目　录

基于 Python 的招聘网站爬虫及可视化的设计与实现

摘要：现在，随着互联网网络的飞速发展，人们获取信息的最重要来源也由报纸、电视转变为了互联网。互联网的广泛应用使网络的数据量呈指数增长，让人们得到了更新、更完整的海量信息的同时，也使得人们在提取自己最想要的信息，过滤掉对自己无用的信息时变得不那么容易，对于应聘者也是如此。由于招聘网站的日益流行，也使得应聘网站成为了应聘者找工作的主要平台。在面对着大量的招聘信息时，就业者不能一目了然的获取自己想要的招聘信息，因此我们需要对海量的招聘数据进行处理，做出一种招聘信息的分析系统。在此基础上本文介绍了基于 Python 的招聘网站的爬虫及可视化的设计与分析过程中的技术线路。

本招聘网站的爬虫及可视化使用 Python 语言编写，使用基于 Flask 的轻量级 Web 应用框架，数据库使用 MySQL，使用 ECharts 进行数据可视化部分的显示。对数据的爬取使用的 Requests 进行爬取数据，本次爬取的招聘网站为拉勾网搜索关键词为 Java、Python、Php 的招聘信息，拉勾网具有较强的反爬虫机制，采用 Cookie 形式进行封装，再进行数据的获取。在 MySQL 数据库中存储爬取的招聘信息，用 Pymysql 包连接 MySQL 数据库将查询的数据使用 ECharts 框架展示到网页。

通过本系统可以用户可以了解到职位的信息概况、薪资分布情况、企业主要招聘城市情况和企业的规模分布、职位的福利待遇和对应聘者的学历要求，工作经验的要求。

关键词：Python 爬虫；数据可视化；招聘网站

Design and Implementation of Recruitment Website Crawler and Visualization Based on Python

Abstract: With the rapid development of the Internet, the most important source for people to obtain information has changed from newspaper and TV to the internet. The wide application of the Internet makes the amount of data on the network grow exponentially. While people get a lot of new and more complete information, it also makes it difficult for people to extract the information they want most and filter out the information that is useless to them. This is also the case for job applicants. With the increasing popularity of recruitment websites, recruitment websites have become the main platform for job seekers to find jobs. Faced with a large number of recruitment information, the employees can not get the recruitment information they want at a glance, so we need to process the massive recruitment data, to make a recruitment information analysis system.On this basis, this paper introduces the technical lines in the design and analysis of the crawler and visualization of the recruitment website based on Python.

The crawler and visualization of this recruitment website are written in Python language, using the lightweight Web application framework based on Flask, using MySQL database, using ECharts for data visualization part of the display. The crawler network has a strong anti-crawler mechanism, encapsulated in the form of cookie, and then collects data. The crawler system can be used to retrieve data from Java, Python, PHP, etc. The obtained information is stored in the MySQL database, and then the queried data is displayed on the web page using the ECharts framework using Pymysql package to connect to the MySQL database.

Through this system, users can understand the information of the position, the distribution of salaries, the main recruitment of enterprises in the city and the size of the distribution of enterprises, the benefits of the position and the requirements of the applicant's educational background, work experience.

Keywords: Python Crawler; Data Visualization; Recruitment Website

引　言

随着互联网的不断发展，网络招聘也更加普遍。招聘网站能使招聘者随时随地了解到招聘信息，同时提高企业招聘的速度。但是面对着大量的招聘信息，应聘者难以在很快的时间内找到适合自己的岗位，做出适合自己选择，也不能根据这些信息直观的看到应聘者比较关心的薪资状况分布，企业的主要招聘城市，公司福利和所要求的学历与经验等方面。因此，当下需要一个能够把招聘信息整合到一起并将信息可视化显示的系统，这样用户就可以通过该平台来进行查看招聘信息的薪资分布，企业福利，所在城市等，从而使求职者可以更快找到心仪的工作。

目前，基于网络爬虫的招聘职位可视化系统在国内外比较少见，有提供该平台的搜索引擎如百度、谷歌等。但是因为招聘网站的招聘信息不能够随意转载，并不能获取到全面的招聘信息，且做不到可视化的效果。因此基于 Python 的招聘信息的爬虫及可视化系统还没有比较成功的案例。

所以，本文通过对拉钩网 Java、Python、Php 相关岗位的公司名称、招聘城市、岗位名称、薪资待遇等进行爬取，然后将招聘信息存入数据库，使用 ECharts 可视化图表将招聘信息以柱状图、折线图等形式展现出来，供用户个性化的获取信息。让计算机相关专业应聘者根据自身优势有选择性的应聘岗位。为广大的社会择业人员和初入社会的应届毕业生提供就业和学习的指导方向。

第 1 章 课题概述

由于近些年互联网的飞速发展，我们所生活的世界正在被数据所淹没，人们面对大量的数据需要从大量数据中快速地提取有效的自己需要的信息。对于求职者来说当查看招聘信息时也是这样，面对招聘网站展示的大量的职位信息，应聘者难以及时选出自己最想要的职位信息，又或者筛选出信息后不能直观地看到招聘所有信息的特征、规律、变化的趋势或者数据之间潜在联系。我们可以借助计算机技术来进行自动获取筛选分析自己想要的职位信息。本文对于基于 Python 的招聘网站的爬虫及可视化的课题研究就显得尤为重要了。

1.1 课题内容

该课题研究的是一种基于 Python 的招聘网站的爬虫及可视化的系统。在开发过程中利用 Python 对招聘信息进行收集和分析。首先，在拉钩网站上爬取招聘信息，然后存入数据库，连接数据库将数据库中的招聘信息从地区、行业、专业、公司规模、要求经验、薪资待遇等维度进行数据分析。最后，利用 ECharts 可视化技术，将有效的数据展示给用户。

1.2 课题背景

近年来随着我国计算机水平的发展，计算机行业的热门，高校也都相继开设了相关课程，越来越多的计算机人才涌入社会，但市场中的一众岗位让人眼花缭乱，同时众多拥有丰富从业经验的从业者，名牌大学与普通院校毕业生共同竞争，致使很多社会中的求职者面临着就业的困扰，而如今的招聘网站信息多，想要获取有效的信息需要的时间太长。为了解决社会二次择业人员和高校应届毕业生获取符合自己的并符合自己意向的招聘岗位信息，利用 Python 对这些招聘信息进行收集和分析势在必行。所以需要一种能够具有分析岗位优势，薪资分布等的系统，可供求职者利用自身优势，分析岗位信息，从而尽快找到心仪的岗位。

通过综合运用互联网数据爬虫技术和图表可视化库，对招聘网站的招聘信息进行爬取，并进行了相关统计分析，从地区、行业、薪酬、经验、岗位素质等方面进行综合分析。从而帮助计算机行业想从事 Java、Python、Php 相关岗位的就业人员了解相关领域的岗位需求和薪资情况、企业招聘城市、招聘企业的规模和学历与工作经验要求等。从而为就业人员的快速选择岗位，在何处选择岗位提供参考，对未来的生活和工作、学习规划等明确方向。

1.3 课题意义

对于即将毕业找工作的应届生和社会择业人员来说，上网快速找到合适的工作，无疑是急需的。而如今的招聘网站信息多，面对着网上形形色色的招聘网站和参差不齐的招聘信息，想要获取有效的信息需要的时间太长，这给就业者根据自身的情况选择自己适合的职业带来了困难。针对以上不足，有必要通过爬虫技术，帮助求职者在杂乱无序的数据中寻找有用的数据，科学分析，缩短求职者找工作的时间成本，帮助求职者快速择业。

本系统爬取了拉勾网站的计算机语言相关多种招聘信息，同学们可以通过选择本身应对的学历和想要的招聘岗位来选择查看相应的招聘信息。同时将这些信息可视化，可以方便同学们快速了解公司需求情况，这些可视化的部分包括薪资情况，企业情况，公司福利情况和学历情况。

1.4 运行环境

开发环境：Pycharm

开发语言：Python+JavaScript+ SQL

后台数据库：MySQL

开发环境运行平台：Windows 7/Windows10

1.5 相关技术

本项目是使用 Python 语言开发编写。使用 request 包进行对招聘网站的数据爬取；用 Pysql 连接数据库，获取数据；使用 Flask 框架将数据返回给前端，用 ECharts 对数据进行可视化展示，使用 Jieba 分词将语句分开。

1.5.1 Python 语言

Python 是由其他多种语言发展而来的脚本语言。Python 具有很强的可读性，比其他语言更容易上手，并跳过了编译的过程，不需要使用编译器。Python 语言是交互式的，我们可以直接运行代码。Python 支持面向对象的风格或者将代码封装在对象的编程技术，是一种面向对象的语言。Python 语言非常适合新手学习，因此作为计算机学生，在步入大学后，专业课程开设的第一门课就是计算机导论——以 Python 为舟，可见 Python 对于初级程序员来说是一种伟大的语言。

爬虫一般来说就是进行网络资源抓取，因为 Python 脚本特性，Python 容易配置，对字符处理十分灵活，Python 有着丰富网络抓取模板，让两者可以很好的链接在一起。对比其

他静态编程语言来说，Python 抓取网页文档接口更加简洁。抓住网页有时候需要模拟浏览器的行为，而 Python 具有很多第三方包。

1.5.2 Flask 框架

Flask 是一种使用 Python 开发的 Web 框架，可以说是以 Werkzeug 和 Jinja2 模板为核心。Flask 相当于一个内核，是一个非常简单的框架且易于扩展。

Flask 调用视图函数后，会将视图函数的返回值作为响应的内容，返回给客户端。一般情况下，响应内容主要是字符串和状态码。当客户端需要访问信息时，通常通过浏览器发起 HTTP 请求。Werkzeug 预处理接收的 HTT 请求进行路由分发，根据 URL 请求，找到具体的视图函数。Flask 框架处理接收的请求，在 Flask 中，通常是用 Flask 程序实例的 Route 装饰器来实现路由。调用视图函数，获取响应数据后，把数据传入 HTML 模板文件中，使用 Jinja2 模板渲染响应数据，然后由 Flask 返回响应数据给浏览器，最后浏览器处理返回的结果显示给客户端，Flask 请求响应图，如图 1-1 所示。

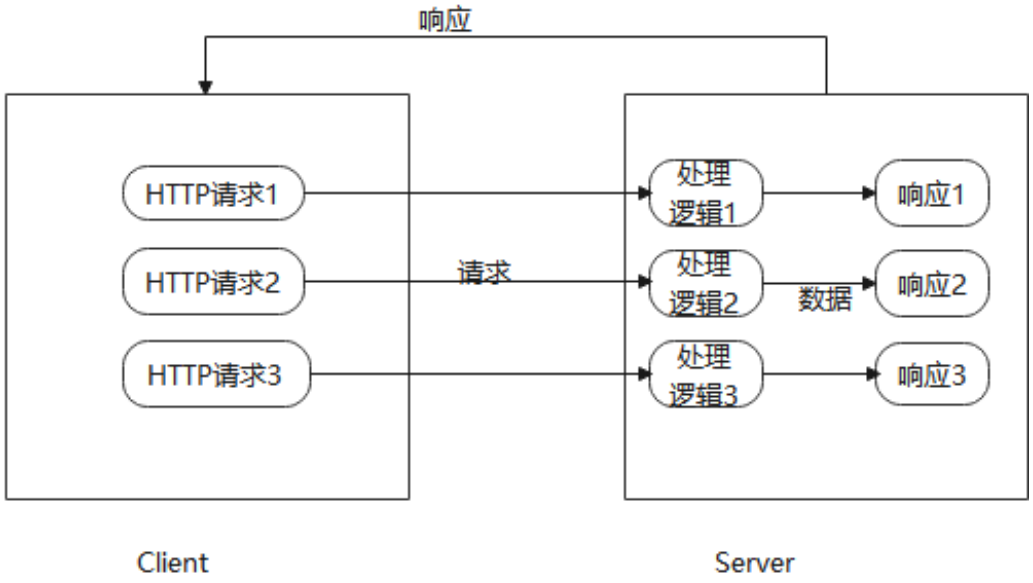

图 1-1 Flask 请求响应

1.5.3 ECharts 图表库

ECharts 是一款使用 JavaScript 实现的开源的数据可视化图表库，可以提供直观的，可交互的数据可视化图表。本招聘网站的爬虫及可视化系统使用 ECharts 做出薪资待遇的柱状图和饼图、折线图的分布展示，公司分布所在城市的饼状图展示，公司规模状况的柱状图和折线图的展示，对学历和工作经验要求的条状图和矩形树图的展示。ECharts 通常数据设置在 SetOption 中，如果我们需要异步加载数据，可以配合 JQuery 等工具，在异步获取数据后通过 SetOption 填入数据和配置项就行。

1.5.4 Jieba 库

Jieba 库是一款 Python 的第三方中文分词库，Jieba 支持三种分词模式：精确模式、全模式和搜索引擎模式。精确模式是将语句最精确的切分，这样不存在冗余数据，可以做文本分析。全模式是将语句中所有可能是词的词语都切分出来，速度很快，但是存在冗余数据。搜索引擎模式是在精确模式的基础上，对长词再次进行切分。

1.5.5 Pysql 包

Pysql 具有历史，完成和行编辑等功能。具有高级功能（搜索表，索引，计数，说明计划，会话列表等），为屏幕和文件提供适当的输出（CSV 可以包含在电子表格中），支持用户定义的 SQL，后台查询，模式数据模型，对象依赖项，PL / SQL 包函数调用树的图形输出，等等。在可视化需要数据连接查询后台数据库时，需要使用 Pysql 包进行连接。

1.6 本章小结

本章主要介绍招聘信息的爬虫及可视化在设计时所使用的软件 Pycharm 和 Mysql 以及平台背景，和介绍本次毕业设计所涉及到的一些技术和技术的相关内容并且阐述了课题意义，讨论了课题背景。为后面的招聘网站的爬虫与可视化系统的设计部分以及系统实现部分打下了坚实的理论基础。

第 2 章 系统设计

本招聘网站的爬虫及可视化系统使用的是 Python 语言编写，采用基于 Flask 的轻量级 Web 应用框架，招聘信息的存储数据库采用 MySQL 设计，使用 ECharts 进行招聘信息的数据可视化显示。

2.1 设计思想

首先使用 Request 爬取拉勾网网页，分析拉钩网网页，将内容进行解析后将招聘信息写入数据库，当我们需要查询信息时，需要连接数据库，将信息查询后读取并写入字典，使用 ECharts 框架，将数据传输到前端网页，以饼状图，柱状图，折线图等形式展示，让用户直观的看到招聘信息的地域，薪资，待遇等分布，让用户直观的获取到最关心的招聘信息。

2.2 需求分析

随着互联网时代的不断发展，各行各业的数据都呈现极为夸张的增长态势，面对毕业找工作，网上有形形色色的招聘网站，招聘信息也参差不齐，这给毕业生和二次择业人员如何根据自身情况选择自己适合的职业带来了困难。

本人想设计一个网站，爬取招聘网站的有关 Java、Python、Php 这三种语言相关职位的信息，然后将这些信息综合，方便同学们可以通过选择学历和想要的招聘职位来选择查看相应招聘信息。同时将这些信息可视化，方便同学们快速了解公司需求情况，可视化的部分包括薪资情况，企业情况，公司福利情况和学历情况。此系统的主要功能需求如下：

1．数据概况

爬取的所有有关 Java、Python、Php 语言的岗位招聘数据都可以看到，也可以通过学历和职位来选择查看满足条件的招聘信息，可以选择学历要求、输入职位来搜索更加精准的职位。

2．可视化

薪资情况：通过选择学历来查看各种岗位对于不同学历的薪资可视化情况，以柱状图、饼图的形式来展示各种职位的薪资分布、所占比例，提供给用户在找工作是作为参考。

企业情况：通过选择职位可以来查看这个职位的主要招聘城市，还可以大概查看一下这个职位的公司规模情况，以及每个职位在各个主要城市所占的比例饼图。

福利情况：通过数据可视化速览公司福利，基于词云进行构造，可以清晰看出所有公

司给出的最核心的福利待遇。

学历情况：可以查看各个职位对学历以及工作经验的要求，以条形图、矩形树的形式进行可视化展示。

2.3 系统可行性分析

对于本系统可行性的分析主要从与系统开发和实际生活息息相关的技术、经济、社会三方面进行分析。

1. 技术可行性

对于技术可行性首先要想到如何运用当前的技术手段可以成功地完成系统开发设计的工作，还要考虑设施以及配置能否契合开发的需要等。本次要开发的招聘数据采集分析网站系统用的是 Python 开发语言，容易编写, 可以直接在服务器上执行端口。并且使用 Pycharm 可以快速创建项目。在软件方面：由于使用 B/S 模型的相对成熟的开发软件, 所以软件开发平台的可行性。并且 ECharts 图表库也已非常成熟且完善，所以其技术可行性非常之高。

2. 经济可行性

Python 是一款开源免费的脚本语言，Pycharm 开发环境也有免费的社区版，而且 ECharts 也是一款优秀的开源的图表。因此开发成本几乎可以忽略不计，因此经济可行性非常高。

3. 社会可行性

本系统的开发符合国家法律进行，也不会触犯到任何人，任何集体的法律权益。只要开发过程中遵纪守法就完全符合法律要求，并且使用计算机的用户都会具有一定的计算机基础，并且本系统操作方法简单，分析的均为计算机相关方面的岗位信息，用户群体也都是计算机方面的人才，所以用户绝对能够熟练使用该系统，并且普通会使用计算机的人群也能使用。因此社会可行性很高。

2.4 功能设计

本项目要对系统功能结构进行设计、系统功能模块爬取网站信息及存入数据库和数据可视化设计、画出系统完整的流程图。

2.4.1 系统功能结构

该系统实现了数据的概览，薪资情况，企业情况，福利情况，学历情况及薪资预测的可视化。系统功能层次图，如图 2-1 所示。

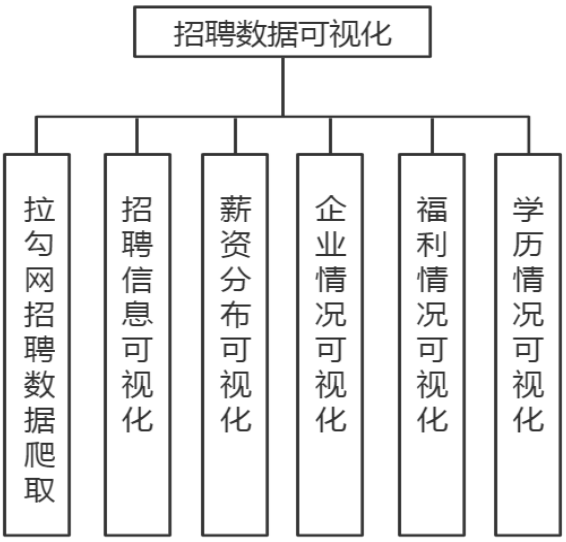

图 2-1 系统功能层次图

2.4.2 系统功能模块设计

本程序使用 Python 语言编写，使用的是 Flask 轻量级 Web 应用框架，数据库采用 MySQL 设计，使用百度开发的开源的 ECharts 图表库进行数据的可视化显示。招聘信息数据的爬取使用 Requests 进行，爬取的招聘网站为拉勾网，拉勾网有较强的反爬机制，所以采用 Cookie 的形式进行封装，再进行数据获取。获取的招聘信息数据存储到 MySQL 数据库，然后使用 Pymysql 包连接 MySQL 将查询的数据展示到页面。系统提供了如下功能：

1．数据爬取功能

程序模拟浏览器访问招聘网站信息获取响应 Json，提取其中招聘岗位的所有数据，并将这些招聘数据进行存储。系统爬取流程图，如图 2-2 所示。

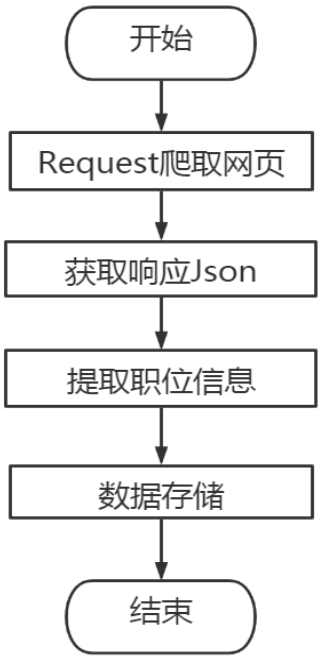

图 2-2 招聘信息爬取流程图

2．数据展示概况

可以通过学历和职位来选择查看满足条件的招聘信息，可以选择学历要求、输入职位来搜索更加精准的职位。

3．数据可视化

通过连接数据处理获取职位信息后，将职位信息传输到 ECharts 前端框架里。

在前端网站框架里放入连接数据后的 ECharts 将各种相关职位的薪资分布情况以柱状图、饼图的形式来展示。

将相关职位的主要招聘城市以所占的比例饼图形式展现；将公司企业的规模分布以折线图，柱状图的形式展示。

通过数据可视化，基于词云进行构造，生成公司福利词云和职位福利词云，展示所有公司给出的最核心的福利待遇。

可视化展现各种岗位对于不同学历和不同经验的薪资情况，以柱状图、矩形树的形式进行可视化展示。

2.4.3 系统流程图

用户登入系统后，通过连接数据库，对招聘信息进行获取，将信息传输到 ECharts 图表里对三种语言相关岗位招聘信息进行可视化的展示。可视化展示流程图，如图 2-3 所示。

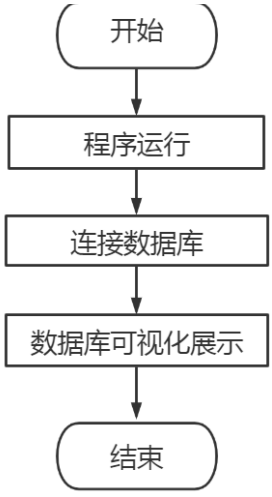

图 2-3 可视化展示流程图

2.5 数据库设计

数据库存储爬取的所有招聘信息数据。在 MySQL 里创建招聘信息表，存储爬取的招聘公司的全称，招聘职位名称，职位福利，薪资，学历要求，所在城市等信息。

当可视化界面展示数据时，查询数据中所有的相关招聘信息。招聘信息数据表，如表 2-1 所示。

表 2-1 招聘信息数据表

字段名	数据类型	备注
companyFullName	Text	公司全称
companyShortName	Text	公司简称
companySize	Text	公司规模
financeStage	Text	融资阶段
district	Text	区域
positionName	Text	职位名称
workYear	Text	工作经验
education	Text	学历
salary	Text	薪资
positionAdvantage	Text	职位福利
industryField	Text	经营范围
firstType	Text	职位类型
companyLabelList	Text	公司福利
secondType	Text	第二职位
city	Text	城市

2.6 本章小结

本章对基于 Python 的招聘网站的爬虫及可视化系统进行了设计思想的阐述，表述了招聘网站可视化的需求分析。对系统可行性进行分析，包括技术、经济、社会。以及详细的功能设计，包括系统功能结构设计、系统功能模块爬取网站信息及存入数据库和数据可视化的设计、系统完整流程图。最后介绍了数据库创建的表及其属性。为接下来的招聘网站的爬虫及可视化的详细实现打下基础。