**基于招聘网站的离线统计及实时分析系统**

**作者**：谢俊炜，刘雨轩，夏余鑫，黄旺辉，黄嘉翔，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：爬虫；数据预处理；Hadoop；Spark；推荐算法。

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 大数据迅速发展成为科技界和企业界甚至世界各国政府关注的热点。随着当前信息不断累加，使得很多信息出现了过饱和的现象。但信息的增多不能很好的服务用户，反而使用户产生巨大信息迷茫之中。本案例选取基于招聘网站的离线统计及实时分析为主题，注重于新时代社会信息人才的缺口，抓住用户的人才需求、就业需求的痛点、痒点，着力于运用大数据对信息人才的就业的赋能。基于以上，本案例中展示了符合分布式大数据体系的一体化的岗位信息统计及推荐系统。在具体推荐系统实现上，案例综合了离线推荐与实时推荐，并采用协同过滤算法以及基于内容的推荐方法构建多维模型提供精准推荐。最后，案例还部署在分布式Web网站进行直观可视化分析，包含大屏展示宏观把控、细粒度数据垂直分析等。

**关键词**：数据挖掘，推荐算法，大数据，分布式

## 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为基于招聘网站的离线统计及实时分析系统，需引导学生进行的主要内容有：（1）使用爬虫工具从热门招聘网站爬取就业信息；（2）搭建大数据平台环境，对数据进行存储、计算；（3）构建符合分布式大数据体系的一体化的岗位信息统计、推荐系统。

## 2 背景介绍

近几年，大数据迅速发展成为科技界和企业界甚至世界各国政府关注的热点。《Nature》和《Science》等相继出版专刊专门探讨大数据带来的机遇和挑战。随着当前信息不断累加，使得很多信息出现了过饱和的现象。但信息的增多不能很好的服务用户，反而使用户产生巨大信息迷茫之中。人们对于大数据的挖掘和运用，预示着新一波生产力增长和消费盈余浪潮的到来。大数据赋能社会的需求已经成为当下最火热、最迫切的改革，它能从就业、教育、智慧城市建设等多方面赋能社会。

因此一定比例的大数据分析案例对于专业硕士的培养是必要的。在本案例中，我们对岗位缺口进行的数据分析，借助大数据分析与处理、数据挖掘的相关方法，以满足培养具备创新能力与实践能力的人才的需求。

## 3 内容

该案例的主要内容主要分为三个小节，分别为数据集收集、大数据平台环境搭建，以及一体化的岗位信息统计、推荐系统。

3.1 数据集收集

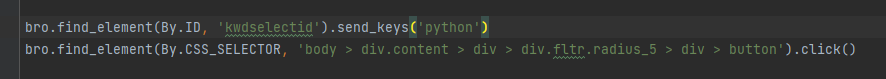
本案例爬取的是51job上的六种热门编程语言（java，python，c语言，c++，c#，php）的全国程序员岗位信息。



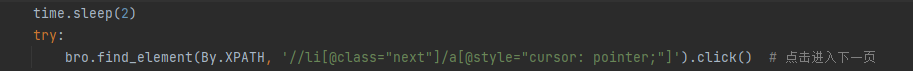
（1）首先利用webdriver调用谷歌浏览器的驱动，用谷歌浏览器打开51job网站首页。



（2）定位如下搜索框，输入想要爬取什么数据，例如python。点击搜索。



（3）跳转到所需页面后，开始爬取数据。为了避免访问太快导致ip被封，我们设置了延时爬取，每两秒爬取一次即每过两秒自动点击一次下一页按钮。



3.2 大数据平台环境搭建

在项目的实现过程中使用到了多种工具进行数据的存储、计算，下面介绍所需搭建的环境，主要包括Hadoop集群搭建、Spark集群搭建、Zookeeper、Kafka、Flume、MongoDB、Redis搭建。

选用Docker容器化技术在一台云服务器上搭建了一个1主3从的完全分布式Hadoop、Spark集群，由于Docker每次重启后，容器的IP会发生变化，因此在搭建环境前需先固定IP，固定后容器的配置如下：

|  |
| --- |
| docker network create --subnet=172.18.0.0/16 mynetworkgroup  第一台主机ip:172.18.0.2 名称：cluster-master 主节点  第二台主机ip:172.18.0.3 名称：cluster-slave1 从节点1  第三台主机ip:172.18.0.4 名称：cluster-slave2 从节点2  第四台主机ip:172.18.0.5 名称：cluster-slave3 从节点3 |

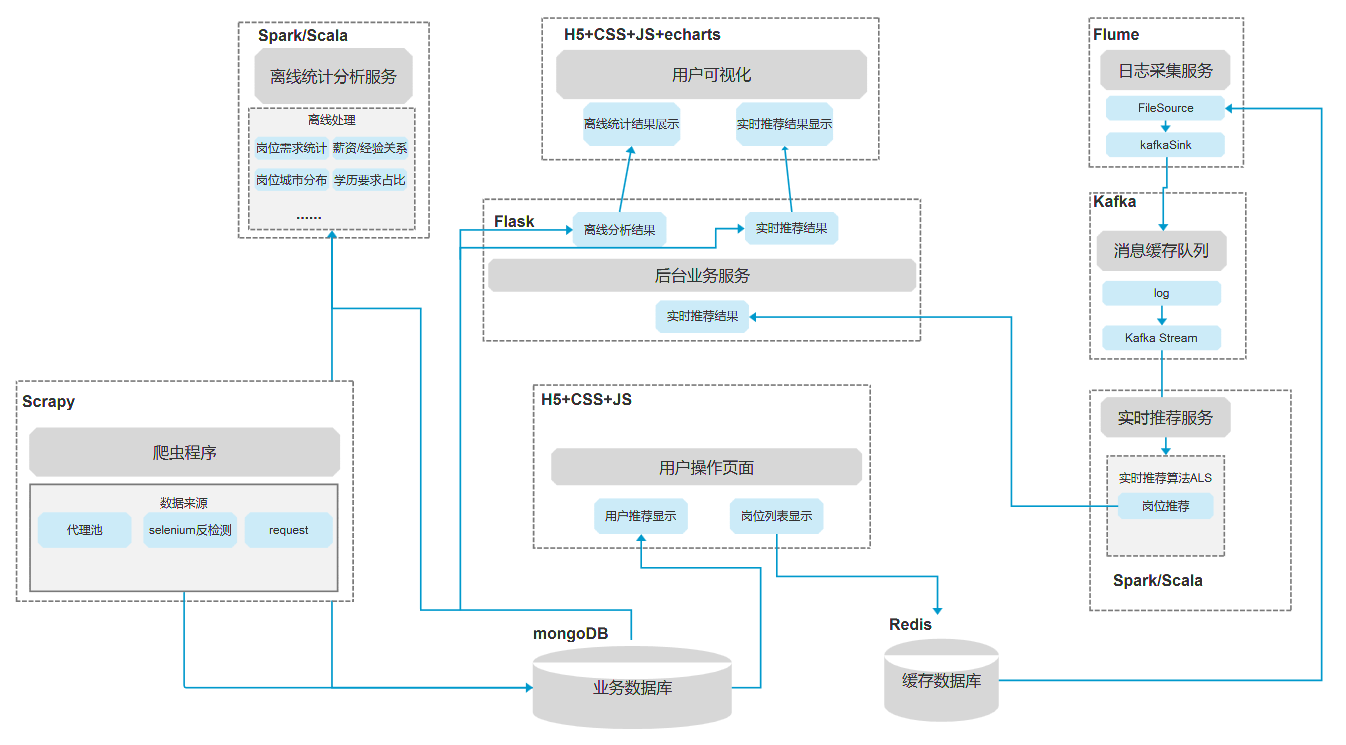
所需安装软件版本：

|  |  |
| --- | --- |
| 软件 | 版本号 |
| Jdk | 1.8 |
| Scala | 2.11.8 |
| Hadoop | 2.7.7 |
| Spark | 2.1.1 |
| Mongo | 4.4.9 |
| Redis | 5.0.14 |
| Zookeeper | 3.5.7 |
| Kafka | 2.11 |
| Flume | 1.8.0 |

3.3 一体化的岗位信息统计、推荐系统

项目以爬虫爬取到的高科技技术招聘领域知名的51job招聘系统数据集作为依托，构建了符合分布式大数据体系的一体化的岗位信息统计、推荐系统，包含了统计以及离线推荐与实时推荐体系，综合利用了协同过滤算法以及基于内容的推荐方法来提供混合推荐。提供了从分布式部署、大屏展示应用、业务系统、算法设计实现等多方位的闭环的大数据业务实现。

本次项目所采用的技术点有：Spark、Scala、MongoDB、Redis、Spark Streaming、Spark SQL、Flume、Zookeeper、Kafka、Flask、Echarts等，项目架构图如图所示：



## 4 小结

本案例选取基于招聘网站的离线统计及实时分析系统，需要使用爬虫工具从热门招聘网站爬取就业信息，搭建大数据平台环境，对数据进行存储、计算，最后构建符合分布式大数据体系的一体化的岗位信息统计、推荐系统。该案例主题新颖，结合了工业界现实需求与大数据分析与挖掘的多种理论与技术，可以充分增强学生的实践能力与理论基础。另外，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、推荐算法完成任务。

## 附录

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），scikit-learn（机器学习算法库），matplotlib（可视化绘图库）。

3. 本案例参考文献如下：

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

参考内容：数据预处理理论；数据挖掘