

基于 Python 的高校在线课程互动数据采集与多维度分析

成纳哲

陕西理工大学 山东淄博 255000

摘要: 随着高校在线课程的普及,课堂互动数据采集与分析对于教学质量监测与优化具有重要意义。本文基于 Python 技术,设计并实施了高校在线课程互动数据的采集流程及多维度分析方法。通过实践案例,详细展示了数据清洗、时间序列分析和文本内容挖掘等操作,揭示了学生互动活跃度的时段性分布及教师授课方式对互动效果的影响。结果为教学策略优化提供了依据,也验证了 Python 在教育数据分析中的可行性与高效性。研究成果可为高校在线教学改进提供有益参考。

关键词: 高校在线课程; Python; 互动数据采集; 多维度分析; 教学优化

引言

随着信息化教学的持续发展,高校在线课程成为培养创新型人才的重要方式之一。在线课程的互动性,直接关系到学生的学习参与度和教师的教学质量。大量研究表明,课堂互动数据的采集与分析对于优化教学过程、提高教学效果有着不可替代的作用。Python 作为一种功能强大的编程语言,具备灵活的数据采集和多维度分析能力,适合应用于高校在线课程互动数据的深入研究。本文将聚焦 Python 在高校在线课程互动数据采集及多维度分析中的应用实践,旨在通过实证案例展示数据采集流程、数据处理方法以及多维度分析结果。论文不仅关注技术层面的具体操作,同时探讨其在教学评价和课程优化中的实际意义。通过构建相应的分析框架并验证其实效,期望为高校在线课程建设和数字化教学改革提供可借鉴的研究思路和技术支持。

1 高校在线课程互动数据采集需求与方法

1.1 在线课程互动数据的内涵与采集需求

随着高校在线课程的普及,教师与学生之间的互动方式不断丰富,互动数据也呈现出多样化的特征。互动数据是指在在线课程教学中,教师和学生通过平台进行各种交流、反馈、讨论、提问、答疑等形成的痕迹化信息,涵盖浏览量、评论、点赞、作业提交、答疑讨论等多维度内容。这些数据不仅直观地反映出学生的学习热情与参与深度,也能够辅助教师评估教学效果和及时调整教学策略。在当前教育数字化背景下,充分采集并分析这些互动数据已成为高校教学改革的关键。基于此,互动数据采集的需求日益突出,包括但不限于以下几方面:一是捕捉教师与学生互动过程的真实

记录;二是多维度呈现学生群体在不同时间段和不同知识点的活跃状态;三是通过数据反映教学环节中潜在的问题和改进之处;四是为学校管理部门的教学质量监测提供客观依据。因此,设计一个高效、可扩展的数据采集方案成为高校数字化教学发展的迫切任务。

1.2 Python 在数据采集中的应用及其优势

在多种编程工具和数据分析方法中,Python 以其简洁易学、库丰富、应用灵活等特点脱颖而出,成为高校数据采集和分析的首选工具。首先,Python 拥有大量成熟的网络爬虫库,如 requests、BeautifulSoup、Selenium 等,可以快速搭建网页数据采集脚本,适应各类教学平台的多样化页面结构。其次,Python 支持丰富的数据处理和存储库,如 pandas、sqlite3、csv 等,可对采集到的原始数据进行灵活的格式转换、缺失值处理、去重和清洗等操作,为后续分析提供可靠的数据基础。此外,Python 的可视化能力强大,matplotlib、plotly 等库能够帮助研究者直观呈现数据特征、趋势和分布情况。相较于传统的手工采集或单一数据库查询方式,Python 脚本化操作可以批量化、自动化地完成数据采集与初步处理,显著提高工作效率和数据处理质量。更重要的是,Python 社区活跃,相关教程和案例丰富,易于高校师生借鉴和二次开发,极大地拓展了其实用性。

1.3 实践案例:高校课程平台的数据采集流程

为了更清晰地说明基于 Python 的数据采集方法,本文以“高校 MOOC 平台”的数据采集为例,描述实践操作流程。首先,确定数据采集的目标模块,例如《Python 编程入门》课程的评论区、作业提交记录和问答讨论版块。接着,

使用 requests 库模拟 HTTP 请求获取网页源代码,如果遇到 JavaScript 动态加载页面的情况,则通过 Selenium 库控制浏览器实现动态加载后的数据抓取。具体而言,编写 Python 脚本,发送请求到评论区 URL,通过 BeautifulSoup 库解析 HTML 文档,提取出评论文本、评论时间、评论者昵称等字段。若需更复杂的数据(如视频播放量、作业提交时间等),可以结合正则表达式灵活提取所需信息。在数据存储方面,考虑到后期分析的灵活性,选择将数据保存为 CSV 文件,或者直接写入 sqlite 数据库,便于在 Python 后期分析中进行快速读写。为了方便教学应用,也可借助 tkinter 库设计简单的图形化界面,供非编程背景的教师进行操作,提升数据采集工具的可用性和友好度。

2 互动数据的预处理与多维度分析框架

2.1 数据预处理技术

在进行高校在线课程互动数据分析之前,首要步骤便是对原始数据全面的预处理。由于数据通常来源于多样化的网页或平台接口,格式各异且往往伴随大量噪声信息,因此预处理对于数据的可用性具有决定性意义。数据清洗的主要目标是确保数据的完整性、准确性和一致性。具体而言,首先需要对采集到的原始数据进行去重处理,删除重复的评论或互动记录,保证每条数据的唯一性。其次,针对部分字段存在缺失或异常值的情况,需进行填充或删除。例如,若评论时间缺失,可考虑用该用户的平均互动时间进行填充;若评论内容为空白,则可将其直接删除以免影响文本分析的有效性。

此外,数据中还可能包含多余的 HTML 标签、表情符号或特殊字符,这些无效内容需要通过正则表达式或字符串替换方法进行清除。数据格式化也是预处理的重要环节,通常将不同来源的 JSON 数据或 HTML 表单结构化为统一的 CSV 或 DataFrame 格式,便于后续的多维度分析操作。Python 中的 pandas 库在此过程中扮演了重要角色,能够快速读取、合并、转换数据结构,支持大规模数据的高效处理和初步展示。例如,使用 pandas 的 dropna() 函数可以直接剔除缺失值较多的行,而 fillna() 函数则可灵活填充空缺。通过这些预处理操作,原始数据才能真正转化为可分析的高质量数据集。

2.2 多维度分析的指标体系构建

高校在线课程的互动数据,涉及到师生、课程、时间

等多个维度,因此在正式分析前,有必要构建科学、合理的多维度指标体系。首先,交互活跃度是最基本的分析维度,常用指标包括活跃人数、评论总数、平均评论长度、回复频率等,能够直接反映课程氛围的热度。其次,教学参与度则着重刻画学生对教学内容的投入程度,典型的衡量指标有观看视频的时长、作业提交率、测验通过率等。

在学生个体维度上,可以通过统计其日常互动频率、评论关键词分布等,了解其兴趣偏好和学习习惯。而教师授课维度的分析重点在于教师的授课风格、与学生的互动强度及其在不同阶段的引导效果。通过这些多维度的指标设计,可以从全局上把握课堂互动的状态,也能深入到具体的子群体或行为模式中,实现更有针对性的教学改进建议。这种分析思路不仅有助于提升课程内容的针对性,还为个性化教学路径的优化提供了可靠的基础。

2.3 Python 实现的多维度分析方法

为了将上述指标体系转化为可操作的分析过程,Python 的多维度数据分析能力得到了充分的发挥。首先,pandas 库的 groupby 和 pivot_table 函数可用于实现按教师、学生或时间等维度的分组统计,计算如日均评论数、周观看量等关键指标。这类操作能够帮助教师把握各个时间段的活跃情况,评估课程设置与互动效果之间的关系。其次,时间序列分析也是多维度分析的重要方面。通过 matplotlib 库的折线图、条形图等基本绘图功能,可以直观展示互动数据在各时间段的变化趋势;而 plotly 库更是提供了交互式的可视化功能,让用户能够动态查看不同时间段的数据细节。

在文本数据处理方面,Python 也有着丰富的工具箱。比如,使用 jieba 库对中文评论进行分词处理,可提取出最常出现的高频词汇;再借助 WordCloud 库,将这些词汇以词云的形式展示,便于快速洞察学生的兴趣点和困惑点。此外,pandas 的 merge 和 concat 操作支持多张表的灵活合并,可将评论数据与学生成绩或提交记录等多表数据进行关联,实现更复杂的维度交叉分析。

在可视化层面,除基本统计图外,还可生成热力图展示活跃度分布,如不同课程模块的评论数或视频观看次数热力图,这对于发现课程薄弱环节具有显著意义。通过这些 Python 多维度分析工具的结合应用,不仅让数据分析结果更具说服力,也让教师和管理者能以直观方式把握教学全貌,及时调整教学策略,优化学生体验。

3 案例应用：某高校在线课程的互动数据多维分析

3.1 案例选取与数据采集实践

为了验证数据采集与分析方法的可行性，本文选取了某高校 MOOC 平台上的《Python 编程入门》课程作为案例。该课程面向计算机及数据科学相关专业学生，活跃度高且互动数据丰富。采集目标涵盖学生评论、弹幕留言、视频播放次数以及作业提交信息。数据采集采用基于 Python 的 requests 库获取页面源代码，并使用 BeautifulSoup 进行 HTML 解析；动态加载部分则借助 Selenium 库模拟浏览器环境。采集的字段包括评论 ID、用户 ID、评论时间、内容、点赞数和所属章节等，便于后期多维度交叉分析。

数据存储方面，为了方便灵活调用，数据既保存为 CSV 文件，也导入 sqlite 数据库。采集中也遇到一些挑战，例如该平台存在反爬策略，需要在请求中添加合适的头部信息和 cookie 等身份验证，部分评论还包含表情和特殊字符，需做编码转换和清洗。最终成功采集到 2000 余条学生评论和 500 余条弹幕信息，以及完整的作业提交数据，为多维度分析奠定了基础。

3.2 数据预处理与多维度分析实施

正式分析前，先对原始数据进行预处理。使用 pandas 对数据进行去重，剔除重复提交的记录。针对部分评论中存在的特殊符号和 HTML 标签，使用正则表达式进行清理，保证文本的可读性。时间戳字段统一转换为标准日期格式，方便后续的时间序列分析。

多维度分析中，首先对评论活跃度进行统计，使用 pandas 的 groupby 方法，按天和章节分组计算评论数量，绘制评论量的时间序列折线图。结果表明，课程更新后 1-2 天内评论数量迅速增加，随后趋于平稳。使用 pivot_table 生成不同班级的评论数据透视表，展现了各群体的活跃度差异。

文本分析方面，借助 jieba 分词库进行分词，并生成词云图展示高频词汇。常见高频词包括“作业难度”“学习进度”“视频问题”等，反映了学生的学习体验和反馈。为更直观展示，利用 matplotlib 绘制热力图，展现不同章节、时间段的互动数据分布。结果显示，部分章节互动更为集中，教师可以关注这些高活跃章节的教学方法。

3.3 结果解读与改进建议

通过上述分析发现，该课程的学生互动活跃度具有明显的时段性，活跃高峰主要集中在视频更新后的 48 小时内。教师可以在这一阶段加强答疑、发起话题，激发学生讨论热情。高频词中“作业难度”问题提示教师要在作业布置中平衡难度，避免学生畏难情绪。此外，弹幕互动量相对较低，说明部分学生更倾向于非实时讨论，教师可在视频中嵌入互动小问答或引导弹幕使用，提升课堂参与度。

同时，数据也显示教师授课节奏对互动效果有影响，语速适中、细致讲解的视频更受学生欢迎。教师可据此优化授课方式，增强课堂吸引力。尽管本次案例验证了分析方法的可行性，但也存在数据样本单一、数据时长有限等不足。后续可在更多课程、多平台采集数据，结合自然语言处理等技术，进一步完善分析框架。整体来看，Python 在数据采集与多维度分析中展现出高效、灵活的优势，为高校教学评价和课程改进提供了有力支撑。

4 结论

本文通过基于 Python 的爬虫和数据处理技术，系统介绍了高校在线课程互动数据的采集流程和多维度分析框架。以《Python 编程入门》课程为例，实践验证了 Python 在数据去重、文本分析、时间序列统计及可视化方面的优势。分析结果揭示了学生互动的时段性和教师授课方式对互动的影 响，提供了有针对性的课程优化思路。研究虽存在样本单一、分析维度尚可扩展等局限，但证明了基于 Python 的互动数据分析在高校教学质量监测和课程迭代改进中的应用潜力，值得在后续研究中进一步深化和推广。

参考文献：

- [1] 曾香金. 基于 Micro Python 与人工智能的互动式教学数据采集方法[J]. 贵阳学院学报(自然科学版), 2024, 19(01): 74-78.
- [2] 王海涛, 齐达. 基于 Python 下的数据采集和分析在融媒体中的作用[J]. 电声技术, 2018, 42(06): 63-64+67.
- [3] 邵杨芳. 健康教育类在线课程的用户需求及评价挖掘分析[J]. 中国大学教学, 2023, (Z1): 100-113.
- [4] 刘云鹤, 李志敏, 阮心明. 基于画像技术的在线课程质量评价方法研究[J]. 中国高等医学教育, 2023, (02): 13-14.