基于大数据的链家网上海二手房市场价格分析

李旺家 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】 随着网上看房的兴起和火爆，链家网作为全国最大的房源信息网站之一，链家网的信息在一定程度上反映了中国各大城市的普遍情况；目前普通人的买房压力过大，相较于新房，二手房比较实惠。但一般很难了解到当地的二手房市场价格的情况，而不能买到经济实惠的房子。本文以上海二手房市场为例，在获取了相关房源信息后，进行数据的处理后，通过可视化的方式展现出二手房市场价格分析。

【关键词】SparkSql,Hive,Python，爬虫，链家网

1 绪论

研究背景

如今的时代已经是信息大数据时代；数据成为一种资产[1]；大数据时代的到来使得数据成为新的战略资源，但其意义不止于此[2]。生活在大数据的信息时代，我们每天都面对着大量的数据。成千上万的数据在我们面前，我们享受着大数据带来的信息便利的同时，也面临着难以从众多的数据中查找出我们真的需要的数据。因为大多数人都想要买房，二手房相较于新房较为经济实惠，因此房价是大多数人时时刻刻都在关心的数据，链家网作为全国最大的房源信息网站之一，上面有着各个地区上万套房源的信息，正常浏览完这些房源信息，不仅耗时耗力，而且还不一定能得出一个正确的大致的信息走向。

1.1 研究意义

链家网中的上海地区二手房房源信息具有数据量较多、数据的结构多，数据类型多、等符合大数据的特性。获取了房源信息之后，再进行数据上的处理、查询、分析以及可视化后，就可以清楚得了解到此地区的二手房的价格所占比。为想要了解上海地区的二手房市场价格的人们提供了一个方便快捷的方式，同时也可以举一反三反映出我们一线城市代表的二手房市场的整体价格情况。

2 整体项目流程介绍

本文的项目是对链家网中的上海地区二手房市场的房价进行大数据处理和分析，所使用的框架软件为：Python+Hive（基于 Hadoop 的数据仓库）+ SparkSql（Spark 架构上的工具）使用 Python 进行数据的爬取，通过谷歌浏览器登录了链家网之后，首先会选择地区，选择上海地区，然后点击二手房市场的跳转页面；首先分析网页页面的链接信息，以便可以循环爬取每一页的数据；分析 url 后编写代码就可以爬取链家网上海地区的二手房每一页的市场房源信息。获取数据后，根据数据的元素，首先使用 Hive 建一个 Hive 表，将数据从本地存储在 hdfs 上，再从 hdfs 将数据导入已经创建好了的 Hive 表中。最后使用 SparkSql 语言从大量的数据中查找我们想要的数据，对查找出来的数据进行分析和统计的操作，最后对处理好的数据进行可视化处理，这样就可以更直观得展示数据。

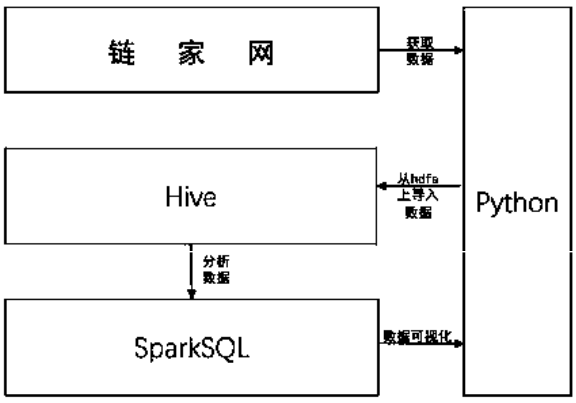

图 1 数据流程图

2.1 项目实现

数据的获取使用的软件是 Python。在获取房源信息之前，首先应该打开网站的网页，进行翻页的点击测试，点击的目的是为了观察网页的变化规律，观察网页的 url 是如何编写的。通过编写伪装 headers 头，构造 url，编写数据爬取代码，进行数据的爬取工作以及将数据存储到本地。

3 数据获取

3.1 使用的工具与环境

开发环境是 win10 与 Python3.8。使用的是 Python 库，有用于向服务器方发送请求的第三方库 requests 库，以及用于解析带着网页源代码的内容返回来的响应的 pyquery 库，我们分析网页源代码的重要工具是谷歌浏览器 Chrome 中开发者工具。

3.2 分析网页

在谷歌浏览器中登入到链家网之后，先选择上海地区，然后选择二手房的房源信息。此时观察网站的网页 url，发现它的 url 如下：https://sh.lianjia.com/ershoufang/pg1/；很明显 pg1 代表的意思是房源信息的页数，利用这一点，可以在代码的实现是使用 for 循环去让代码自动爬取链家网上海地区的二手市场所有的房源信息；使用谷歌浏览器 Chrome 的开发者工具可以进一步分析出网页的 HTML 结构，知道网页请求的各种类型的参数信息以便于代码中编写请求头的信息。

3.3 构造请求头

在请求网页的数据之前，我们必须先设置好我们的 headers（请求头）。这是因为如果没有 headers（请求头），网页会禁止我们的访问行为。请求头 headers 是一种伪装，通过设置请求头 headers 的一些信息（例如 User-Agent）将我们的请求伪装成浏览器的请求去获取网站的信息。

```
headers={ ‘User-Agent’ :Mozilla/5.0(Windows NT 10;Win64;x64)AppleWebKit/537.
36(KHTML,like Gecko)Chrome/78.0.3904.108/Safari/537.36}
```

3.4 构造 url

url，是资源标志符。它表示的是网页上的各种可用资源（例如图像、文本和视频）都会有一个独特的“标识”。因为本文是通过 Python 爬虫去获取链家网上海地区二手房市场的房源信息，所以就要通过每个房源的 url 就定位到每一个房源。通过观察链家网上海地区二手房页面的 url:

https://sh.lianjia.com/ershoufang/pg1/ 明显观察到了 pg1 是控制网页翻页的，通过简单的实验将 pg1 改为 pg2。果然页面跳转到了第二页。知道了链家网的 url 是如何实现翻页的之后，就可以构造 url 了。将 url 分为两部分实现拼接：

```
url=url1+str(i);
url1= https://sh.lianjia.com/ershoufang/pg
```

传参函数：def get_outer_list(maxNum)。这样就实现了利用 for 循环函数中的爬虫代码不断遍历爬取链家网上海地区二手房网页每一页的房源信息了。

3.5 使用 PyQuery 解析网页

PyQuery 库是 python 中的一个类 jQuery 库。它的作用是在请求返回的信息中找寻到我们真正需要的信息；PyQuery 的使用方法与 jQuery 在很大程度上都相同。但是 PyQuery 与 Python 自带的 urllib 库相比，它使用起来会更加方便简洁。

在寻找真正所需要的数据前，初始化一个 PyQuery 对象是必要的。使用 PyQuery 对于发出请求后接收到的 HTML 文本进行初始化，然后利用得到的 HTML 文本完成这个初始化。这个过程它本质上的原理就是用网页的源代码以字符串的形式传递给 PyQuery 来进行初始化，这样以便进行后续的深入挖掘我们所需要的文本信息操作：

```
response=requests.get( ‘https://sh.lianjia.com/ershoufang/’ + str(i) + ‘.html’, headers=headers)
doc = PyQuery(response.text)
```

通过 for 循环再进行基本的属性解析：

```
base_li_item = doc( ‘.base .content ul li’ ).remove( ‘.label’ ).items()
base_li_list = []
for item in base_li_item:
    base_li_list.append(item.text())
```

通过每个房源信息的观察，了解到每个房源数据的数据结构是一样的，因此构造出一个字典，以便临时存入解析出来的数据；字典的每个字段依次是：‘id’，‘Construction area of’，‘Set within the area’，‘Building head’，‘Decorate a situation’，‘Equipped with an elevator’，‘floor’，‘Door area’，‘Building type’，‘Building structure’，‘Ladder household proportion’。

对于已经解析出来的数据再进行一层循环遍历解析，得到我们想要的房源信息，并放入一个 data 字典中：

```
transaction_li_item = doc( ‘.transaction .content ul li’ ).items()
transaction_li_list = []
for item in transaction_li_item:
transaction_li_list.append(item.children().not_( ‘.label’ ).text())
```

4 数据存储

4.1Hive 与 Hadoop

由于本文所获取的房源信息数量庞大，组成简单。因此使用 Hadoop 模块中的分布式文件系统 HDFS 存储，再将数据从 HDFS 导入 Hive 表中以方便我们使用 SparkSql 处理数据。

Hive 的本质并不是一个数据库。它是运行在 Hadoop 上，建立在 Hadoop 上的一个数据仓库。它可以将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能 [3]。因为 Hive 这个数据仓库是和 Hadoop 关系紧密，它的底层计算原理仍然是 MapReduce。并且 Hive 之中的数据都存放在 Hadoop 中的 HDFS 上，可以通过 Hive 映射出来的表查看数据。

对于存放在 HDFS 中的大量的数据，Hive 可以进行加载，查询和提取等一系列操作。并且 Hive 操作较为简单，为我们查询提取提供了便利。因此通过 HDFS 和 Hive 和互相配合，就可以很好的解决数据的存储，查询，提取等工作。

4.2 存入数据

先在 Hive 使用命令创建一张结构与所获取下来的房源信息相同的 Hive 表：

```
create table data(Id int, unitprice int, totalprice int, area string, community string, doormodel string, floor string, construction double, structure string, elevator string) row format delimited fields terminated by '\t';
```

再使用 Hive 命令将原本存储在 HDFS 上的数据导入到已经创建好的 Hive 表中：

```
Load data local inpath '/root/usr/lib/hive/data/data.txt' into table data;
```

5 数据查询与分析

5.1SparkSql

虽然数据的映射表是在 Hive 中，但是本文并没有使用 Hive 的类 Sql 语句是进行数据的查找与分析，而是使用 SparkSql 来进行处理所获取到的数据的工作（提取、查询和分析）。

虽然 Hive 是数据仓库，同时也可以是查询数据的引擎。但 Hive 的查询速度无法与 SparkSql 的查询速度相比，所以 SparkSql 是完完全全可以取代 Hive 的查询这一部分业务。Spark SQL 是 Spark 用来操作结构化数据的组件，通过 Spark SQL，用户可以使用 SQL 或者 Apache Hive 的 HQL 语言来查询数据 [4]。根据两者的特点一般都是给两者进行分工。

5.2Hive: 负责廉价基础的数据仓库。

SparkSql：负责高速度高效率的数据查询计算。

使用 SparkSql 进行数据查询

SparkSql 之所以可以从 Hive 的数据映射表中处理数据，是因为 SparkSql 中的查询语句就是执行的 SQL 查询语句。之前的操作中，我们已经将获取下来的数据映射表创建在 Hive 上了，现在就是查询数据，进行分析了；根据我们的条件去查询数据。

```
spark.sql("select id from data where zongjia < 200).show
spark.sql("select id from data where zongjia >= 200 and zongjia <= 400).show
spark.sql("select id from data where zongjia >= 400 and zongjia <= 600).show
```

使用这样的查询语句就可以根据我们想要分析的数据依据条件来查询，在这里以链家网上海地区二手房价格区间来查询。

5.3 数据可视化

数据可视化就是使用图形图表等方式来呈现数据，图形图表能够高效、清晰、直观地表达数据包含的信息[5]。将查询分析出来的数据通过 Python 进行可视化处理后就会得到直观的图表以便于我们了解到更为明显的数据信息。下图为可视化后的一部分结果。

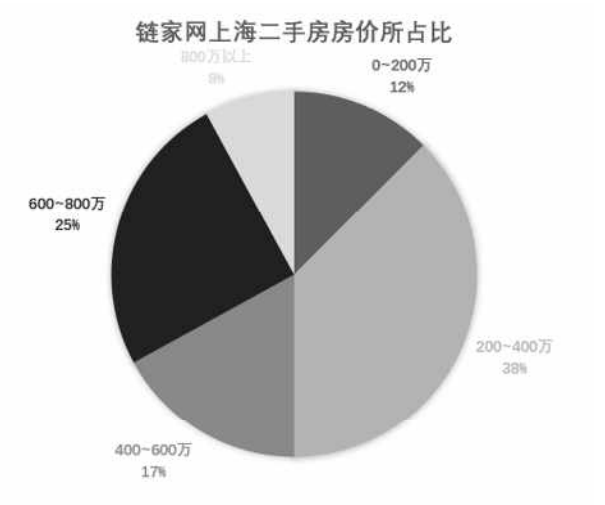

图 2 链家网上海二手房房价所占比图

结论

当下社会飞速发展，经济飞速增长，但物价也不断在上涨。一线城市上海的二手房房价也不例外，可以看到 200 万以下的房价，占比十分小才仅仅十分之一，而 200~800 万的房子占了绝大多数；举一反三，在这里可以想象其他一二线城市每个城市的房价是否也会像本文所分析出来的这样。房价与普通公民的生活息息相关，与幸福感有着千丝万缕的联系，中国在进入老龄化社会的今天，年轻人不敢结婚不敢生育是否也与房价有关系。这些经过查询、分析以及可视化后的数据都可以为我们提供一些关于一线城市的房价数据参考。

【参考文献】

[1] 冯勤群 . 大数据背景下数据库安全保障体系研究 [J]. 软件导刊 , 2013,12(01):156-158.

[2] 张晋 , 董亚君 , 王鹏宇 . 大数据对大学生网络舆情研究的意义 [J]. 佳木斯职业学院学报 , 2020,36(10):45-47.

[3] 张岩 , 王胤祥 , 胡林生 . Hive 大数据仓库构建与应用——以大陆在美上市股票数据为例 [J]. 数字通信世界 , 2021(04):186-187.

[4] 史媛 . 基于文本信息的 SparkSQL 处理研究 [J]. 电子技术与软件工程 , 2020(15):213-214.

[5] 刘艳玲 , 姚建盛 . Python 在数据可视化中的应用 [J]. 福建电脑 , 2020,36(03):68-70.