项目说明文档

所属学校：青岛科技大学

参赛项目：齐鲁软件大赛

团队名称：GOODLUCK

团队成员：张莹、刘桐源、邵华、乔善宝、寻宇星

指导老师：周艳平

2017 年 9 月

第一章 前言

1.1 系统名称

基于互联网数据的大数据分析（v2.0）。

1.2 命题企业

山东乾云启创信息科技股份有限公司。

1.3 内容简介

本文档对“基于互联网数据的大数据分析（v2.0）”项目进行介绍及技术性说明。包括对项目的总体介绍、适用范围、开发流程、效果展示、技术说明补充等内容。具体代码实现参见源程序文件夹。

1.4 参考资料及数据来源

参考资料：

《python 网络数据采集》 Ryan Mitchell 著

《C++ GUI Qt4 编程》Jasmin Blanchette、Mark Summerfield 著

HTML5+CSS3 从入门到精通 --李东博 著

响应式 Web 设计:HTML5 和 CSS3 实战 --BenFrain (作者), 王永强 (译者)

JavaScript DOM 编程艺术 --基思 (Jeremy Keith) (作者), 桑布尔斯 (Jeffrey Sambells) (作者), 魏忠 (合著者)

JavaScript 高级程序设计 --泽卡斯 (Zakas. Nicholas C.) (作者), 李松峰 (译者), 曹力 (译者)

jQuery API 中文文档

Bootstrap3.0 官方文档

ECharts3 官方配置项手册

数据来源网站：智联招聘、前程无忧、拉勾网、腾讯招聘等。

第二章 总体介绍

2.1 开发目的

随着云时代的来临，大数据技术吸引了越来越多的关注，并且迅速兴起，为数据分析及前景预测等带来了极大便利。本项目基于互联网数据，对计算机行业的各方面信息进行分析整理，使计算机行业的职位需求、薪资状况、能力要求等数据直观表现出来。

2.2 技术背景

2.2.1 网络爬虫(python)

网络爬虫是搜索引擎抓取系统的重要组成部分。爬虫的主要目的是将互联网上的网页下载到本地形成一个或联网内容的镜像备份。

爬虫框架如图 2-1-1（来自网络@博客园 wawlian）。

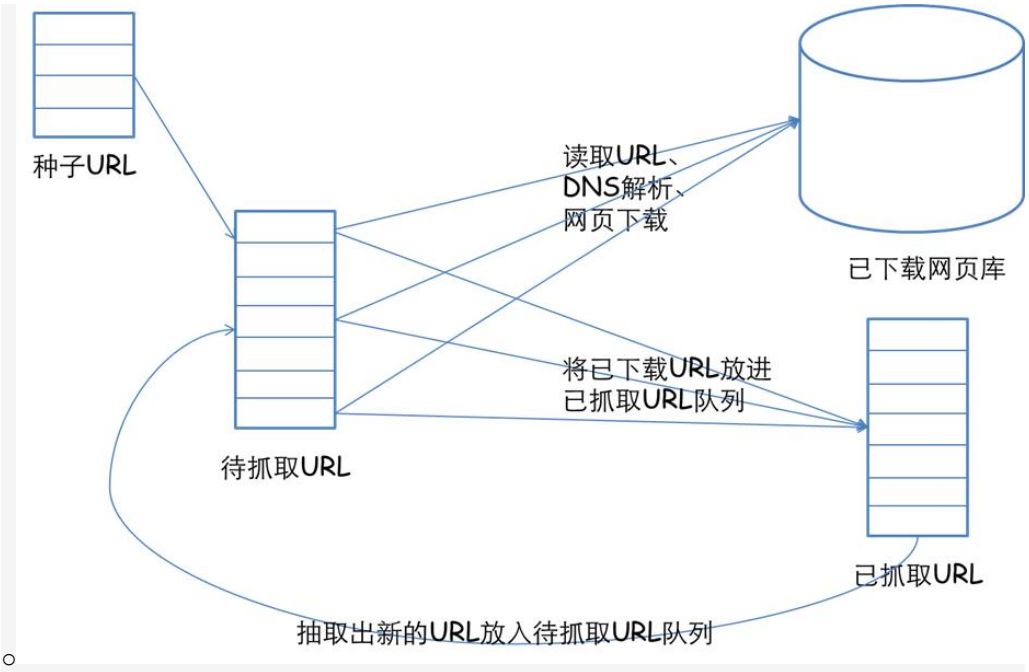


图 2-1-1

我们使用了 scrapy 构建了分布式网络爬虫，通过对热门招聘网站的爬取，获得了近一百五十万的数据， 我们将爬虫的速度尽可能的放慢，从而减轻被爬取网站的负担。

2.2.2 MySQL 数据库

MySQL 是一个关系型数据库管理系统，由瑞典 MySQL AB 公司开发，目前属于 Oracle 旗下产品。MySQL 是最流行的关系型数据库管理系统之一，在 WEB 应用方面，MySQL 是最好的 RDBMS (Relational Database Management System，关系数据库管理系统) 应用软件。

MySQL 是一种关系数据库管理系统，关系数据库将数据保存在不同的表中，而不是将所有数据放在一个大仓库内，这样就增加了速度并提高了灵活性。

我们将抓取的数据存储在 MySQL 中，便于进一步的数据清洗、分析以及可视化。

2.2.3 HTML

HTML 用来定义了网页的内容。 HTML 是用来描述网页的一种超文本标记语言 (Hyper Text Markup Language) ，而不是一种编程语言。标记语言是一套标记标签 (markup tag)， HTML 使用标记标签来描述网页。Web 浏览器（如谷歌浏览器，Internet Explorer，Firefox，Safari）是用于读取 HTML 文件，并将其作为网页显示。浏览器并不是直接显示的 HTML 标签，但可以使用标签来决定如何展现 HTML 页面的内容给用户

2.2.4 CSS

CSS 描述了网页的布局。CSS 指层叠样式表 (Cascading Style Sheets) ，样式将定义如何显示 HTML 元素 ，通常将样式存储在样式表中 ，而添加到 HTML 4.0 中，是为了解决内容与表现分离的问题 。外部样式表可以极大提高工作效率 ，通常把它存储在 CSS 文件中 ，多个样式定义可层叠为一 。

2.2.5 JavaScript

JavaScript 定义了网页的行为。JavaScript 是脚本语言，是一种轻量级的编程语言，是可插入 HTML 页面的编程代码。JavaScript 插入 HTML 页面后，可由所有的现代浏览器执行。

2.2.6 jQuery

jQuery 是一个 JavaScript 函数库，可以通过一行简单的标记被添加到网页中。Query 库包含 HTML 元素选取、HTML 元素操作、CSS 操作、HTML 事件函数、JavaScript 特效和动画、HTML DOM 遍历和修改、HTML DOM 遍历和修改 、AJAX、Utilities 等特性：

2.2.7 Bootstrap3.0

Bootstrap 是一个用于快速开发 Web 应用程序和网站的前端框架，是基于 HTML、CSS、JAVASCRIPT 的。响应式设计：Bootstrap 的响应式 CSS 能够自适应于台式机、平板电脑和手机。它为开发人员创建接口提供了一个简洁统一的解决方案，包含了功能强大的内置组件，易于定制。 它还提供了基于 Web 的定制，是开源的。

2.2.8 ECharts 3

ECharts，一个纯 Javascript 的图表库，可以流畅的运行在 PC 和移动设备上，兼容当前绝大部分浏览器（IE8/9/10/11，Chrome，Firefox，Safari 等），底层依赖轻量级的 Canvas 类库 ZRender，提供直观，生动，可交互，可高度个性化定制的数据可视化图表。而 ECharts 3 中更是加入了更多丰富的交互功能以及更多的可视化效果，并且对移动端做了深度的优化。

2.3 功能介绍

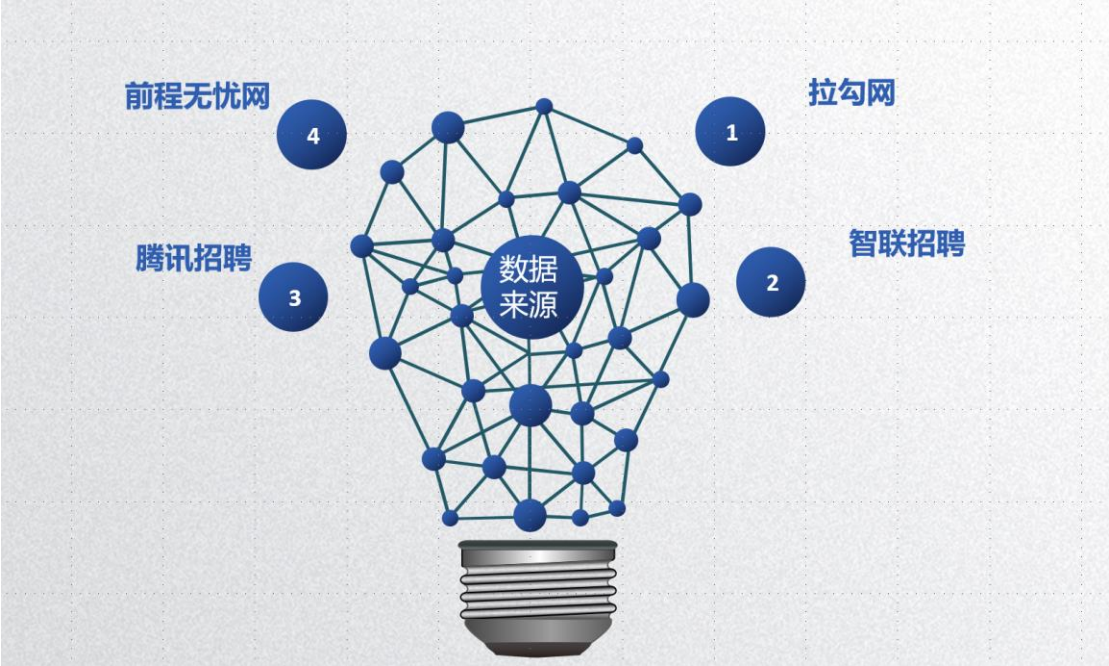
对以下数据进行整理分析以及图表可视化：
1. 公布有效数据：
爬取数据的来源分布 --> 极坐标系下的堆叠柱状图
不同网站不同领域比例(大数据，开发，测试，运维) -->嵌套环形图
2. 全国范围内的数据分布：
2.1 全国（平均）薪资分布情况 --> 地图散点图，薪资高的颜色深
2.2 不同领域的平均薪资所占比例 --> 【饼图】或者南丁格尔图
2.4 全国公司的城市分布 --> 地图散点图
2.3 在全国岗位需求量分布 --> 饼图或者地图散点图，岗位多的颜色深，
2.5 大数据在全国的岗位需求量分布 --> 柱状图或者地图散点图
2.6 计算机专业工作经验要求分布情况（应届、1～3 年、3～5 年） --> 嵌套环形图
2.7 全国各行业占比
3. 排行表：
3.1 计算机专业薪水最高的前 10 名招聘职位+岗位需求量 --> 折线图
3.2 大数据职位需求量最高的前 10 名城市+大数据前六个岗位需求量排行 --> 极坐标系下的堆叠柱状图-polar-stack-radial
3.3 大数据职位需求量最高的前 10 名行业（如互联网、金融、电子商务等） --> 南丁格尔玫瑰图
3.4 企业对哪类大数据人才需求最为迫切（大数据分析师、大数据架构师等）
3.5 计算机专业编程语言职位需求量前十名饼状图 --> 矩形树图
3.6 计算机专业编程语言平均薪资 –>柱状图
3.6 计 算 机专业不同编程语言平均薪资柱状图

4. 不同要素的关系图表：
4.1 学历和薪资(min,max)关系 --> 折线图
4.2 工作经验和薪资(min,max)的关系 --> 折柱图

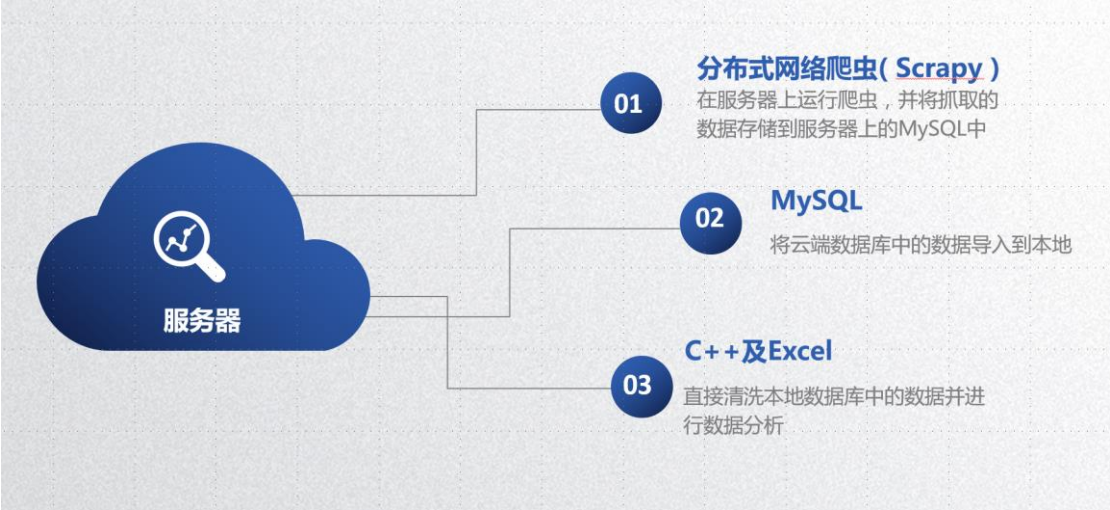
数据图可在柱状图、折线图、饼状图、表格等模式下切换。与地域分布有关数据则使用地图模式进行展示。

第三章 技术性说明

3.1 数据来源



3.2 数据爬取，存储，清洗以及分析



3.3 数据可视化