|  |
| --- |
| 全国云计算应用创新大赛 |
| **职位智能分析系统** |
| 系统简介概述 |

|  |
| --- |
| 参赛团队：EFISH  项目开发者：刘港  liu920848992@163.com  2018-3-10 |

# 项目背景

大数据和“云计算”像是一枚硬币的正反面一样慢慢勾勒出当今世界的财富价值风向。大数据的出现得益于互联网行业的快速发展、计算机硬件和软件能力的不断提升。大数据技术现已被应用到各行各业，而在招聘求职领域，我们希望通过爬虫技术、机器学习、文本挖掘、统计分析等手段帮助求职者更好的了解市场需求，从而有一个清晰、明确的求职方向。

# 系统需求分析

**通过对信息的挖掘分析，可以更加精准、清晰的指导求职者所在行业的待遇**

**水平、自身可能的待遇、以及对公司、行业的选择。**

## 需求一：网络爬虫

网络爬虫根据选定的网站列表（如智联招聘，51job 等）进行爬取，爬取网站公布的招聘信息。并将爬取到的信息进行整理后存储在数据库中。

## 需求二：数据存储转换

解决爬取下来的数据抽取、转换、加载到大数据平台的问题。即把MySQL的数据通过sqoop工具加载到Hadoop平台中的HBase数据库。

## 需求三：数据分析处理

解析爬取的数据，进行行业招聘态势的分析。通过文本信息提取、分词、关键词提取、词频统计等数据分析算法，得到有价值的结论，同时为求职者的职位智能推荐及网页展示功能做准备。

## 需求四：智能检索和推荐

分析选手的求职信息，设计数据分析算法为求职者推荐多元化的信息和匹配的职位。

## 需求五：应用开发和部署

开发可与用户进行交互的推荐系统，有快速和精准的推荐和良好的用户体验效果。

### 系统创新点

目前国内的大多数招聘网站在职位推荐上都采用协同过滤算法和基于内容的推荐算法。所用到的基于用户进行过滤和基于职位进行过滤虽然极大地提高了推荐的质量，但用户往往受限于职位的具体要求，或者对行业、所处的环境没有清晰的认识，无法更好地进行职位的选择。本系统根据这些需求作了以下创新：

**一、利用大数据与云计算技术，通过海量的数据分析，利用数据分析算法，将职位的行业分布情况，城市分布情况，薪酬分布情况、企业信息等进行数据分析。用户可以更加清晰地了解到各地区、各行业、各企业的薪酬分布情况，使用户能更加清晰地认识到自身所处环境，有利于求职者职业的选择。**



**二、利用jieba分词系统分词、提取关键字等技术，将职位的具体要求进行关键词提取，将提取到的关键词和其他要求与薪酬结合，形成一个list，使用欧氏距离或者均方根误差来计算出职位的符合程度，进而为求职者更加智能地推荐。**



**三、基于分词、提取关键字技术，对不同的工作职位进行对比选择，使用户更加科学，更加智能地对比选择工作。**



### 系统核心技术

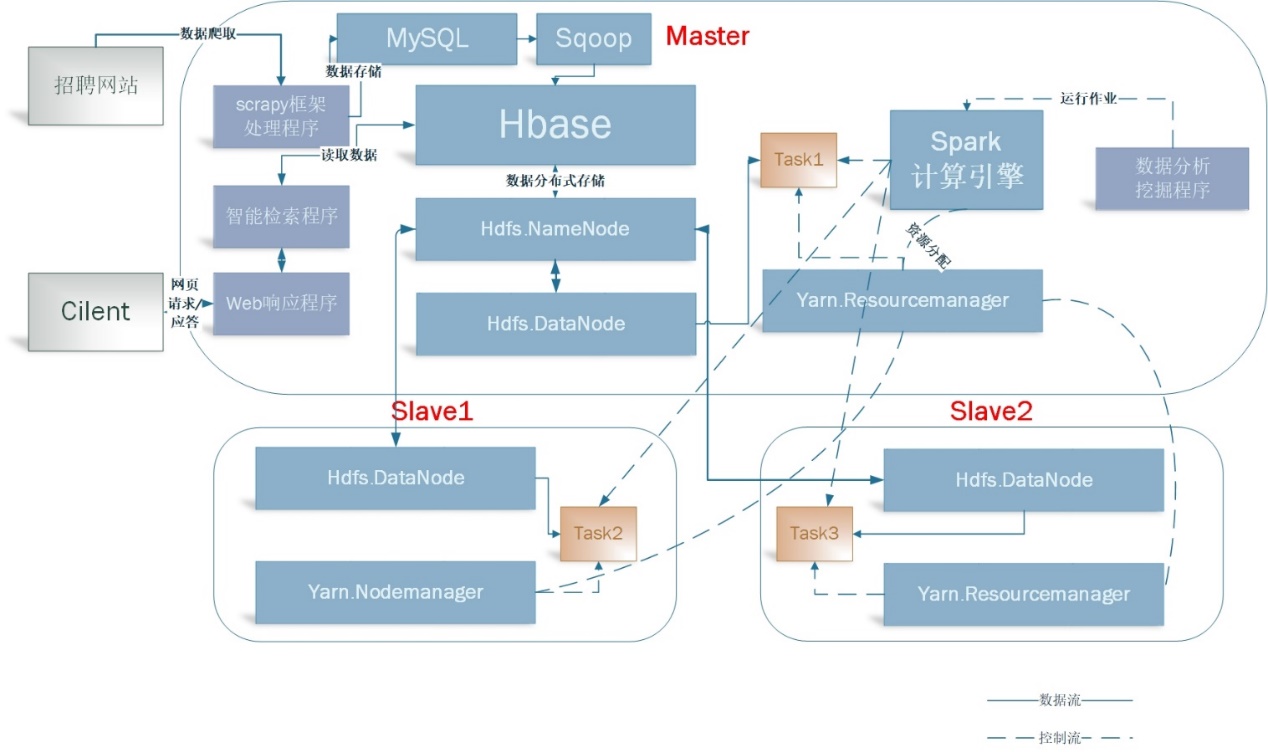
**基于Scrapy（Python）框架的信息收集的爬虫框架**

**基于分词，关键字提取、推荐算法和数据分析的相关算法**

**基于Hadoop生态圈的分布式存储和计算的框架**

**基于Django（Python）框架的网页部署和展示**

# 系统总体设计



系统应用环境采用3台服务器构成。

Master(主节点服务器):

1. HDFS分布式文件系统的NameNode节点

SecondaryNamenode节点、DataNode节点（实验条件限制，最大化利用资源）

1. Yarn资源管理的Resourcemanager节点
2. Spark计算引擎的Master节点，同时也作为Worker节点
3. Web应用部署服务器

Slave1 & Slave2（节点服务器）:

1. HDFS分布式文件系统的DataNode节点
2. Yarn资源管理的Nodemanager节点
3. Spark计算引擎的Worker节点

### 实现一、网络爬虫

1. 目标

对招聘网站公布的招聘信息进行爬取

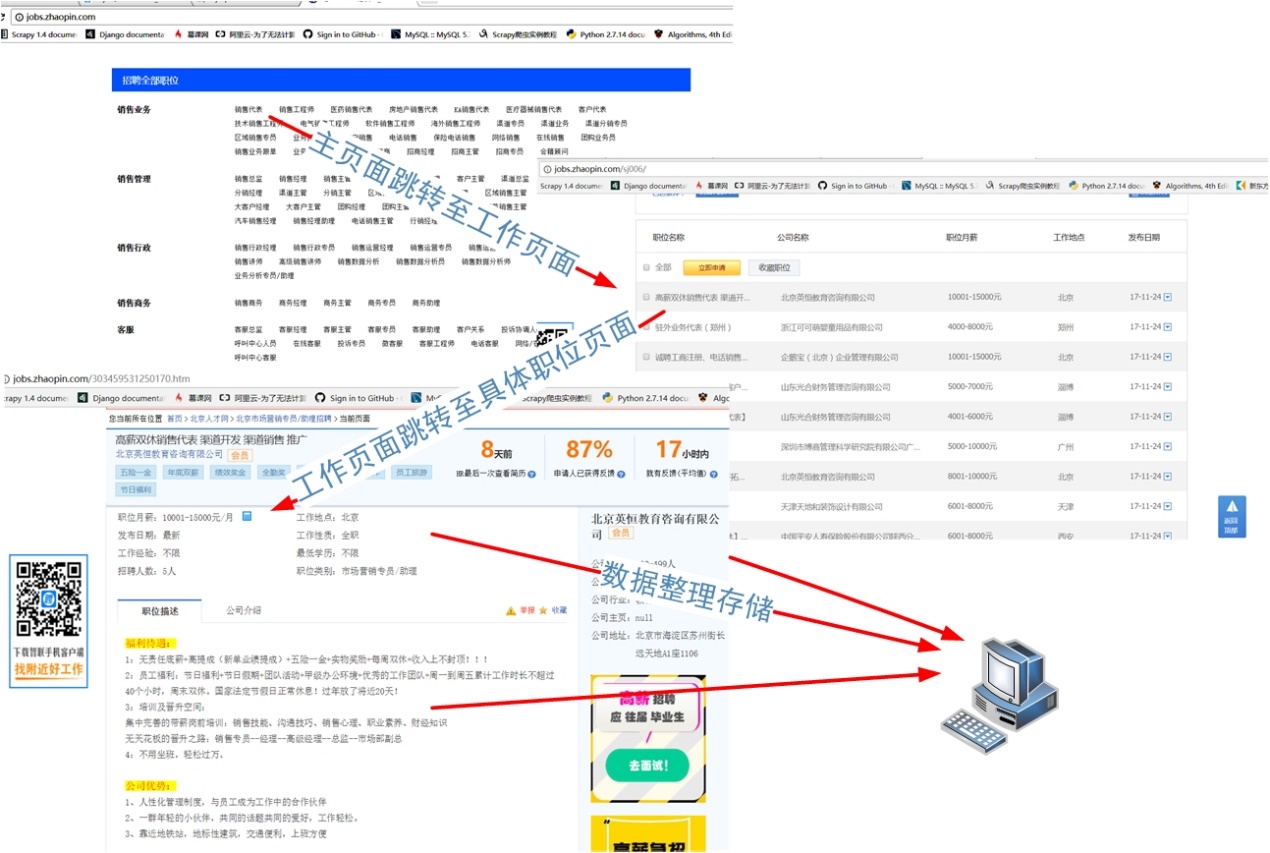
1. 工具

Python开源框架Scrapy、MySQL数据库

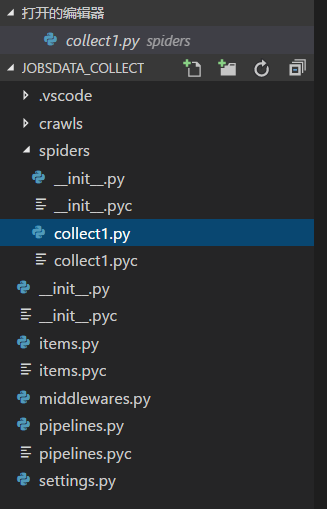
1. 爬取网站

智联招聘职位信息网（<http://jobs.zhaopin.com/>）

1. 爬取流程概述：



1. 爬虫框架Scrapy：

jobsdata\_collect

--crawls

----co1 #保存scrapy爬取状态

--spider

----collect1.py #爬取主程序

--items.py #爬取数据类

--middlewares.py #爬取中间件

--pipelines.py #管道程序，用于连接数据库等

--settings.py #爬虫主设置

### 实现二、数据存储

网络数据的存储分为两个方面：本地MySQL存储和分布式HBase存储

MySQL存储：

MySQL数据存储类型



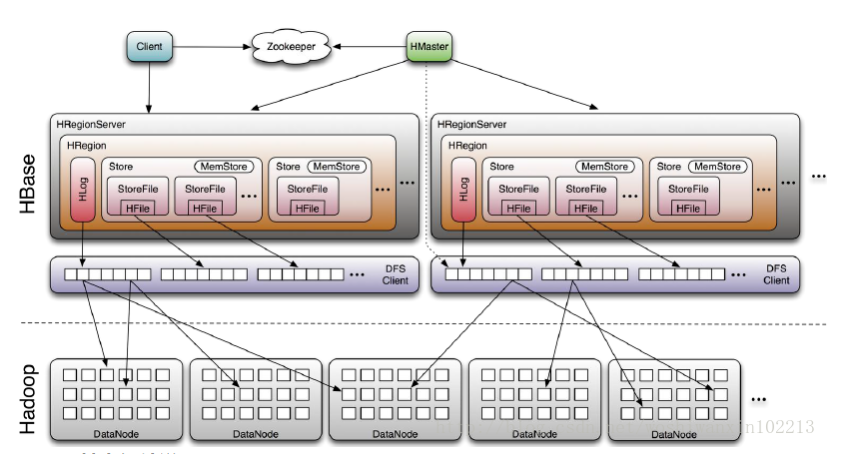
大数据的存储使得数据的查询更新等操作变得很缓慢，于是进行了索引优化和分表优化。

1.分表优化：把数据分为多个表，并对这些表创建一个索引表，每个表的大小为500000行，对于5000000行的数据，即可拆分为10个表。使得查询等操作更快。

2.索引优化：对每一个表进行b-tree索引优化，其操作速度大幅度加快。对于500000行数据，即使采用二叉树构建，速度也可加快100倍以上。（实际建立的树取决于字段的类型大小，字段类型占的空间越小，树的高度越低，查询次数越少。）

HBase存储：

HBase基于HDFS分布式存储系统，基本原理流程图如下：



HBase是一个构建在HDFS上的分布式列存储系统，为典型的key/value系统，主要用于海量结构化数据存储。从逻辑上讲，HBase将数据按照表、行和列进行存储。  
与Hadoop一样，HBase目标主要依靠横向扩展，通过不断增加廉价的商用服务器，来增加计算和存储能力。

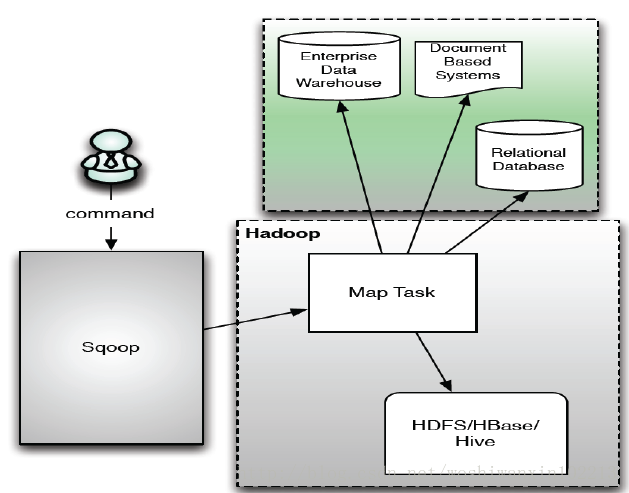
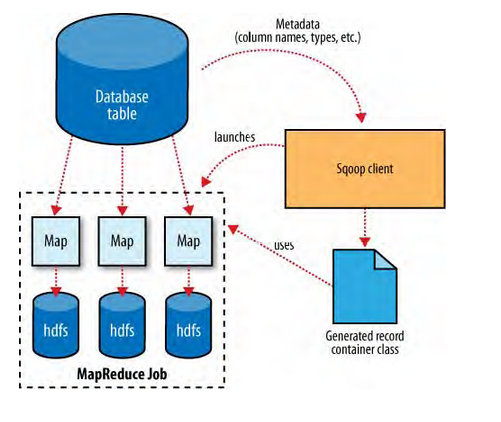
### 实现三、数据ETL

目标：

解决把爬取下来的数据从MySQL数据库中，把数据抽取、转换、加载到Hadoop平台，即把MySQL的数据通过etl工具加载到Hadoop平台中的HBase数据库。

工具：sqoop1.4.6

流程概述：



从传统数据库如MySQL获取元数据信息(schema、table、field、field type)，把导入功能转换为只有Map的MapReduce作业，在MapReduce中有很多map，每个map读一片数据，进而并行的完成数据的拷贝。

主要步骤：

1、 数据清洗：

·数据补缺：对空数据、缺失数据进行数据补缺操作，无法处理的做标记。

·数据替换：对无效数据进行数据的替换。

·格式规范化：将源数据抽取的数据格式转换成为便于进入仓库处理的目标数据格式。

·主外键约束：通过建立主外键约束，对非法数据进行数据替换或导出到错误文件重新处理。

2、 数据转换

·数据合并：多用表关联实现，大小表关联用lookup，大大表相交用join（每个字段家索引，保证关联查询的效率）

·数据拆分：按一定规则进行数据拆分

·行列互换、排序/修改序号、去除重复记录

·数据验证：lookup、sum、count

实现方式：

·在ETL引擎中进行（SQL无法实现的）

·在数据库中进行（SQL可以实现的）

3、 数据加载

方式：

·时间戳方式：在业务表中统一添加字段作为时间戳，当OLAP系统更新修改业务数据时，同时修改时间戳字段值。

·日志表方式：在OLAP系统中添加日志表，业务数据发生变化时，更新维护日志表内容。

· 全表对比方式：抽取所有源数据，在更新目标表之前先根据主键和字段进行数据比对，有更新的进行update或insert。

·全表删除插入方式：删除目标表数据，将源数据全部插入。

4、异常处理

在ETL的过程中，必不可少的要面临数据异常的问题，处理办法：

1、将错误信息单独输出，继续执行ETL，错误数据修改后再单独加载。中断ETL，修改后重新执行ETL。原则：最大限度接收数据。

2、对于网络中断等外部原因造成的异常，设定尝试次数或尝试时间，超数或超时后，由外部人员手工干预。

3、 例如源数据结构改变、接口改变等异常状况，应进行同步后，在装载数据。

### 实现四、数据预处理分析

目标：

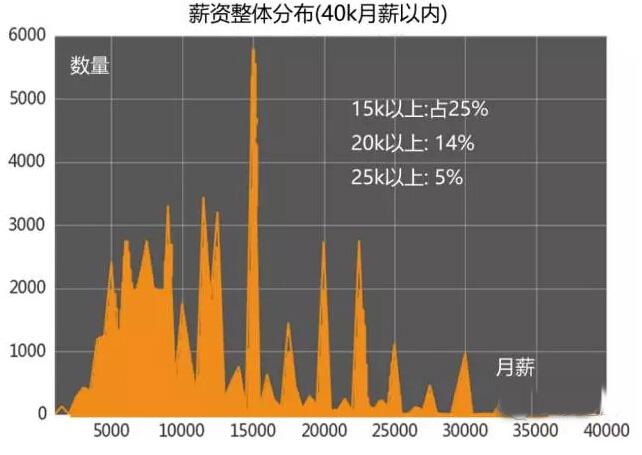
对所爬取的数据进行初步的认识，包括分析各行业，各地区的企业信息，职位的各种信息等，利用可视化技术进行直观认识，为数据挖掘和智能推荐作基础。

相关信息前500行如下：



主题一：职位的分析：

1. 职位薪酬分析：对于相同行业的职位进行横向分析。主要分析其薪酬的分布函数，极大（小）值，平均值等。



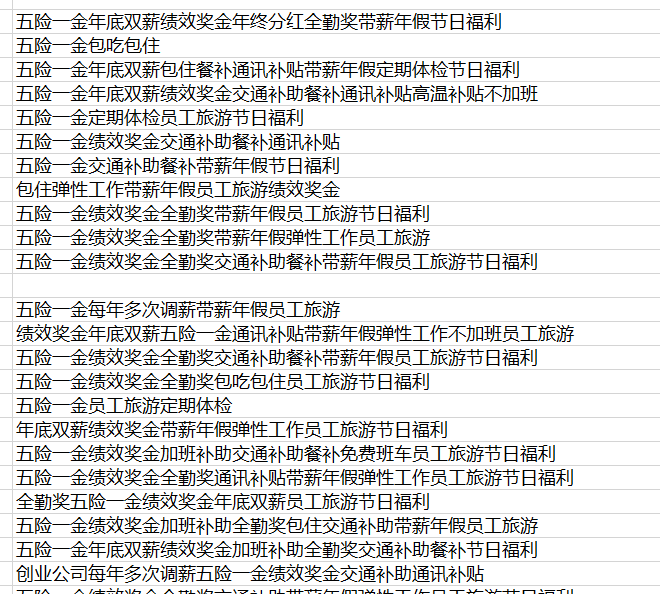
1. 职位地域分析：对于该行业的职位，分析各城市的需求状况。



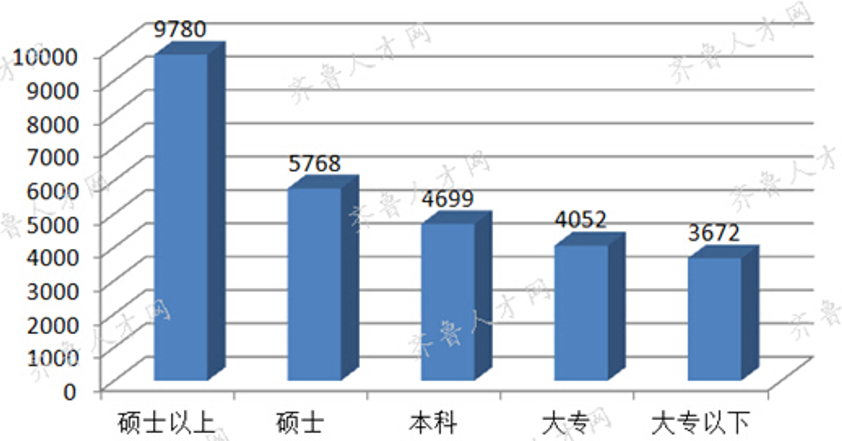
1. 职位工作经验要求分析：对于收集数据的工作要求，对于各行业工作经验的分布进行系统的分析



1. 职位福利分析：对于各个职位的工作福利进行提取关键字等方式，得到每个职位的工作福利信息。



1. 职位其他要求分析：如对学历、年龄及特殊要求的分析



使用语言及相关包：python。Pandas(数据结构化)、matplotlib(可视化)

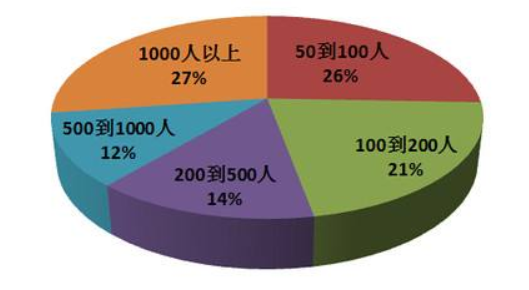
相关算法：文本数据结构化，提取关键字，分类算法，概率密度分布等。

主题二：企业的分析

1.对各职位的企业性质进行分析，企业性质包括民营、国企、合资、上市等



2．对企业的规模、行业等进行分类



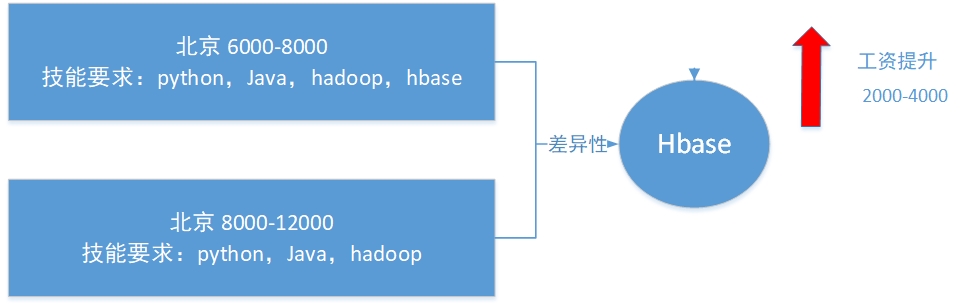
主题三、分词、关键词提取

利用jieba分词系统分词、提取关键字等技术，将职位的具体要求进行关键词提取，将提取到的关键词和其他要求与薪酬结合，形成一个list，并存储在数据库中。



### 实现五、智能推荐和检索

1. 对于提交的个人信息，根据基于内容的推荐算法。将提取到的关键词和其他要求与薪酬结合，形成一个list，使用欧氏距离或者皮尔逊相关度评价来计算出职位的符合程度，预测出求职者的薪酬水平及智能推荐工作。
2. 对于职位薪酬进行分段处理，对于不同的薪酬段进行纵向对比，挖掘出其中的共同点和差异性。并为职位的提升突破找出关键点，相关实例如下：





### 实现六、应用开发部署

目标：

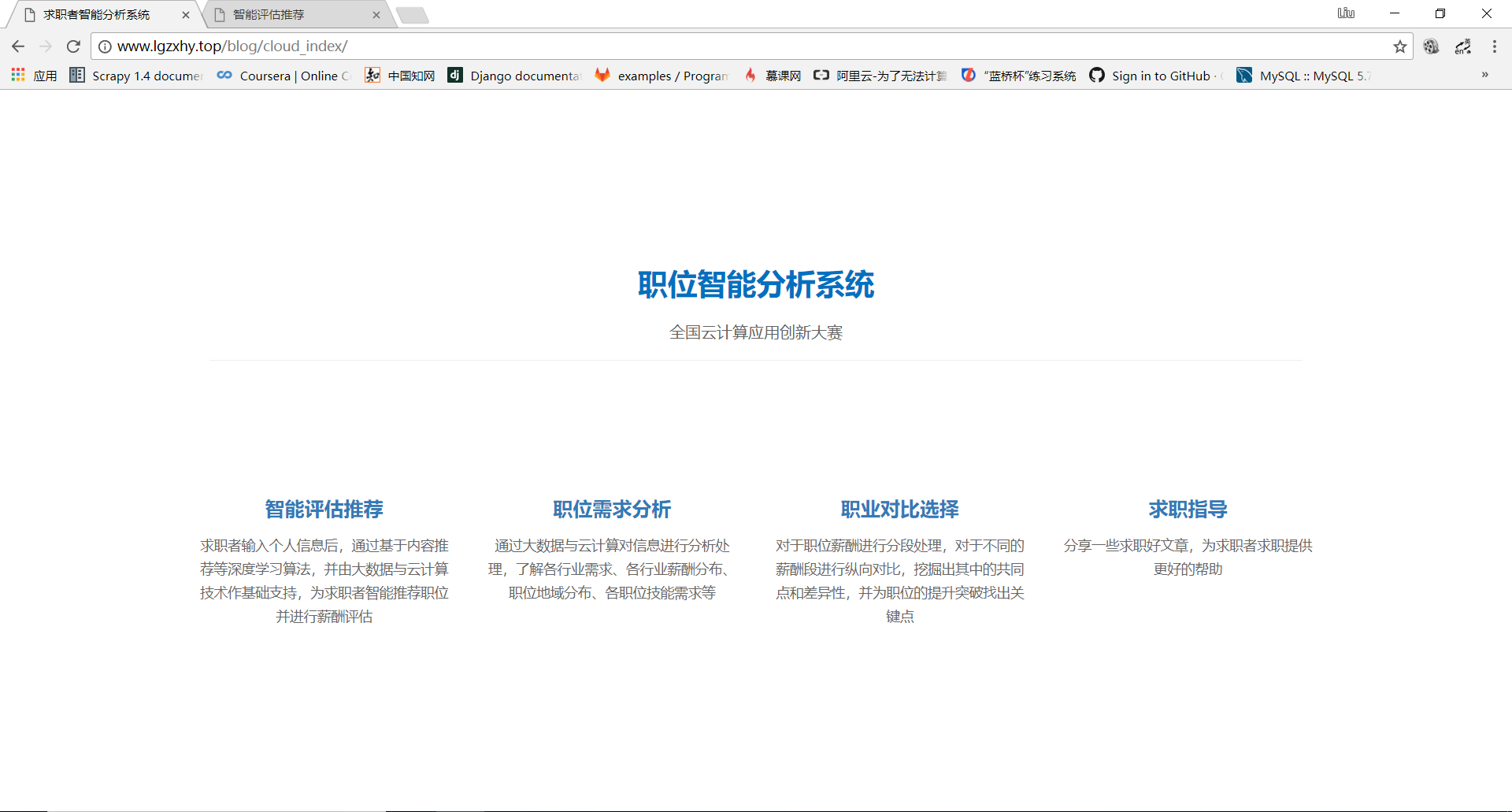
实现与用户进行交互的推荐系统页面，具有良好的网页交互体验。

工具与框架：

网页创建：Python开源网页框架Django

源码：../程序代码及说明/应用开发部署/\*

实现页面：











网址：<http://www.lgzxhy.top/blog/cloud_index/>