基于二手房网站大数据分析系统的设计与实现

李治 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】随着时代的发展，互联网越来越成为了人们日常生活的不可分割的一部分，互联网越来越成为了人们日常生活中不可分割的一部分。而大数据的出现，更是让人们的日常生活更加便捷。二手房网站大数据分析系统是一个涉及数据处理、存储、查询、可视化分析等数据处理的完整流程项目。其使用了 Linux、MySQL、Hadoop、Hive、Sqoop、idel、ECharts 等系统和软件。使用 SSM 框架集开发。它的功能是可以通过前端界面，来实现随机生成出数据、实现历史记录的查看以及实现二手房数据的可视化等。

【关键词】大数据分析；可视化；随机生成

1 概述

随着大数据走进人们的视野，越来越多的行业会选择使用大数据来对事物进行分析。如今市面上有很多关于二手房交易的网站和信息，而面对铺天盖地的信息，二手房买家面对海量的信息难免会陷入不知如何选择的境地。为了解决这个问题，二手房网站大数据分析系统可以对大量的二手房数据进行分析。通过不同维度，不同的统计图类型，例如扇形图、折线图、柱状图等，通过这些统计图来实现数据的可视化。让人们能够清楚的、快速的看见现如今二手房的形式，提取出自己想要的信息，并以此作为辅助做出购房选择。

二手房大数据分析系统，通过生成、分析、查询这一套流程，使用户可以清楚的看见现如今二手房交易市场的形式。避免在购买过程中买到自己不如意的二手房，避免不必要的损失。这个项目可以让一个对二手房市场并不是很了解的人能够快速的获得有效信息，并以此做出决定。

2 网站功能介绍

首先用户需要通过用户名和密码进入网站。然后可以通过随机生成页面，点击随机生成按钮，将会随机生成一定数量的二手房数据并传入 HDFS。这个数量可以由用户自由选择。然后通过 Hive，对数据进行查询和分析，主要对比 4 个维度：楼层数量对比、户型数量对比、不同年建房屋数量对比以及不同城市平均总价、平均单价、平均面积对比。然后在成果展示界面，通过 ECharts 绘制扇形图、折线图、柱状图等，通过这些统计图，可以将二手房各个维度的数据清楚的展示给用户，实现数据可视化。在历史记录页面，将记录每个用户所进行过的操作。用户可以通过历史记录页面查看到自己之前所进行过的一些操作。同时网站拥有一个数据来源选项，点击这个选项可以到达数据的爬取页面。网站还有一个设计思路选项，通过点击将在网页上通过图片展示整个项目的设计思路以及运行的过程。

3 开发环境

这个项目几乎全部在 Linux 下进行，以 Hadoop（分布式系统的基础架构）作为基础。在这个项目中，需要将数据存入到数据库 MySQL，同时需要 MySQL 为 Hive 提供元数据存储服务，也需要 MySQL 为前端 ECharts 的数据可视化图片提供数据支持。ECharts 名称源于 Enterprise Charts 的缩写，含义是商业级数据图表[1]。本项目还需要用到 Sqoop。Sqoop 支持在 Hadoop 和其他数据库之间进行数据互导操作，在这个项目中我们需要使用 Sqoop 将 Hive 中的数据传入 MySQL，前端开发选择了 SSM 框架。SSM 实现与数据库的动态交互，SSM 架构使用 MyBatis 持久层框架，该框架专注与 SQL 本身，是一种灵活的 dao 层优化方案，适用于性能要求高、需求多变的项目[2]，非常适合本项目的开发。前端框架选择了 Bootstrap 来搭建和编写 JSP 页面，BootStrap 包含 HTML、CSS 和 JavaScript 开发工具集，在开发过程中使用这些工具集不仅可以使 Web 页面自适应移动设备，还可以提高代码的复用率，提高开发效率[3]。项目管理使用 MAVEN。开发工具选择 IDEA。

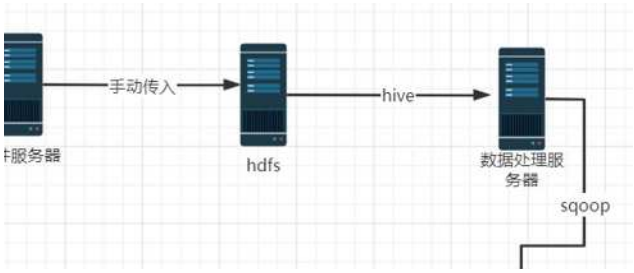

图 1 数据流程图

Fig.1 Data flow chart

4 网站功能的实现

4.1 SSM 框架搭建

要实现网站的各个功能，首先是数据库建表。其中表中的维度含有：房屋名称、地址、高、中、低楼层、户型、朝向、单价、总价。以此开对应项目中的实体类。然后要完成 SSM 框架的搭建。

当今流行的 SSM 框架中 Spring MVC 对应网络层，mybatis 对应持久层，Spring 统筹全局，业务逻辑层也交给 Spring 框架处理 [4]。先修改 pom.xml 来导入本次项目中可能会用到的 jar 包，然后编写 Web 项目的配置文件、Spring 容器配置文件、SpringMVC 配置文件、 MyBasits 配置文件。做好这些之后，SSM 框架就算是正式搭建完成。

然后是前端页面的搭建，在前端的多个页面搭建过程中，我们可以分别将其中公共部分的代码提取出来，然后将公共部分的 HTML 代码用 JSP 实现，这样当我们编写其他页面的时候就可以将公共部分的代码进行引用。

4.2 随机生成页面

关于随机生成页面的实现，首先将随机生成数据的代码放入 service 层。然后通过 controller 层用 autobean 自动装载类，然后在 controller 里面生成一个对象，再用对象调用 service 层中用于随机生成的方法，然后在 jsp 文件中写一个按钮，通过 ajax 点击一个按钮来调用随机生成数据的方法。从而实现在随机生成页面点击按键进行随机文件的生成。 还有另一个办法，先进行实体类的实现，先写一个 bean 的 class 文件，里面写上随机生成所需要的属性的 get 和 set 方法，然后编写 mapper 文件以及 Mapper.xml 实现使用 mybatis 自动代理实现持久层。然后编写生成属性的 class，设定各种属性的初始值以及生成规则，在 service 层设定按类型注入属性并获取属性。再通过 controller 层的实现控制。最后在 JSP 文件中编写页面，设定下拉框、按钮等。这样也能实现数据的随机生成。

在早期制作这个项目的时候，我们选择直接从二手房交易网站中使用 python 爬取二手房相关数据，但这样做会出现 2 个问题。第一，python 爬取速度有限，如果我们需要极大量数据的时候，随机生成的速度将远高于爬取信息的速度。第二，直接爬取下来的数据很难直接通过 Hive 进行分析。爬取下来的数据信息，有一小部分是错位、空值等异常值，必须先通过数据清洗才能使用。某些维度中也包含有中文，这样也不便于 Hive 进行分析。由于随机生成的数据需要放进 Hive 里去分析，所以我们将生成文件的维度中，能去掉的中文去掉，能转换为数字的转换为数字。例如，我们将建造年份中的年建去掉。如 2010 年建改为 2010。同时我们将房屋朝向维度中的东、南、西、北，分别用 1、2、3、4 替代。户型也可以通过数字替代，例如两室一厅可以写作 2.1。这样可以方便 Hive 进行查询和分析。在随机生成的过程中，大部分维度都可以通过生成随机数的方式进行生成。但是有些维度全是中文，例如地址，这些维度就不能通过随机数生成进行生成。但是我们可以先通过设定大量的街道名称、小区名称、门牌号等，然后在随机生成的过程中将不同的街道名称、小区名称、门牌号进行自由组合，以此来实现地址的随机生成。然后将随机生成的 HDFS 上的数据传输给 Hive。

```
public void wd() {

    List list=new ArrayList();
    for (int i=0;i<100000;i++){
        //          System.out.println(getList());
        list.add(getList());
    }
    System.out.println(list.size());
    System.out.println(list);
    //写入txt
    FileSystemView fsv = FileSystemView.getFileSystemView();
    File com = fsv.getHomeDirectory();
    String deskPath = com.getPath();
    System.out.println( deskPath );
    File file = new File( pathname: deskPath + "\\" + "aa.txt" );
    BufferedWriter bw = null;
    try {
        bw = new BufferedWriter( new FileWriter(file) );
        for(int i = 0; i < list.size(); i++ ) {
            bw.write( list.get(i).toString() );
            bw.newLine();
        }
        bw.close();
    } catch (IOException e) {
        e.printStackTrace();
    }
}
```

图 2 随机生成代码图

Fig.2 Randomly generated code diagram

4.3 Hive 处理

Hive 可以通过类似 SQL 语句的 HiveQL 语句对数据进行四个维度的查询和分析。最后需要将查询的结果存

入一张新的表传输给 MySQL。然后通过 Sqoop 将数据从 Hive 导入到 MySQL。首先我们要在 MySQL 中建立数据库和建表，这里要注意数据库的编码，这里我们可以通过在 MySQL 中输入代码 show variables like “char%”；来查看当前数据库的编码。我们一定要保证我们当前的编码为 utf8，否则将无法正常的导入中文。然后我们就可以通过代码将数据从 Hive 复制到 MySQL 当中来。用于最后通过 ECharts 读取 MySQL 中的数据，然后展示到前端页面的统计图里。以此来实现数据的可视化。

4.4 ECharts 可视化

ECharts 可以提供直观，生动，可交互，可高度个性化定制的数据可视化图表。ECharts 支持很多的图表，如折线图、柱状图、散点图等。完美契合了本项目的需求。前端页面想要获取服务端的数据，首先需要导入相关的包，前端 JSP 页面使用 ECharts 来展现可视化。在 JSP 文件中，需要指定图表的配置项和数据，每个 JSP 页面都需要导入相关 ECharts.js 文件。在每个 JSP 的底部还需要编写可视化逻辑代码。这样就能在前端看见我们所编写的可视化图形。

图 3 可视化图

Fig.3 visualization diagram

5 结束语

对于生活在这个数据大爆炸的时代的我们来说，通过使用大数据来丰富或方便我们的生活是很有必要的。随着网络上有关二手房的信息越来越多，二手房买家可能会因为面对海量的信息而发愁，不知道该怎么快速、有效的提取自己所需要的、有用的信息。而本项目拥有完整的数据采集、数据处理、数据展示的过程，将会为这一类的二手房买家提供极大的帮助。

这个对于一个二手房分析系统来说，首先界面要简洁大方，功能要完善，这样才有利于用户的使用。同时还需要有完整的实现过程。本项目的界面十分清晰，可以让初次使用的用户也能快速的找到自己所需要的功能。二手房分析系统，有利于用户从繁杂的数据当中提取出用户自己最想要的部分。以此来辅助用户更好的做出自己的购房选择。本项目还有很多地方可以继续完善，后期可以增加查询和分析的维度来丰富整个项目。也可以增加随机生成的增、删、改等功能使用户能够更加灵活的控制随机生成的数据。本项目还有一个缺陷，由于没有购买域名，本项目只能在本地运行。后期随着项目的完善可以通过购买域名的方式让更多的用户能使用到本项目。

【参考文献】

[1] 赵海国 .Ajax 技术支持下的 ECharts 动态数据实时刷新技术的实现 [J]. 电子技术 ,2018,47(03):25-27+57.
[2] 蔡明月 , 甄勇 , 苏林晗 . 基于 SSM 框架的组工管理互联信息平台的设计与实现 [J]. 铁道运营技术 ,2020,26(03):49-51.
[3] 王燕贞 , 沈毅波 . 基于 SSM 框架的高校学生综合测评系统设计与实现 [J]. 通化师范学院学报 ,2020,41(04):58-63.
[4] 赵志成 . 基于 J2EE 协同办公管理系统的设计与实现 [J]. 哈尔滨师范大学自然科学学报 ,2015,31(01):85-87.