网络爬虫课程内容探索

李洋[1] 孙轲[1] 李伟[1] 张杰[1] 舒静[2]

（1.电子科技大学成都学院；2.成都郫都区博瑞实验学校）

摘要：针对工科类型学科与课程思政融合存在困难的现状，以网络爬虫技术课程教学内容作为研究对象，为后续课程设计红色知识图谱做“知识”储备。对红色数据元素进行挖掘，围绕爬虫技术、课程内容、教学设计提出方法，以期为工科类型学科科目与课程思政融合提供思路与参考。

关键词：课程思政，网络爬虫技术课程，工科类型学科。

Exploration of Web Crawler Course Content in Curriculum Civics Perspective

Li Yang, Sun Ke, Zhang Jie

（Chengdu College of University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731）

Abstract: In view of the current situation of difficult integration of engineering disciplines and courses ideological and political, taking the teaching content of the web crawler technology course as the research object, a red knowledge map is designed for the follow-up courses as a ″knowledge″ reserve. Mining red data elements, and putting forward methods around crawler technology, course content, and teaching design, in order to provide ideas and references for the integration of engineering disciplines and courses ideological and political.

Key words: Course Ideological and Political, Web Crawler Technology Course, Engineering Type Discipline.

1 引言

2016 年 12 月，习近平总书记在全国高校思想政治工作会议上强调，高校要坚持把立德树人作为中心环节，把思想政治工作贯穿教育教学全过程，实现全程育人、全方位育人、努力开创我国高等教育事业发展新局面[1]。“大学之道，在明明德”，高校作为育人之地更要育德，更要全方位育人。而工科类型的学科往往具有较强的理论要求、逻辑要求，知识方面相对枯燥，将思政教学与学科进行融合存在多方面困难。在教学设计上，需要将思政教学内容与课堂知识进行紧密衔接，一味生搬硬套不仅会影响课堂教学设计，还会影响学生的学习兴趣，甚至弱化学生思想觉悟水平。

针对工科类型学科与课程思政融合存在困难的现状，结合网络爬虫技术课程，筛选知识挖掘目标，引入红色知识作为挖掘目标，同时为后续课程设计红色知识图谱作“知识”储备。

2 网络爬虫技术与知识挖掘

近年来，“大数据”一词不断出现在生活中，与此同时，在生活中、学习中、工作中也往往需要“大量数据”的支持，而这些数据往往来自于互联网。互联网中的数据呈现出数据量大、参杂广告、数据正确性未知、数据冗余等问题。能够自动完成甄别、筛选重复的页面或者找到高质量的网站并将需要的数据进行捕获、清洗、本地化保存就显得十分有意义，而爬虫就是实现相关功能的重要手段。网络爬虫技术应用范围广泛，尤其是百度一类的检索网站。在搜索引擎中，爬虫可以在互联网中对网站进行检索筛选，将符合检索的目标进行排列展示，以便用户后续处理。另一方面，对目标服务器需要获取部分关键数据并希望本地化持久存储时，网络爬虫技术可但当重任。现实情况是并非所有的网站均开放可爬，一般可以通过检查 robots 协议以及法律声明获取目标网站的隐私信息。部分网站服务器会进行反爬虫处理，一般会检查发起者的身份信息，对非浏览器的发起者进行屏蔽。与之对应的也有反反爬虫处理，如将在发起请求的数据包中修改 UserAgent 信息，将其伪装为浏览器形式等。一般而言，不能对有明确声明禁止爬取的网站进行数据爬取，应当遵循相应的协议。

3 课程内容与数据爬取

课程内容选取红色知识，在视频、文本、图片等多方面的素材中进行数据提取。在获取数据、分析 web 结构的同时，对数据也进行归类本地化持久处理，在潜移默化中树立同学们正确的人生观、价值观、弘扬爱国主义精神，培养积极向上、富含正能量、富有爱国情怀的新时代年轻学者。

3.1 Robots 协议

使用爬虫程序对互联网中的数据进行爬取需要格外谨慎，网页上往往会对该网页的爬取权限进行说明、限制，对此爬虫设计者们就应该了解 robots 协议。

多数网站都进行了 robots 协议声明，该协议对当前网站的数据权限进行了说明，对于数据权限应当遵循。查看 robots 协议一般在官网地址后面增加“/robots.txt”即可。以 bilibili 网站为例，网址 url=https：//www.bilibili.com/，爬虫设计者们只需要在网站的最后面加上 robots.txt 就可以查看网站的 robots 协议。访问 https：//www.bilibili.com/robots.txt，之后网站就会返回这样的界面。

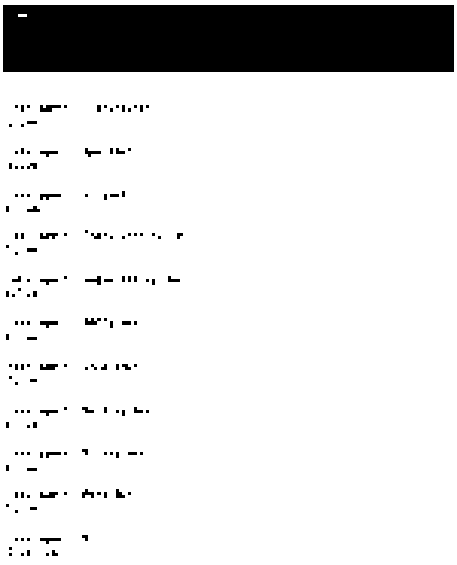

图 1 B 站的 robots 协议

图 1 中展示了 B 站的 robots 协议内容，UserAgent 代表爬虫的所

有者，对相应的爬虫设置了允许权限与禁止权限。一般来说，爬虫设计者需要遵循 robots 协议中的权限说明，在允许访问的数据范围内保证不影响对象服务器正常运行的情况下合理获取数据是较为友好的行为。

对于允许访问的数据也不能够毫无限制去访问爬取数据，过多的访问量可能会占用服务器的资源并造成服务器请求拥挤，甚至导致服务器死机，给目标网站和其他用户带来麻烦。课程设计中要求同学们对目标服务器的访问应略快于正常用户在浏览器中的访问速度，通过延时或睡眠形式降低程序对目标服务器的访问频率，确保服务器不受影响。

3.2 数据网站选取

结合课程定义，数据选取以红色知识为目标。课堂中筛选开放的网站获取红色知识，并完成本地化持久存储。以党史学习教育网为例，查询 robots 协议后结果为 404，表现为没有 bobots 协议。一般来说，在不影响对方服务器正常运行，获取目标知识为学习的前提，且不存在 robots 协议的情况下，可以考虑以温和的形式完成对数据爬取。

3.3 UA 伪装

在课程设计中，引入相应的基础反爬虫措施，即伪装 User-Agent，本文简称 UA 伪装。设计 UA 伪装后，能够对部分服务器筛查请求头时不被发现为程序而不是浏览器。

使用浏览器附带的抓包工具，点击网络之后再刷新目标网站界面。在数据中可随意任意点击选择，在右方数据框中有三个选项，分别是常规，响应头，请求标头。在常规里得到浏览器对网站的请求方式等信息，在响应头中可以查看响应标头的部分信息，在请求标头中可以看到 User-Agent 等信息。

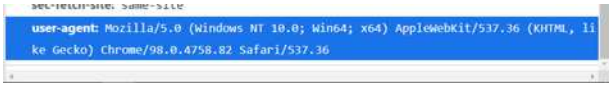

图 2.USER-AGENT 信息

3.4 数据解析

以获取邓小平故居陈列馆数据为例，利用抓包工具查看数据信息，找到对应的 url，拼接网址即可对目标爬起请求。网页源代码如下图所示：

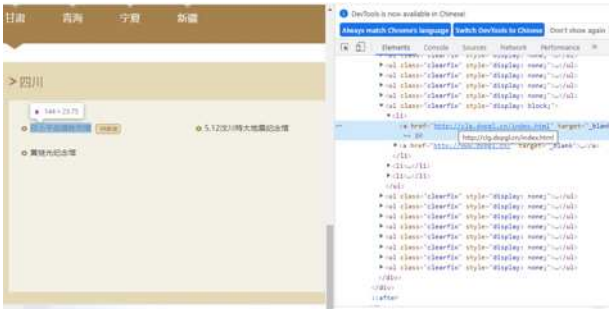

图 3.邓小平故居陈列馆 web 结构图

利用 a 标签的 href 属性值作为新的 url，对目标发起请求可以获得对应服务器响应数据。在新目标中，以生平简介为例，在标签 <spanstyle="font-size：16px；line-height：1.5；">中可以获取到目标对象。内容如下图所示：

图 4.个人信息 web 结构图

对于一次爬虫行为存入列表的数据，为了数据可视化爬虫设计者们需要将数据进行可视化处理，这时候就需要将爬虫设计者们通过爬虫所爬取到的文件存放到本地的文件中。爬虫设计者们可以使用 python 自带的 csv 的库来再本地创建一个 csv 文件，并将爬虫爬取到的数据存放到这个 csv 文件当中。

4 结束语

本课程以开放的红色知识获取为目标，课堂强调以温柔、合理、善意的方式对目标服务器进行访问，反复强调技术、工具的正确使用，树立正确的价值观。课堂内容上案例丰富，对 python 的爬虫支持库、框架、解析库、正则表达式进行了充分的演练，对网站的反爬虫和 ip 封禁机制有了更加明显的认识，从专业技术上对学生有较为全面的培养。

课堂以网页分析、知识讲解、案例测试、项目实战四个环节完成，考核采用课上课后练习结合期末大任务完成，采用答辩方式让学生在知识输出、讲演等多方面进行锻炼培养，有效提升学生工程能力。以 2018 级同学为例，对网络爬虫课程学习完成后在课后反馈中获得了学生的一致好评，在后续学习中也反馈多次运用到爬虫知识，超过百分之 50 的同学在毕业设计中运用了爬虫技术完成设计前期数据积累。后续课程中，采用 neo4j 图数据库为基础工具，结合爬虫知识、深度学习、微信小程序设计等知识，以设计“红色知识问答系统”为目标开展实训课程，以一个完整项目展开，以三人小组构建团队，并在期末进行项目答辩验收。对前后课程、知识进行衔接，以实际案例结合定期进度汇报、难点阐述、自由讨论完成对同学的工程能力培养。

爬虫终究只是人们手中的一个工具，课堂中除开思政内容引入外，需要对学生强调工具的正确使用途径。在了解爬虫带来的方便同时也需要清晰认知到与道德、法律的关系，善用工具，遵纪守法，网络从来不是法外之地。

参考文献：

[1]方芳.网络爬虫课程思政元素挖掘与融入的实践研究[J].电脑知识与技术，2022，18（23）：125-126+180.DOI：10.14004/j.cnki.ckt.2022.1564.

[2]刘业，吴建平.动态可配置网络爬虫系统的形式化研究[J].福建电脑，2022，38（08）：1-4.DOI：10.16707/j.cnki.fjpc.2022.08.001.

[3]王国华. 基于 python 的豆瓣电影网络爬虫设计与分析[C]//.第三十六届中国（天津）2022’IT、网络、信息技术、电子、仪器仪表创新学术会议论文集.，2022：212-215.DOI：10.26914/c.cnkihy.2022.015025.

[4]龙辉.基于 Ajax 的聚焦网络爬虫技术在科研项目管理系统中的应用[J].电子元器件与信息技术，2022，6（07）：8-11.DOI：10.19772/j.cnki.2096-4455.2022.7.003.

[5]李森.非法提供网络爬虫技术行为的刑法规制[J].南京航空航天大学学报（社会科学版），2022，24（03）：69-74.DOI：10.16297/j.nuaass.202203012.

[6]张正阳，任保见，刘娜.Python 网络爬虫在农业网络数据获取中的研究[J].现代化农业，2022（07）：50-53.

[7]李洋，兰元帅，徐康，王国臣，郭澜，米豪.基于树莓派的访客指纹门禁打卡系统的设计与实现[J].电脑编程技巧与维护，2021（06）：39-40.DOI：10.16184/j.cnki.comprg.2021.06.013.

[8]习近平在全国高校思想政治工作会议上强调：把思想政治工作贯穿教学过程中开创我国高等教育事业发展新局面[N].人民日报，2016-12-09（1）