二手房网站大数据分析系统的设计

张标枫 张桂花
四川大学锦城学院计算机与软件学院 四川 成都 611731

【摘 要】随着社会的发展，房子对大家越来越重要，其中二手房是最活跃的，可以在短时间内满足人们的住房要求，不同的楼层，户型，小区，地理位置决定了房价的单价，而这些都随着社会的各种情况而改变，房价的走势依然是是大家关注的焦点。为了能够简单直接地看到房价的变化，本文运用 JavaEE 的 ssm 框架结合 hadoop 的 hdfs+hive+sqoop+mysql 搭建了一个大数据分析系统，实现了对二手房信息的管理，并利用 python 技术对各大二手房网站的房屋信息进行爬取，对各种影响房屋价格的因素做了可视化分析。

【关键词】大数据分析；管理系统；ssm 框架；python；hive；sqoop

本文首先阐述了系统的设计需求，定义了实现的功能，介绍了设计方案，分别阐述了使用到的相关技术，开发工具和环境，系统流程图，以及具体的设计实现。然后使用 JavaEE 的 ssm 框架实现了一个前后端交互的系统，并且利用 Python 对原始数据进行爬取，最后利用 java 随机生成 10 万条数据对系统进行测试。

1 设计需求

1.1 二手房网站的大数据分析系统设计与实现的现实意义

最近几年二手房市场变得格外火热，房价变动幅度较大，市场宣传眼花缭乱，人们往往无法提取有用的购房信息。通过对二手房网站数据的挖掘，分析，可视化的整合，能够为消费者提供购买房屋面积、单价、地段、总价等因素的横向和纵向比，解决了消费者找房难，难以掌握市场价格变动等问题。同时也能有效地为二手房中介机构提供极大的助力。通过二手房网站的大数据分析系统，他们能直观了解消费者购买房屋的需求以及哪些房屋比较受消费者青睐，掌握市场变化，及时采取措施。

1.2 设计需求

系统环境及工具要求：centos7,vmvare,Hadoop2.7.0,sqoop,mysql,hive,IntelliJ IDEA 2019,sublime。

系统功能：系统登录，python 爬虫，随机生成数据，数据分析，数据可视化，日志记录。

2 设计方案

2.1 相关技术

Linux 的操作命令；Idea 的 maven 搭建 ssm 框架；前端 + 后端开发；mysql 数据库操作；hive 操作；Sqoop 数据传输；java 随机数据；Python 数据挖掘；Hadoop 的分布式系统。

2.2 系统流程图

图 1 数据流程图

Fig.1 Data flow chart

3 系统的具体实现

3.1 Python 数据爬取

爬取某平台某市十个区共 100 页二手房的交易信息，获取房屋相关的信息，主要包括：title, positon, houseinfo, totalprice ,unitprice，将获取的信息保存到 Excle 表格中。

首先引入 python 的 xlwt，requests，BeautifulSoup 第三方库，然后，我们设置了一个 headers，在构建 request 时传入，在请求时，就加入了 headers 传送，服务器若识别了是浏览器发来的请求，就会得到响应。另外，我们还有对付”反盗链”的方式，对付防盗链，服务器会识别 headers 中的 referer 是不是它自己，如果不是，有的服务器不会响应，所以我们还可以在 headers 中加入 referer，再加上 HOST，服务器根据 Host 这一行中的值来确定本次请求的是哪个具体的网站。再定义

函数 get_info() 获取房屋的文本信息，定义几个列表装各种房屋信息，通过嵌套循环爬取 100 页的信息，构建 BeautifulSoup 实例，再调用 findAll 获取 div 标签的房屋信息，通过 for 的嵌套循环写入列表中，再将信息保存到表格中，设置初始化样式 ，字体，通过 worksheet.write() 向工作表单元格写入数据，最后调用 workbook.save 生成 excle 文件。

3.2 随机数据的生成

3.2.1 要确定我们整个项目要使用的数据的字段

City、Describe、Address、info、Buliddate、Layout、Area、Towards、Totalprice、Unitprice 分别代表了城市、描述、地址、楼层、修建日期、户型、面积、朝向、总价、单价。

3.2.2 编程思路

首先根据数据字段建立多个字符串数组，并且调用 split 方法以逗号进行分隔，将一个字符串分割为子字符串，然后将结果作为字符串数组返回。定义每个字段的随机生成的函数，调用 Math.random()。定义一个动态数组把所有随机生成的字段放入数组中。最后利用 for 循环并且调用 get 方法得到数据，控制随机生成的数据条数。

3.2.3 生成数据文件

先写一个获取当前操作系统桌面路径的代码。然后写一个生成文件的代码 File file = new File 生成一个文件，用for循环把数据写入进去。最后在桌面查看生成的文件。

3.3 数据维度

编写数据维度：楼层分布，户型分布，不同时间二手房数量分布，不同小区的平均总价，平均面积，平均单价分布。

3.4 数据处理传输与加载

首先把网络爬虫的数据并入随机生成数据的文件中进行数据的整合。再利用 SSH Secure File Transfer Client 工具把文件上传到虚拟机系统。最后进行数据的预处理：先删除文件第一行记录，即字段名称。获取数据集中的前一万条数据作为小数据集，使用 vim 编辑器建立一个脚本文件使用 awk 命令进行文件的截取。把截取的小数据集上传到HDFS/dataset/user_log。在Hive上创建数据库，创建对应的表，把 HDFS 中的数据加载到数据仓库 Hive 中。

3.5 数据分析

3.5.1 启动 Hadoop 和 Hive 以及 mysql

因为需要借助于 MySQL 保存 Hive 的元数据，所以请首先启动 MySQL 数据库。由于 Hive 是基于 Hadoop 的数据仓库，使用 HiveQL 语言撰写的查询语句。最终都会被 Hive 自动解析成 MapReduce 任务由 Hadoop 去具体执行。因此，需要启动 Hadoop，然后再启动 Hive。执行下面命令启动进入 Hive：cd /usr/local/hive，执行 ./bin/hive。

3.5.2 创建数据库和表

创建数据库 Create database house；再执行命令建立表格 inner_cd1, 关键加载数据 ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION 'dataset/user_log'; 再查询数据 select * from inner_cd1 limit 10; 看数据加载进来没有。

3.5.3 简单查询分析（根据维度）

1)select info,count(*) from cd group by info ;2)select layout,count(*) from cd group by layout;3)select builddate,count(*) from cd group by builddate;4)select address,avg(totalprice),avg(area),avg(unitprice) from cd group by address limit 10 .

3.6 存入数据库

登录 MySQL:mysql –u root –p；创建数据库 :create database house; 创建表，cd 数据字段类型为 varchar 类型；最后导入数据 : 使用 Sqoop 将数据从 Hive 导入 MySQL，用 sqoop 命令直接把数据全部加载。

3.7 网页设计

左边框：随机数据生成模块。上边框：a. 项目介绍模块 b. 设计思路 c. 数据来源 d. 日志记录 e. 成果展示。中间部分：主要的展示区域。

3.8 SSM 框架的搭建

3.8.1 新建 maven 项目

File－>New－>Project, 勾 选 Create from archetype，选择 webapp，指定 maven 的位置，指定项目名称和地址。

3.8.2 依赖配置好之后，开始整合

整合 spring 和 mybatis，新建 mybatis–config.xml 的文件，配置数据源连接信息。新建 applicationContext.xml 文件，配置 Druid 数据源，SqlSessionFactory 会话工厂，Mybatis 的 Mapper 接口扫描等信息 , 然后整合 Spring 和 SpringMVC, 新建 spring–mvc.xml 文件：配置事务。配置文件统一放在 Resources 目录下，方便管理配置视图解析器路径前缀 <property name="prefix" value="/WEB–INF/jsp/"/> 文件后缀 <property name="suffix" value=".jsp"/>

3.8.3 目录结构建立

在 java 文件下建立 Aspect、dao、bean、controller、mapper、service、util 等文件，在 webapp 下新建一个 jsp 文件，在 jsp 文件下建立 common、css、js 文件，在 resources 下建立 UserLogMapper.xml 写入各维度的查询语句。

3.8.4 配置 tomcat 服务器

配置名称: Tomcat 8.5.54; URL: http://localhost:8080/; HTTP port: 8080; Deployment; 添加 artifacts。

3.8.5 模块功能实现

1) 系统登录

设置账号 admin、密码 123, 先创建一个工具类 JsonMsg, 来验证是否登录成功, 定义 private int code;private String msg;private Map<String, Object> extendInfo = new HashMap<>(); 实现三个方法, 登录成功 success( ) 实现 setCode(100); 登录失败 fail ( ), 登录失败后的提示方法 extendInfo ( ) 。设计 jsp 登录界面关键代码, 写个 if, else 语句 if(result.code==100) 跳转主页, 否则报错。通过控制层与页面进行交互的关键代码: if(!admin123.equals(username+password)){return JsonMsg.fail().addInfo(login_error, 输入账号用户名与密码不匹配, 请重新输入!);}return JsonMsg.success()。

2) 随机数据生成模块

在 service 下建立 suijiS 编写相随机生成的代码。然后在 controller 下建立 suiji 的 java 文件, 使用注解 @Controller 在 tomcat 启动的时候, 把这个类作为一个控制器加载到 Spring 的 Bean 工厂并且对其实例化, @RequestMapping 这个注解, 添加路径的跳转 ,@Autowired 标注 suijiS 的实例, 让 spring 完成 bean 自动装配的工作, 最后再写一个方法调用随机生成数据的 jsp 页面。

3) 项目介绍模块

在 Controller下建立 js 的文件, 同样使用 @Controller 和 @RequestMapping 注解, 再写一个 jsp 页面跳转的方法, 实现页面的跳转功能。页面跳转到项目介绍的展示页面。

4) 设计思路模块

在 Controller下建立 sjsl 的文件, 同样使用 @Controller 和 @RequestMapping 注解, 再写一个 jsp 页面跳转的方法, 实现页面的跳转功能。设计思路跳转页面的展示为本项目流程图。

5) 数据来源

在 Controller下建立 sj 的文件, 同样使用 @Controller 和 @RequestMapping 注解, 再写一个 jsp 页面跳转的方法, 实现页面的跳转功能。页面直接跳转到 python 爬取数据的网站。

6) 日志记录

日志记录的编写用 Aop 切面编程利用注解的方式, 实现一个切面日志功能。先定义一个自定义注解类 @Aop。然后基于注解, 编写一个切面类, 实现对 ip, 操作时间, 访问的地址, 访问耗时, 访问的方法的获取 , 难点在于获取 ip, 我使用了一个工具类 ToolUtil , 工具类的关键是获得 ip 地址。然后, 在 controller 的类或方法的头部加一个注解: @Aop。在 mysql 建立日志表格 sys_log, 用于存放获得的日志信息: id int(20) 日志记录主键 ,visittime, datetime 访问时间 ,ip varchar(64) 访问的 IP,url varchar(300) 访问的地址 ,executiontime bigint(20) 访问耗时 ,method varchar(100) 访问的方法名称。编写一个服务层把获取到的信息插入数据库。最后在 Controller 下建立 rizhiController 的文件, 同样使用 @Controller 和 @RequestMapping 注解, 再写一个 jsp 页面跳转的方法, 实现点击日志记录页面就跳转到日志展示的页面。

7) 成果展示

同样在 Controller 下建立 chengguoController 的文件, 同样使用 @Controller 和 @RequestMapping 注解, 再写一个 jsp 页面跳转的方法, 实现点击成果展示页面跳转到 echarts 表格展示。

8) 各个模块的 jsp 页面设计

a. 公共部分

网页底部栏: 按照 idea 提供的 jsp 页面框架在 <body> 下写一个版权信息的标志。

导航框: 写项目介绍、设计思路、数据来源、日志记录、成果展示模块, 能够点击实现相应的页面跳转。

左边框: 随机生成数据的模块, 实现点击随机生成数据。

b. 各模块实现自身功能且每个模块引用公共的 jsp 页面

项目介绍的 jsp 页面, 直接在 <body> 写项目成员, 人数, 分工。设计思路的 jsp 页面, 添加一个流程图。数据来源的 jsp 页面, 在实现跳转功能的函数添加爬虫的网站地址。日志记录的页面, 加入一个表格, 查询数据库信息, 在表格展示出来。成果展示的 jsp 页面, 在此页面下建立几个子页面, 每个页面对应一个维度的分析展示, 在 echarts 官网下载 js 文件, 使用对应的表格代码。在 controller 下建立一个类, 初始化各自段的数据, 并且初始化几个 list, 把 UserLogMapper.xml 查询出的信息传到 list, 再在 echarts 表格代码中进行调用, 实现数据的可视化。

4 数据测试

首先启动虚拟机, 启动 idea 的 tomcat, 浏览器弹出系统登录页面, 登录系统。点击随机数据生成, 启动 python 爬虫, 把数据总合在一起。然后上传虚拟机, 再启动虚拟机的 hadoop, 把数据上传到 hdfs。启动 hive 创建表格, 加载数据, 进行分析, 最后启动 sqoop 加载数据到 mysql。点击网页的成果展示, 查看 echarts 表格的数据分析展示。完成数据测试。

【参考文献】

[1]Tom White. 华东师范大学数据科学与工程学院译 .Hadoop 权威指南（第 3 版）（修订版）. 北京：清华大学出版社 ,2015.
[2] 黑马程序员 . Java EE 企业级应用开发教程 [M]. 北京 : 人民邮电出版社 ,2017.
[3] 杨开振 , 周吉文 , 梁华辉 , 谭茂华 . Java EE 互联网轻量级框架整合开发 SSM 框架 [M]. 北京 : 电子工业出版社 .2017.
[4] 张良均 . Hadoop 大数据开发基础 [M]. 人民邮电出版社 ,2018.
[5] Edward Capriolo.Hive 编程指南 [M]. 北京 : 人民邮电出版社 ,2013.