基于网络爬虫和数据可视化技术的应用型高校人才培养方向研究

接 辉

（江西应用科技学院，江西 南昌 330100）

摘要：随着我国高等教育的普及，高校毕业生就业问题日益突出。人才培养能否适应社会需求，已经成为制约普通应用型高校发展的重要因素。本文从社会人才需求着手，使用网络爬虫和数据分析技术，采集并分析某知名招聘网站的招聘信息，研究各行业人才需求数量、工资待遇、学历和经验要求、知识技能要求等情况，为应用型高校人才培养提供有益的启发。

关键词：网络爬虫；数据分析；招聘信息；人才培养

近年来，我国高等教育的普及，高等教育规模急剧扩大。国家教育部、统计局数据显示，2023 年以来全国普通高等院校每年本专科招生和毕业数量均已突破 1000 万，在校学生超过 3500 万。如此大规模的学生该学习哪些专业知识来更好地满足社会发展需要，是高等教育届，尤其是应用型高校亟需深入研究、认真对待的问题。当今社会，各类人才服务网站和招聘信息平台已普及使用，使得利用网络爬虫、数据可视化等技术，能够及时采集分析市场需求信息，为应用型高校人才培养提供决策依据。

一、招聘信息采集分析系统

（一）系统总体设计

系统采用 Python 网络爬虫技术从国内某知名招聘网站较为全面地采集北京、上海、深圳、南昌 4 个热门城市招聘信息。将信息进行预处理后，按照招聘岗位的行业、城市、学历、工作经验、工资待遇、技术要求等进行分析研究。

在信息采集过程中，系统采用 Selenium 自动化测试工具采集页面数据，使用 XPath 工具解析页面数据，在一定程度上规避了网站反爬虫策略、提升了数据解析效率。

（二）数据采集过程

通过观察分析，目标网站招聘信息分为互联网 /AI、电子 / 电气 / 通信、产品、客服 / 运营等 22 个行业，每个行业又分为若干职位。观察易知，用户查询招聘信息时，网站以 post 方式提交用户请求。为此，我们将信息采集分为行业信息采集和招聘岗位信息采集两个部分。行业信息采集部分负责从网站首页获取页面数据，解析出所包含其中的行业、职位等信息，并将其保存。招聘岗位信息采集部分，将采集到的行业、职位信息生成查询字符串，使用 Selenium 操作浏览器打开相应页面，然后 XPath 工具进行解析，采集到招聘岗位数据并保存下来。

招聘岗位信息采集部分主要程序代码如下：

```
with open（'./nc_boss_industry.json'，'r'） as f：
  d=json.load（f）  # 打开行业数据
for p，positions in d.items（）：  # 对所有行业进行遍历
  position_names=list（positions.keys（））
  for position_name in position_names：
    for jobs in list（positions[position_name]）：
      for job_type，job_url in jobs.items（）：  # 对所有职位进行遍历
        wd.get（job_url）  # 使用 selenium 获取招聘岗位列表页面数据
        while joburl_divs==[] or jobname_divs==[]：  # 等待读取到数据
          try：WebDriverWait（wd，60，1）.until（EC.presence_of_element_located（（By.CSS_SELECTOR，'div[class="job-card-body clearfix"] span[class="job-name"]'））） # 等待页面元素加载成功
          time.sleep（15）  # 暂停 15 秒，减轻服务器压力
          sub_tree=etree.HTML（wd.page_source）  # 构造 xpath 文档树
          joburl_divs=sub_tree.xpath（'//div[@class="job-card-body clearfix"]/a'）  # 使用 xpath 解析页面数据
          # 将新采集到的数据追加写入文件
```

（三）数据采集结果

本文共采集到共计 50 余万条招聘信息，包含工作岗位名称、工资、技能要求、学历要求、公司名称、公司概况等信息。

二、数据分析过程

（一）数据预处理

采集到的数据中，工资、工作经验、学历等都是以自然语言表达的，表达格式相对固定。我们使用正则表达式对文本进行处理，得到相应的数据。然后删除工资数值、工作经验年限等数据中的异常或缺失较多的数据。最后对数据进行标准化，便于后续分析处理。

（二）总体情况

经过预处理后，最终得到有效数据 495716 条，分为 24 个行业、142 职位类别，总体平均工资 13998.06 元，工资中位数 10000.00 元。

（三）分行业分析

招聘岗位最多的行业为互联网 /AI、零售 / 生活服务、生产制造，其平均工资水平在 22 个行业中分别排名第 6、22 和 23。招聘工资最高的行业为采购 / 贸易、金融、能源 / 环保 / 农业，其招聘岗位数量排名分别为 21、17、23。

为平衡工资和岗位数量对分析结果的影响，我们将这两列数据进行极差标准化后放大 100 倍，作为招聘岗位的工资指数和岗位指数。然后将工资指数和岗位指数按照各 50% 的权重构成就业指数，兼顾工资收入和行业人才需求。

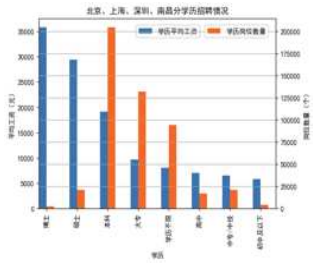
2.4 分学历分析

目标网站将学历要求分为博士、硕士、本科、大专、高中、中专 / 中技、初中及以下、学历不限共 8 个层次。总体来看，工资随着学历提升程指数级提高，岗位数量则呈现以本科、大专为中心的大致正态分布。唯一的特例是学历不限的岗位招聘数量和平均工资均明显高于高中以下学历层次。如图一

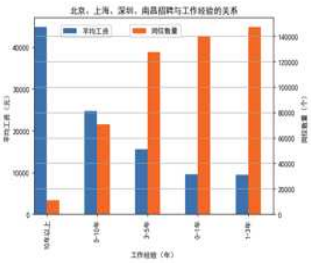
（五）按经验要求分析

目标网站将工作经验分为 0–1 年、1–3 年、3–5 年、5–10 年和 10 年以上 5 个层次。招聘工资随工作经验程指数级增加，岗位

数量则随工作经验增加而明显减少。有一个特例是工作经验要求 3–5 年的岗位数量最多，但工资却略少于 0–1 年。如图二



图一



图二

（六）招聘岗位关键词分析

采集到的招聘岗位描述关键词共计 124685 个，其中一些是对岗位的宽泛要求，与专业技术无关。筛选出出现次数大于 400 的专业技术热门关键词，并对内涵基本一致的关键词进行合并，列出对应的课程，前 25 个如下表一。

表一

序号	关键词	出现次数	对应课程
1	3D Max、SketchUp、SolidWorks、Maya、C4D、UG、Catia、3D 设计	15116	3D Max 建模基础、SketchUp 应用、SolidWorks 三维产品设计与建模等
2	C/C++、Objective C	11974	C 语言程序设计
3	AutoCAD、CAD	11200	AutoCAD 计算机辅助设计
4	PhotoShop	10702	PhotoShop 图像处理
5	Python	10222	Python 程序设计
6	短视频、抖音	8965	短视频运营、Premiere 视频剪辑等
7	数据分析	8679	Python 数据分析
8	Java	7014	Java 程序设计
9	SQL、MySQL、SQL Server	6624	数据库原理与应用、数据库技术等
10	软件测试、自动化测试	5777	软件测试技术
11	前端开发、HTML、CSS、JavaScript	5152	Web 前端开发基础
12	人工智能、AI	5142	机器学习、深度学习、计算机视觉、自然语言处理等
13	AE	2825	After Effects 动效设计
14	Unity3D、UE4	2682	Unity3D 虚拟现实开发、Unreal Engine 4 开发入门等
15	深度学习、TensorFlow、PyTorch	2652	深度学习
16	Premiere	2579	Premiere 视频剪辑
17	视觉设计	2342	PhotoShop 图像处理、Illustrator矢量图形编辑、Premiere 视频编辑、广告创意设计等
18	大数据、Hadoop	2324	Hadoop 大数据技术
19	项目管理	2201	软件工程
20	Linux	2182	Linux 操作系统
21	机器学习	1978	机器学习
22	Go、Golang	1849	Go 语言程序设计
23	C#	1776	C# 程序设计
24	运维、网络运维、日常运维	1660	Linux 操作系统、计算机网络、云计算与虚拟化技术、网络安全等
25	大模型	1451	AIGC 基础与应用等

表中内涵较为单一、对应课程较为具体的热门关键词为 C/C++、AutoCAD、PhotoShop、Java、Python、数据分析、软件测试、前端开发、AE、深度学习、Premiere、大数据、Linux、机器学习、Go、C# 等。

三、结论

（一）数据采集技术

采用 Selenium 自动化测试工具采集数据比较接近真实用户访问网页，能够较好地规避反爬虫策略、降低服务器压力，但采集效率较低。本文采用 Selenium 无头浏览器模式，并采用 XPath 进行数据解析，数据采集效率明显提升。

（二）学科专业建设

从平均工资和岗位数量两个方面考虑，互联网 /AI 行业具有明显的优势，医疗健康、金融、采购 / 贸易、市场 / 公关 / 广告、能源 / 环保 / 农业等行业相对靠前，高校专业建设可重点考虑相关专业；产品、人力 / 行政 / 法务、财务 / 审计 / 税务、酒店 / 旅游、餐饮等行业比较靠后，高校专业建设应尽量规避此类专业。

（三）人才培养层次

大专以上学历对工作收入影响显著，高中、中专 / 中技以下学历对工作收入没有帮助。3 年以上工作经历对工作收入影响显著，3 年以下工作经历对工作收入没有帮助。院校在人才培养中，应鼓励提升学历，注重培养专业热情和工匠精神。

（四）课程体系建设

C 语言程序设计、AutoCAD 计算机辅助设计、PhotoShop 图像处理、Python 程序设计、Java 程序设计、软件测试技术、Web 前端开发基础、After Effects 动效设计、深度学习、Premiere 视频剪辑、Hadoop 大数据技术、软件工程、Linux 操作系统、机器学习、Go 语言程序设计、C# 程序设计、3D Max/SketchUP/Solidworks 等课程对应的关键词较为热门，相关专业可重点考虑开设此类课程。

四、研究展望

本文根据网站招聘信息，分析当前人才市场需求情况，作为应用型高校人才培养的参考依据。思路和结论有一定的新意，但数据资源还不够丰富，研究方法还比较粗略，研究结果还比较肤浅。后续应继续丰富数据来源、充分发掘数据内涵，以便为人才培养提供更加精准、深入的决策依据。

参考文献：

[1] 黄媛 . 基于网络爬虫技术的网络招聘信息分析 [J]. 长江工程职业技术学院学报，2024，41（03）：30–34.

[2] 蔡文乐，秦立静 . 基于 Python 爬虫的招聘数据可视化分析 [J]. 物联网技术，2024，14（01）：102–105.

基金项目：江西应用科技学院 2021 年度校级科学技术研究项目“基于网络爬虫和数据可视化技术的应用型高校人才培养方向研究”（编号：JXYKKJ–21–10）。