**不同房产交易平台上学区房数据爬取与转换**

1. 数据爬取与转换程序整体处理过程

处理流程主要包括三步(具体过程见图1):

1. 从网页上爬取原始房屋信息
2. 用原始信息中的地址信息获取百度坐标中的经纬度信息
3. 将百度坐标转换成谷歌坐标

对于不同的房产交易平台来说，第二步和第三步的过程是相同的，程序代码不需要修改，由于不同房产交易平台的网页构造不同，因此不同平台需要特定的爬虫程序。每种平台的爬虫程序将在对应的平台中介绍。

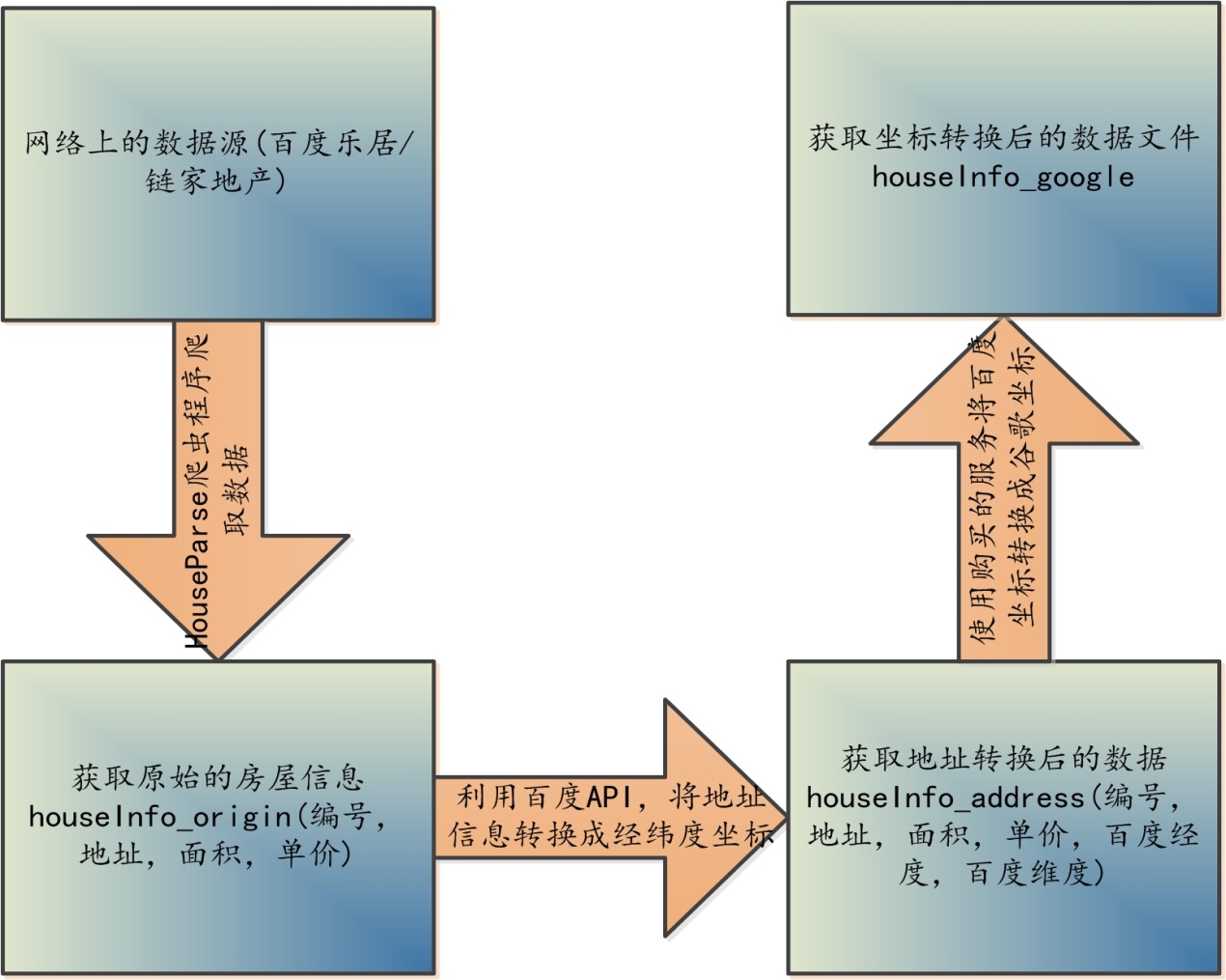
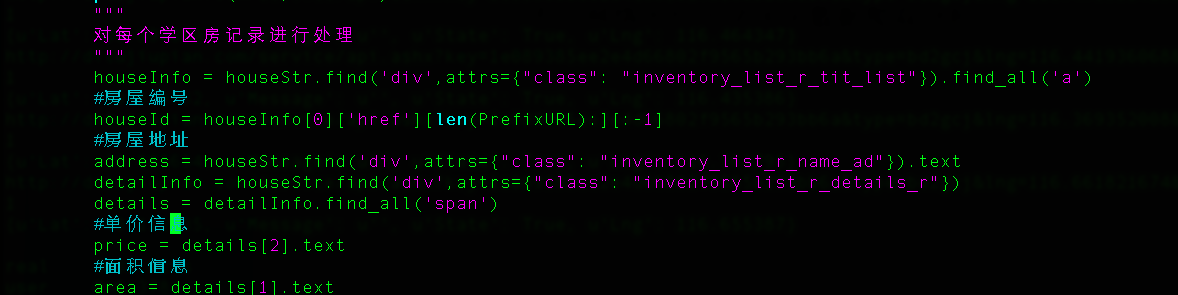


图1.不同房产交易平台上学区房数据爬取与转换过程

1. [百度乐居学区房](http://esf.baidu.com/bj/house/a100-m6-k5-n1)(总计获取1700条数据)
2. 具体页面



从图中可以看出，该页面中包含了具体的地址信息，因此我们只需通过程序解析出相应的地址，面积以及单价信息即可。程序处理源代码如图



1. 百度乐居存在的问题

百度乐居一定数量页面之后的页面都是相同的。如图:

正因为百度乐居有大量这样的重复页面，导致最后可以爬取的数据只有1700条左右。

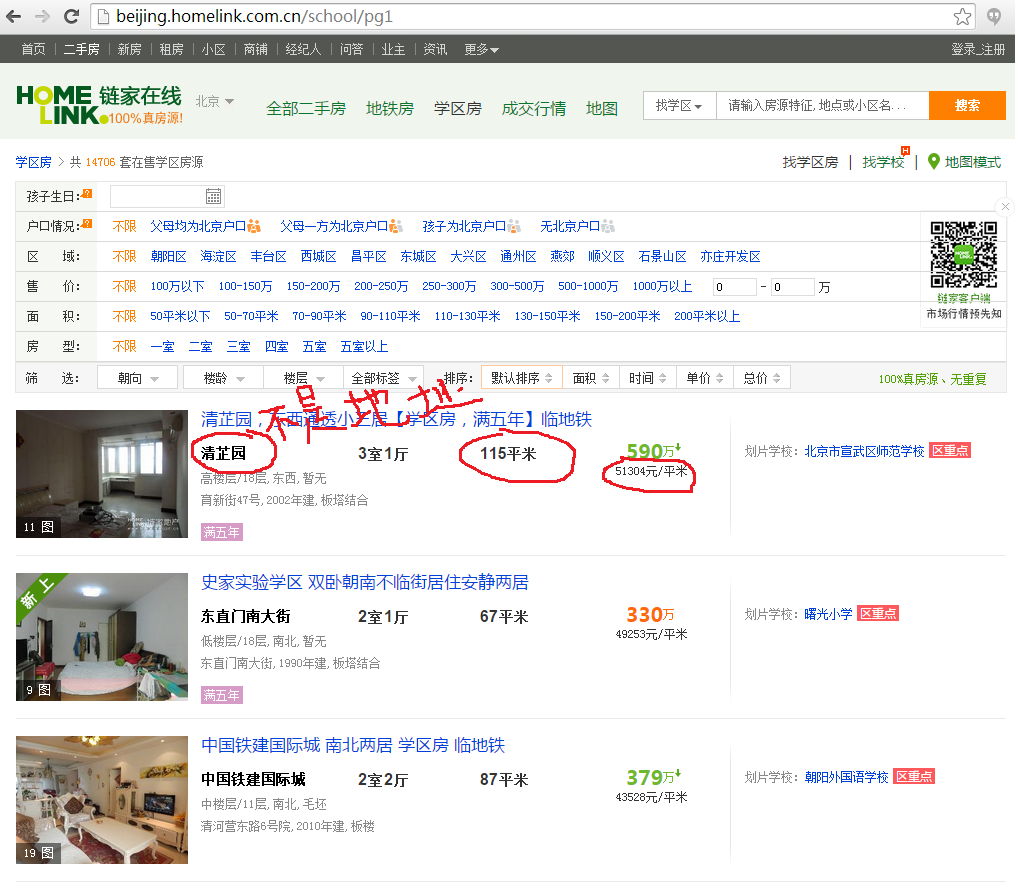
 

第1000页 第1002页

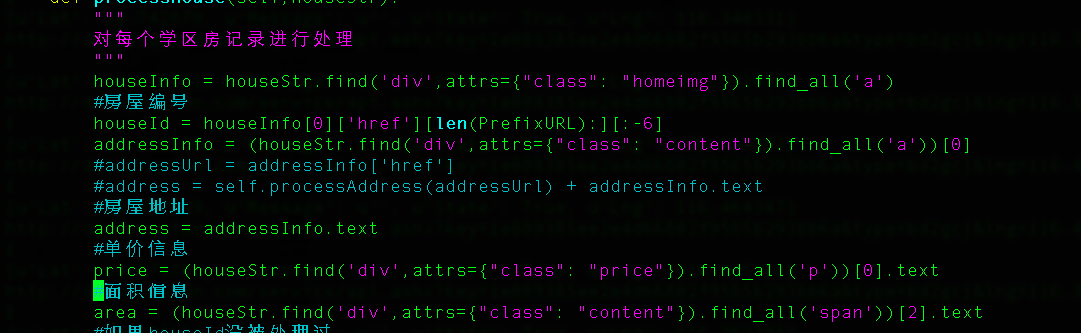


第4458页(最后一页)

1. [链家地产学区房](http://beijing.homelink.com.cn/school/pg1)(总计获取6789条数据)
2. 具体页面



从图中可以看出，链家中展示的页面中没有具体的地址，只包含小区名称，而且没有具体的地址位置。如果直接用小区名通过百度获取经纬度坐标(7814条数据（页面说14000条数据有重复数据）中有6788条能够获取到,成功率86.9%)。因此我们需通过程序解析出相应的小区，面积以及单价信息即可。程序处理源代码如图:



1. 链家地产存在的问题

因为只有小区没有地址不够准确，导致只有86.9%的房源能够获取到经纬度信息。为了解决这个问题，有一种比较耗时且相对麻烦的办法。

我们可以通过小区链接获取小区具体地址，如图:



但是这种方式会导致处理请求增加很多，增加了8000次请求，时间是原来基础上的12倍左右。原来大概需要3-4个小时跑完程序，这种方法需要将在一天左右，而且程序修改较麻烦。

1. 其他房产交易平台

如果要获取其他房产交易平台的数据，需要修改第一部分的爬虫数据，针对特定房产交易平台的网页结构具体分析即可。