CSDN 搜索引擎

需求规格说明书

作者：吴宇涛 李德宏

2021 年 11 月 3 日

華南師範大學
SOUTH CHINA NORMAL UNIVERSITY

1 引言

1.1 编写的目的

搜索引擎是指根据一定的策略、运用特定的计算机程序从互联网上采集信息，在对信息进行组织和处理后，为用户提供检索服务，将检索的相关信息展示给用户的系统。

网络中用于各种功能的搜索引擎非常多，对于计算机专业的同学，寻找博客自学，代码题解等方面，国内的博客园、CSDN比较优秀。但数据都过于集中，缺少重点。且界面不是很纯净，会有一些广告。对于计算机专业的大学生，要花费大量的时间寻找质量不一的博文，无疑增加了学生们的工作量。这次软件工程实验项目，我们想要实现的就是整合CSDN网站的有益信息，搭建起一个更为方便、便利的搜索引擎，让同学们在学习的时候能够获得更多的帮助。

1.2 背景

搜索引擎技术是互联网上面用的最普遍之一的技术。主要是以用户为中心。当客户输入查询的请求时候，同一个查询的请求关键词在用户的背后可能是不同查询要求。例如用户输入的是“苹果”，那么作为一个想要购买iPhone的用户和一个果农来说，那么要求就是大大的不一样。甚至是同一个用户，所查询的关键词一样，也会因为所在的时间和所在的场合不同而返回的结果不同的所有主流搜索引擎，都在致力于解决同一个问题：怎样才能从用户所输入的一个简短的关键词来判断用户的真正查询请求。当代搜索引擎主要是以用户为中心。

1.3 定义

- **Django：**Django是一个由Python写成的Web应用框架。Django的主要目的是简便，快速的开发数据库驱动的网站。它强调代码复用，多个组件可以很方便的以“插件”形式服务于整个框架，Django有许多功能强大的第三方插件。

- **HTML：**HTML是超文本标记语言（Hyper Text Markup Language),HTML不是一种编程语言，而是一种标记语言。

- **Web crawler:** 网络爬虫（又称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。

- **MySQL:** MySQL 是最流行的关系型数据库管理系统，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)应用软件之一。

1.4 参考资料

[1] https://www.runoob.com/mysql/mysql-tutorial.html

[2] https://www.runoob.com/django/django-tutorial.html

[3] Python 测试驱动开发：使用 Django、Selenium 和 JavaScript 进行 Web 编程.第 2 版

[4] Python 网络爬虫权威指南（第 2 版） by 米切尔

2 任务概述

2.1 产品描述

这个项目的功能是利用爬虫获取信息，获取的数据缓存到csdn搜索引擎。该搜索引擎可以根据浏览者输入的字符串，对搜索信息进行推荐，然后展示搜索信息。

本课程设计由2人组成一组，李德宏同学负责前台网页界面的设计与美工以及搜索出来的文字界面设计，网站后台的功能模块的实现，以及数据库的录入设计和数据的获取；吴宇涛同学负责网页之间各模块之间的调用，搜索功能的实现。

2.2 产品功能

功能	具体描述
模拟登录	使用ChromeDriver驱动，通过http的 post请求 方式来提交用户信息的(用户名和密码)用到的库有“selenium”和“requests”。通过selenium进行模拟登陆，然后将Cookies传入requests，最终用requests进行网站的抓取。
爬取数据	通过http将web服务器上协议站点的网页代码提取出来，根据一定的正则表达式和xpath提取器取出所需的数据。
搜索功能	用户想要查询某些文章的时候，搜索框输入文章关键字，搜索引擎返回含有相关关键字或标签的文章列表。
分类排序	文章有三种排序：按相关度排序、按时间排序、按阅读量排序。用户点击不同的排序方式可以把文章按想要的方式排序显示
热门搜索	为用户显示搜索次数最高的五个关键词，点击热门搜索的关键词，搜索引擎按点击的关键词查询返回文章信息，输出含有相关关键词的文章信息

2.3 用户特点

此搜索引擎主要面对有一定计算机专业知识，经常使用 CSDN 平台交流学习的用户使用，主要为他们提供 CSDN 文章搜索，软件的界面简洁，操作简单，无任何广告，可以提高效率。 维护人员有爬虫和数据库的经验，可以定期对软件功能进行检查和更新。编写程序时常常会出现问题，搜索网上的文章是一种很好的学习方法，因此本系统使用次数预计比较高。

3 需求规定

3.1 功能需求

3.1.1 模拟登录

I. 概述

登录 CSDN 账号，到主页爬取主页所有文章链接

II. 输入

无

III. 流程

程序自动打开浏览器输入账号密码登录账号，爬取主页的文章链接

IV. 输出

文章链接保存到数据库等待下一步爬取

3.1.2 爬虫功能

I. 概述

取出数据库文章链接，每一个链接爬取获取文章内容

II. 输入

文章链接

III. 流程

以文章链接为爬虫url发送请求获取文章源码，进行解析，保存到数据库

IV. 输出

文章解析后的信息，保存到数据库

3.1.3 搜索功能

I. 概述

当用户想要查询某些文章的时候，搜索框输入文章关键字，搜索引擎返回含有相关关键字或标签的文章列表。

II. 输入

文章的关键字、标签

III. 流程

点击搜索按钮，服务器查询并返回对应文章的信息到网页进行显示。

IV. 输出

相应文章的链接，标题，阅读量等。
[用列表的方式，逐项定量和定性地叙述对系统所提出的功能要求，说明输入什么量、经怎么样的处理、得到什么输出，说明系统的容量,包括系统应支持的终端数和应支持的并行操作的用户数等指标。]

3.1.4 排序功能

I. 概述

文章有三种排序：按相关度排序、按时间排序、按阅读量排序。用户点击不同的排序方式可以把文章按想要的方式排序显示

II. 输入

点击排序按钮

III. 流程

将搜索结果进行相应的排序，再显示出来

IV. 输出

按相应方式排序好的文章信息

3.1.5 热门搜索

I. 概述

为用户显示搜索次数最高的五个关键词

II. 输入

点击热门搜索的关键词

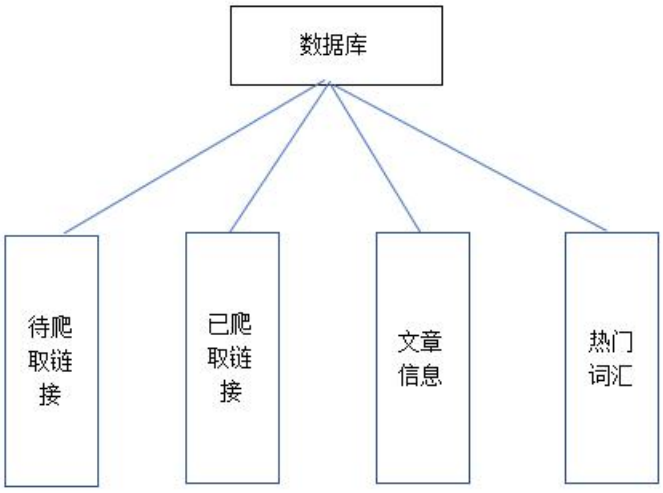
III. 流程

搜索引擎按点击的关键词查询返回文章信息

IV. 输出

含有相关关键词的文章信息

3.2 数据库结构



3.2.1 表信息

crawler_csdnblog：记录文章信息

search_engine_query：记录热门词汇信息
url_queue：待爬取文章链接
visited：已爬取文章信息

现在已爬取文章有 1059 篇，一篇文章信息占一条记录，一条文章链接占一条记录，由于爬取太频繁数据量太大可能会对服务器造成压力，有被限制请求的风险，不考虑使用代理，因此爬取文章的数量不会很多，文章数量大概在 2000 篇，管理数据不会很困难。

4 运行环境规定

4.1 设备：PC笔记本

4.2 支持软件：

操作系统：Microsoft Windows10

IDE：VsCode

数据库：Mysql，Navicat

编程语言：Python 3.7.8，Django

4.3 接口

（1） TCP/IP 通信协议接口
（2） GSM/CDMA 无线通信协议接口
（3） SMS 短消息通信协议接口
（4） 联通网关通信协议接口
（5） 防火墙通信接口
（6） 路由器通信接口
（7） 交换机通信接口

4.4 控制

在vscode调试模式下选择django环境，或者在terminal中输入manger.py runserver，运行整个项目.在本地浏览器输入127.0.0.1即可看到前端页面。