. ACM SIGCSE Bulletin, 2013: 29-219.

DONG Rongsheng, born in 1965, professor, is a senior member of China Computer Federation. His main research interests include knowledge graph and machine learning.

LI Fengying, born in 1974, Ph.D, professor, is a member of China Computer Federation. Her main research interests include knowledge graph, machine learning and symbolic computing.

（责任编辑：何杨）