

# 教师可干预的在线讨论质量分析工具的设计与实现

吴昊

(内蒙古师范大学 呼和浩特市 011599)

**摘要:** 在线学习以是当下必不可少的学习模式, 其在线讨论产生的海量数据蕴含着巨大的教育价值。然而, 当下的对于在线讨论的分析多数是依靠人力或是依靠简单的特征提取部分信息, 分析不够科学和严谨, 从而大大降低数据的价值。基于此, 本研究综合现在何将人工智能领域早已成熟的自然语言处理技术、数据挖掘技术应用到教育领域中, 使数据呈现更多价值, 从而为教育工作者提供一种新的观察数据的工具, 同时也使得在线讨论质量分析的相关研究有新的应用与发展, 具有一定的研究意义。本文通过文献分析, 整合现有对于在线讨论的研究以及分析框架, 使用 Python 语言设计开发供教师干预的半自动化的在线讨论质量分析工具。

**关键词:** 在线学习、在线讨论、自然语言处理

Design and implementation of quality analysis tools for teachers' intervention in online discussion  
Wu Hao

(Inner Mongolia Normal University, Hohhot 011599)

**Abstract:** Online learning is an indispensable learning mode at present. The massive data generated by online discussion contains great educational value. However, the current analysis of online discussion mostly relies on human resources or simple feature extraction of partial information, which is not scientific and rigorous enough, thus greatly reducing the value of data. Based on this, this study integrated the application of natural language processing technology and data mining technology already mature in the field of artificial intelligence to the field of education, so that the data presented more value, so as to provide educators with a new tool to observe the data, but also make the online discussion quality analysis related research has a new application and development, has a certain research significance. Through literature analysis and integration of existing research and analysis framework on online discussion, this paper uses Python language to design and develop a semi-automatic online discussion quality analysis tool for teachers' intervention.

**Key words:** online learning, online discussion, natural language processing

## 引言:

21 世纪以来, 信息化进程在不断推进, 在教育领域中信息化同样取得了不错的成功, 不少在线学习平台已有完备的功能和技术供学习者与教育工作者学习和使用, 同时软件系统和平台中存储着海量的学习者信息及学习过程数据。而如何让数据成为精准教学的重要依据, 是当前教育研究者亟待解决的问题。

随着教学形式的多样化, 包括在线讨论、学习计划、教学反思等在内的以文本为主要形式的非结构化数据日渐积累。非结构化数据往往包含完整而连续的信息, 其价值得到越来越广泛的认可。由于语言表达具有灵活性和多样性等特点, 因此文本类非结构化数据在分析上会受限于文本处理的难度。现有以教学文本为对象开展的研究大都是通过人工编码分析的方式, 即由若干名经过培训的编码员对研究数据进行独立编码, 随后通过讨论确认不一致或一致性分析等方式统一数据, 并在此基础上, 根据研究目的进行合理的分析推断。由于非结构化数据处理及分析的难度较高, 耗费的人力和时间成本无疑是比较大的, 同时也很难拓展到更大量级的数据研究层面。

## 问题的提出:

当下在大规模在线学习普及的同时, 也促进了信息的生产 and 创造。多数学生产生的过程性数据大多以文本的形式留存。教师使用人力进行文本质量的标注或者是仅仅依靠简单的特征对文本的质量进行分析评价是当下非常常见的方法。然而, 在大数据时代下, 这些海量的数据是很难做到使用人力去进行筛选分类的。同时, 简单的指标也无法有效的区分出文本的质量与学生个体间的差异。所以通过技术赋能的方式, 提供教师一种新的观察数据的手段, 从而实现人力不可为的分析, 如何这样一款工具是当下值的研究的问题。

## 本文关注的问题是:

1. 设计什么样的在线讨论分析模型, 使不同维度的在线讨论分析方法与自然语言处理技术相结合, 从而提供教师新的观测、评价在线讨论的手段

2. 如何设计开发一款软件, 具有通用性、可视化、易操作等特点, 用以在线讨论质量分析, 从而提升教师工作效率。

在线讨论分析工具设计:

在线讨论质量分析的第一步就是如何实现科学准确的对文本进行分类编码, 这个过程首先需要制定分类指标, 依据指标, 将讨论数据归类到特定的类别之下。本论文以在线学习当中产生的过程性讨论数据作为分析对象, 通过文献分析总结已有的在线讨论分析的分析指标, 使之结合当下的文本分类技术, 同时考虑到这些指标与技术落实到在线学习领域的特殊性与适用性, 从而设计在线讨论质量分析的一般模型。

文本分析模型的训练是在线讨论质量工具的软件总体框架, 其实现依赖于文本分类技术。文本分类技术主要包括数据预处理、特征选择、词性标注等步骤, 其中数据预处理又包括分词和去停用词。在自动编码模型的训练过程中, 利用少量的已标注在线讨论数据, 即已完成归类的数据集, 采取一定的技术处理手段, 达成构造自动分类模型的目的。首先, 通过预处理, 将文本数据划分为由多个字段组合的有序序列, 并过滤无关信息; 然后, 通过各字段的出现频次和字词间的词性结果的等信息, 基于不同维度的质量算法, 选择并放大有效文本中的特征信息; 最后, 利用特征信息和一定的分类算法, 深挖文本内容与编码之间的潜藏映射关系, 构造数据对应的分类模型。自动编码模型, 依托计算机的强大运算能力, 能够有效解决人工编码中人力和时间成本耗费大、数据量大分析难的问题。

一般数据分析工作的输入为基础语料集合, 在本论文研究中, 除未标注过的“生语料”集合外, 还需输入研究选定的编码量表。编码量表是内容分析过程中编码实施的基础, 在短文本分类中表现为问题核心的分类体系。而质量分析的实质就是完成原始输入的两类数据间的映射, 基于编码实现对讨论数据的分析理解。利用训练好的质量分析模型, 使计算机替代人力快速完成“生语料”的编码, 进行输出。然而, 输出的大量复杂数据被紧紧地组织到一处, 尽管可以通过列表的形式呈现, 但不同标准的数据任意排序, 不同编码的数据数量分布以及特征表现并不明显, 这对理解与分析是十分不利的。因此, 在线讨论质量分析一般模型中加入了统计可视化的步骤, 利用计算机的高速运算对编码数据进行梳理、汇总, 并依托图表强大的视觉传达能力对统计后的结果进行精加工, 使编码结果易于理解, 从而辅助教师进一步推断总结。

### 功能模块设计:

本研究对学生在线讨论的评价工具设计采用行为投入、认知投入以及情感投入三个维度,这三个维度的评价结果也能够判断学生是否达到了深度学习。其中,学生在线讨论的行为投入评价使用在线讨论数量评价表;学生认知投入评价使用在线讨论质量评价表;学生讨论情感投入评价使用在线讨论交互性质评价表。

教师投入影响学生在线讨论的效果,全身心地投入到学生的小组讨论会促进学生的深度讨论,但是大量的教师投入对教师本人来说是一个巨大的压力,因此也需要对教师投入进行评价,目前较为公认的教师投入包括情感投入、时间投入和精力投入。在本研究中,结合实际讨论情况,将教师投入评价分为三个维度,包括三个:时间投入、精力投入和情感投入。

### 在线讨论质量分析工具的实现

在线讨论质量分析工具的前端是依靠 PyQt5 完成基本界面结构的搭建,通过 pandas、xlrd、xlwt 完成文本数据的导入。使用 Python 语言中 jieba、jiagu 实现文本分词,以及前期的数据预处理工作,数据库有 MySQL 提供支持。

### 数据预处理模块的实现:

在线讨论质量分析工具中,数据预处理模块主要分为两个部分,分别是文本分词和去停用词。

文本分词采用 Python 中的 jieba 库实现,jieba 分词库具有全模式、搜索引擎模式、精确模式三种分词模式,在本研究中为保证分词的准确以及减少冗余特征和无关特征的产生,因此本研究采取精确模式进行分词。同时,由于在线讨论数据的特殊性,研究也会采用 jiagu 分词库进行分词,jiagu 分词库可以根据用户自定义语料库进行分词处理,更精准的实现对讨论语句的切分。

文本预处理的第二部分工作就是去停用词处理,在学生讨论的过程中,往往会出现如表情、错别字等,去停用处理的主要就是清理数据中的无法识别数据、无关数据,达到更精确、更高效的处理文本。

### 数据分析模块的实现:

不管是在线讨论还是面对面交流,都是一个学习的过程,过程中学习者的认知、情感变化和行为表现都是以文本的形式来呈现。在线学习过程中,学生参与讨论的积极性与参与性很大程度可以通过学生的外显行为进行表现,即学生发言频次,发言字数特征。因此行为投入主要关注学生在小组讨论中的主动性、参与程度以及发言状况,转化为量化评价,主要关注的是学生发言频次和发言字数。在设计过程中,随机选取了部分小组的讨论,对学生发言次数和发言字数进行统计,学生发言次数最低为 1 次,最高为 22 次,发言字数最低 56 个字,最高 874 个字。基于该统计结果,将该范围平均分成 5 个等级并进行赋值。

根据学生行为投入量表设计学生行为投入功能,将学生讨论数据以 Python 编程语言中字典类型进行存贮,学生姓名与其发言形成键值对,信息输入程序之后根据键值对识别学生姓名与讨论数据,从而累计学生发言频率与发言字数。

认知投入主要关注在线讨论时学生发言达到的认知维度和目标维度。本研究采用的学生认知投入评价工具是杨虹丽(2018)在“翻转课堂促进大学生深度学习的实践研究”中所建立的评价工具。该评价工具不仅结合了深度学习的定义,还融合了该领域常用的评价工具,主要从认知和目标两个维度设计了深度学习评价,认知维度参考了 SOLO 分类法,目标维度参考了我国对认知领域教学目标的分类型体系。具体评价指标如表所示。

根据学生认知评价分析表设计学生软件认知评价功能,在该模块将使用基于词频(TF)的以及基于词频—逆向文档频率(TF-IDF)做特征选择,做特征选择之后学生的讨论数据将形成学生讨论特征词,学生讨论特征词与同义词库对比,形成学生讨论特征范围,教师围绕讨论主题预设主题词,主题词同样对比同义词库,形成主题词范围,接着做基于互信息的关联性根据分析结果以及特征词的对量进行评价

社会交互投入主要关注学生在线讨论时发言的交互性质,分析的是学生在线讨论的过程中的行为数据和行为关系数据。小组讨论

是协作学习的一种形式,在小组讨论中,最直接最有效的方式就是同伴之间的交互。同伴交互的质量对小组讨论的质量和个人的思维发展以及学习深度有重要影响。基于此,该部分的研究工具借鉴了黄荣怀等构建的协作学习过程中实时讨论的评价因素及对应的交互言行体系表,在该表原有的“建构”、“互动”和“争辩”3种水平基础上,根据在线讨论的实际情况增添了“无效”这一最低级的交互水平。在使用该评价工具时,研究者依据首句词和相应解释对讨论文本进行相应编码。

根据自然语言处理技术做具有的词性标注功能,将句子拆分成不同字词,判断其中语气词转折词,识别是否与其他同学发生讨论,设计独立预料库,做建构、互动、争辩不同层级的预料识别功能,提取学生讨论中涉及到的不同维度是句首词,完成社会交互投入功能开发。

### 统计可视化模块的实现:

通过文献阅读以及相关材料的收集,在当前提供质性分析的工具当中可支持用户去选择多种可视化图表呈现统计分析数据,如词云图、饼状图、直方图、列表等。其中列表可以呈现对数据编码之后每一条数据的明细,饼状图以呈现不同类别的数据在数据整体中的分布,直方图可以更好展现数据的统计结果,以及不同类别数据之间的数值差距。以此,本研究也会根据以上图表的特点以及功能,为使用本软件的用户提供相应的图表来呈现数据,词云图将采用 jiagu 库中的函数实现词云图的可视化,使用 Python 中的 matplotlib 库提供直方图、饼状图的可视化实现,matplotlib 是建立在 Numpy 之上的一个 Python 图库,它提供了一个面向对象的 API 和一个过程式的 MATLAB API。

### 总结:

教育信息化时代下,技术赋能教育已是当下的必由之路,但我们始终在路上,需要我们教育工作者不断努力与探索使之成为可能。在线学习平台是教育工作者以及研究者获取教育资源以及学生反馈的优质渠道,尤其在当前疫情之下,在线学习平台的使用更为普及,数据的产生更广泛、更密集,而其中的数据以文本作为主要形式。因此,如何将人工智能领域早已成熟的自然语言处理技术、数据挖掘技术应用到教育领域中,使数据呈现更多价值,从而为教育工作者提供一种新的观察数据的工具,同时也使得在线讨论质量分析的相关研究有新的应用与发展,具有一定的研究意义。

在线讨论通常是学习者以及教师围绕某个话题进行意见分享、问题解答、学习交流的活动,这种方式对于学习者思考理解问题有很大的促进作用,同时也大大的降低学习者在在线学习过程中的疏离感,因此这种方式的到广泛教育工作者的认可。在线讨论主要以文本数据为主,本研究以在线讨论中的文本数据为分析切入点,结合自然语言处理技术,对在线讨论质量分析方法进行探索与实践,有针对性的设计开发一款在线讨论质量分析工具,实现从不同维度对在线讨论数据进行编码、分析以及可视化呈现等功能,从而促进在线讨论质量分析研究的的深化发展,期望能够为技术赋能教育提供新思路。

### 参考文献:

- [1]张娜.国内外学习投入及其学校影响因素研究综述[J].心理研究,2012,5(02):83-92.
  - [2]徐丽华.国外形成性教学反馈的研究及启示[J].当代教育科学,2010(07):27-31.
  - [3]李银玲.远程学习中在线讨论的设计与主持[J].中国电化教育,2010(01):59-64.
  - [4]余明媚,李文光,王新辉.学生投入在线讨论的影响因素研究[J].现代教育技术,2010,20(04):85-90.
  - [5]何玲,黎加厚.促进学生深度学习[J].现代教学,2005,(05):29-30.
  - [6]冷静,刘黄玲子,黄荣怀,李乾.在线协作讨论中成员表现的综合评价研究[J].现代教育技术,2007(03):51-56.
- 作者简介:吴昊(1997.10—),男,汉族,内蒙古鄂尔多斯人,内蒙古师范大学教育学院,三年级在读研究生,硕士学位,专业:教育技术学,研究方向:教学设计与绩效技术。