

基于房产交易网站的数据获取与在线工具开发

王荇梓, 赖雯洁

(华东师范大学 地理科学学院, 上海 200241)

摘要:房屋交易网站提供了每个交易房产的详细信息,自动获取这些数据并进行在线分析可以帮助人们更好地分析一个地区房产情况,更有利于决策。开发网页分析工具是分析大数据发展的趋势,其具有更少的代码,同时拥有不亚于应用程序的功能实现数据采集与数据分析的实时对接,使得其成为工具开发的新宠。房产交易网站在线工具利用 Python 语言结合 Scrapy、ArcPy 等第三方模块开发,可自动提取房产数据,并针对不同数据类型,对某一地区的房产进行空间分布分析和规律监测等。以链家网、安居客两个房产交易网站为例,从中获取上海市的新房、二手房等房产数据,通过统计图表的形式显示上海市房产的空间分布情况,房价涨幅,各地区房产数量分布比例等,实现用户对大数据的进一步分析认识。

关键词:房屋交易网站;网络爬虫;地理编码;ArcPy

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2017)05-0154-06

doi:10.3969/j.issn.1673-629X.2017.05.032

Data Acquisition and Development of Online Analysis Tools Based on Real Estate Transaction Websites

WANG Jin-zi, LAI Wen-jie

(School of Geographic Sciences, East China Normal University, Shanghai 200241, China)

Abstract: Real estate transaction websites provide every detail about the real estate data. It would be helpful for people to know better about real estate, which is helpful for final decision. Development of online analysis tools meets the trend to analyze big data on real estate with less code to expect the integration of data mining with data analysis in real time owning the function no less than practical programs as new favorite in tool development. This analysis tool online has been developed with Python language as well as third party modules including Scrapy, ArcPy etc. to acquire information on real estate and to monitor and analyze spatial distributions and variations of real estate information from various types of data. Taking two trade websites of real estate, Network for Linking Family and Dwelling Guest, as examples, real estate data in Shanghai has been acquired to display spatial distributions and variations of housing prices as well as percents of real estates in diverse districts for convenience to analysis.

Key words: real estate transaction websites; website spiders; geocoding; ArcPy

0 引言

随着房产市场的快速发展和互联网技术的广泛应用,目前网上已有很多房屋交易网站,如链家网、安居客、yes515、爱屋吉屋、我爱我家等,这些网站为购房者和售房者提供了交易平台,用户可以从网站上查看每个交易房产的详细信息,如房产位置、房价、房屋的建造年代、楼层等。由于房屋交易网站覆盖面广,反映的信息实时性强,因此,对房屋交易网站上的房产信息进行分析能实时掌握一个地区的房地产市场情况。但网站上的信息是以 Web 页面的形式呈现,并不是直接可以

用于分析的数据,如以浏览网页的方式来进行分析,显然效率很低,很难实现对房产市场的实时变化监测。从网页中抓取原始数据,并处理成可以直接用于分析的数据这方面已有很多研究。较流行的抽取工具有 MDR^[1]、改进方法 Depta^[2]等,但其更希望目标网页是结构化的,因为抽取的信息主要是在列表或表格中。梅雪等^[3]基于网页模板的设计准则,提出了全自动生成网页信息抽取包装器 Wrapper 的方法—PSNT (extraction based on temPlate Structure aNd Tag tree),该方法同时实现了对网页中严格和松散的结构化信息的自

收稿日期:2016-06-10

修回日期:2016-09-15

网络出版时间:2017-03-07

基金项目:国家自然科学基金资助项目(41001270);上海市自然科学基金项目(14ZR1412200);闵行区中小企业技术创新计划项目(2014MH011)

作者简介:王荇梓(1995-),女,研究方向为数据挖掘、GIS 开发与应用;导师:吴健平,教授,博导,研究方向为地理信息系统开发。

网络出版地址:http://cnki.net/kcms/detail/61.1450.TP.20170307.0922.074.html

动化抽取,在相似网站中模板生成的匹配效果较好。例如主网站及其各个子网站,针对不同开发商的网站,还需要重新匹配模板。欧健文等^[4]使用多个网页对模板进行训练,以得到较为普适的模板,而后对归类爬取网页的主题信息,这对于搜索引擎十分实用。在地理信息提取方面,王曙等^[5]针对同一地理要素有不同描述的语言特点,建立地理语料库,使用搜索引擎与通用主题相结合的爬虫抓取网页。该方法没有事先训练样本,是先广泛获取相关网页,而后从筛选下来的网页中再次爬取内容,才可获取数据。这几种方法都是大面积爬取地理信息,并没有真正意义上利用 Web 中的原始数据,因此为了对地理数据进行统计分析,定点定抓的轻量级主题爬虫更为适合,功能全面且获取的是原始数据。

从网站抓取房产信息不仅是文本数据,还需要转换成 GIS 数据,以方便对房产信息进行空间分析。除

此之外,由于在大城市及特大城市中房产相应特征指标变化较其他城市迅速^[6-8],比起耗时长的精确研究,实时监测可以掌握房产变化最新动态。以链家网和安居客网站为例,研究基于房屋交易网站的房产数据获取与在线分析工具开发,并演示在上海市的应用。

1 主要房产交易网站介绍

1.1 链家网网站

链家网 (<http://www.lianjia.com/>) 是链家房地产经纪有限公司在 2009 年成立的房产交易线上平台,其主要业务领域为新房、二手房和租房。房产数据包括小区名称、地址、小区房屋均价、建造年代、楼栋总数、房屋总数、容积率、绿化率等信息。不仅如此,链家网中有百度地图提供的定位显示功能,可以直接得到小区的经纬度信息,如图 1 所示。

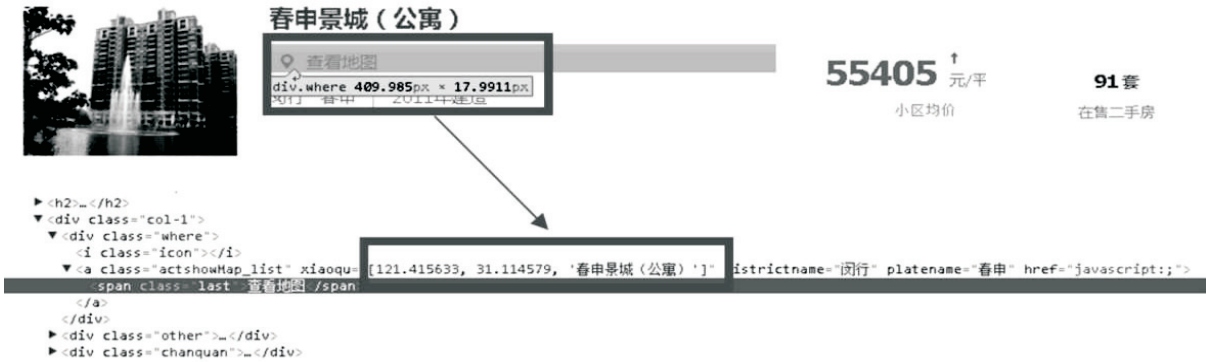


图 1 链家网显示信息

1.2 安居客网站

安居客 (<http://shanghai.anjuke.com/>) 是 2007 年成立的互联网房产交易平台,主要业务领域为新房、二手房、租房、商业写字楼四部分,2015 年进入 58 同城旗下。与链家网相比,安居客的房产没有地理坐标,需要经过地理编码得到地理坐标;在数据细节方面,安居客的数据较链家网更为全面。以春申景城为例,在链家网中搜索结果只有“春申景城(公寓)”1 条数据,如图 1 所示,但在安居客中搜索结果有 4 条,如图 2 所示,分为一、二、三期和 MID-TOWN,分类更为细致。另外,安居客房产信息中有物业类型、物业费用、总建面、停车位、出租率等,数据种类更为全面。

2 原始房产数据获取方法

从网站中获取数据主要包括三个过程:确定网页地址(URL),解析网页文件,存储数据规范格式。下面介绍利用 Scrapy 爬虫框架实现上述过程的方法。

2.1 Scrapy 简介

Scrapy 是 Python 的第三方软件包,是一个简单轻

量级的爬虫框架,操作简便,适合爬行简单网页数据。如果 HTML 格式复杂,含有 JSON,或需要用户身份验证等,可以考虑使用 Java 的分布式 Nutch 或稳定性更好的 Heritrix^[9-15]。Scrapy 规范了完整爬虫应有的核心:服务请求与返回、网页解析、数据存储。同时提供给用户足够的空间来完善爬虫,用户锁定目标网站后只需分析网页结构,即可快速编写爬虫。

使用 Scrapy 需要安装第三方扩展包,Python2.x 版本中需要安装:zope.interface, pypiwin32, pyOpenSSL, twisted, libxml2dom, lxml, Scrapy。Scrapy 框架的工作原理是:首先传入 URL,调度器(scheduler)将其传入下载器(downloader)对服务器发出访问请求,返回结果传入爬虫(spider)中进行解析。如果含有超链接,则传回调度器,否则传入解析器(ItemPipeline),利用 ScrapySelector 对 HTML 文件进行解析。

2.2 确定 URL

通常情况下,传入爬虫的是网站的主网页,即用户最先浏览的主页,而后根据不同的需求在主页相关的网页之间切换。因此确定 URL 的关键是网页的相互

切换,其可分为当前网页的切换和超链接跳转两种,而本质上两者都是通过对 URL 的改变来实现的。以安居客网页为例演示确定 URL 方法。



春申景城(一期)

【闵行·春申】莲花南路1111弄 兴梅路1199弄

竣工日期:2007年05月

57815元/平米

10.25%

二手房(1857)

查看地图

生活配套

价格行情

小区相册



春申景城(二期)

【闵行·春申】兴梅路1199弄

竣工日期:暂无数据

57705元/平米

10.77%

二手房(793)

查看地图

生活配套

价格行情

小区相册



春申景城三期

【闵行·春申】锦梅路1398弄

竣工日期:2011年12月

58392元/平米

10.56%

二手房(440)

查看地图

生活配套

价格行情

小区相册



春申景城MID-TOWN

【闵行·春申】锦梅路1398弄

竣工日期:2011年12月

56663元/平米

10.41%

二手房(106)

查看地图

生活配套

价格行情

小区相册

小区概况

春申景城三期

小区名	春申景城三期	总建面	900000平方米(大型小区)
所在区域	闵行·春申	总户数	442户
地址	锦梅路1398弄	建设年代	2011-12
开发商	上海连申房地产有限公司	容积率	1.81
物业公司	锦城怡安(上海)物业管理有限公司	出租率	约58%
物业类型	公寓	停车位	300
物业费用	3.15元/平方米·月	绿化率	36%(绿化率高)

春申景城MID-TOWN是春申景城三期锦梅路。春申景城是一座集国际公寓、万豪旗下高标准酒店、5A甲级写字楼、绿色LOFT、复合商业、精品公寓酒店、院常式住宅于一体的多业态城市综合体(City Economic Poly...

图 2 安居客网站房产信息网页

在 Scrapy 中免去了爬虫需要编写的请求返回命令,用户只需直接传入 URL 即可。观察安居客主网站为目录界面,为抓取全网数据需要机器模拟翻页。在 view-source 中查找“下一页”的源码,所在主标签为<div class=“multi-page”>,当在第一页时,“下一页”所在标签为<a href=“http://shanghai.anjuke.com/sale

```
▼<div class="wrapper">
  <!--begin: 筛选条件-->
  ▶<div class="filter-box noTab"></div>
  <!--end: 筛选条件-->
  <!--begin: 搜索结果-->
  ▼<div class="con-box">
    ▶<div class="list-head clear"></div>
    ▼<div class="list-wrap">
      ▼<ul id="house-1st" class="house-1st">
        ▶<li></li>
        ▶<li></li>
        ▶<li></li>
        ▶<li></li>
        ▶<li></li>
        ▶<li></li>
        ▶<li></li>
      </ul>
    </div>
  </div>
</div>
```

图 3 HTML 文档 DOMtree 结构图

Scrapy 提供了基于 XPath 和 CSS 的 Selector 方式对网页文件进行解析,根据系统自动选择最优的解析方法,可解析 HTML 和 XML 两种文件类型。在 XPath 中,“//”表示从文档节点开始抓取,“/”表示从上一级标签节点开始抓取,@ href 表示提取标签超链接。实现提取小

/p2/#filtersort”,class=“aNxt”>,标签中含有超链接,即需要抓取的 URL。而在最后一页时,“下一页”所在标签为<i class=“iNxt”>,标签中不含超链接。可以将其作为循环条件,遇到无链接的情况则跳出循环翻页。

实现模拟翻页代码如下:

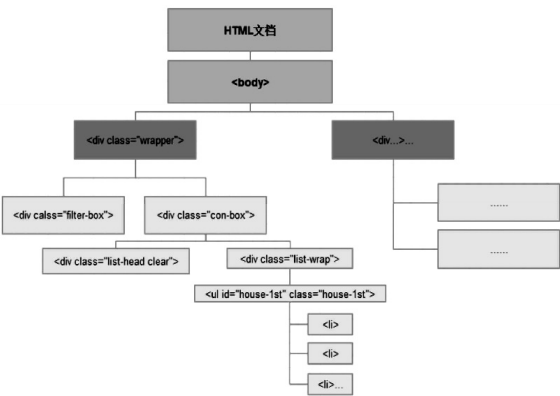
```
classConSpider(Spider):#创建 URL 池类 ConSpider
name="con" #爬虫名称为 con,必须唯一
allowed_domains=["shanghai.anjuke.com"]#域名
start_urls=[]

tpath="//div[@ class="page-content"]/div[@ class="multi-page"]/a[@ class="aNxt"]/@ href'

npath="//div[@ class="page-content"]/div[@ class="multi-page"]/i[@ class="iNxt"]'

while True:
    #rpage 为当前网页传入 selector 的解析,利用 XPath 寻找路径,
    nxtpage=rpage.xpath(npath)
    #限制循环条件,如果最后一页的“下一页”标签不存在,即为空,未到最后一页
    ifnxtpage==[]:
        turl=rpage.xpath(tpath)[0]
        #读取“下一页”标签中的超链接,@ href 读取属性,结果为只含有一个元素的 list
        start_url.append(turl) #将 URL 加入链接池
    else:
        break
```

在当前目录页一般含有的小区信息是不全面的,需要跳转到该小区网页进行抓取。首先对当前网页进行 DOMtree 分析。小区信息包含在中,嵌套于<ul id=“house-1st” class=“house-1st”>标签中,以此类推,DOMtree 结构如图 3 所示。



```
houses=selec.xpath('//ul[@ id="house-1st"]/li') #从当前网页中提取<li>
foreachhouse in houses:
    #从<h2>提取 href 的属性值
    xq=eachhouse.xpath('div[@ class="info-panel"]/h2/a/@ href')
    xq_href='http://sh.lianjia.com'+str(xq[0])
```

2.3 解析网页文件

在 Scrapy 中,解析函数命名为 parse,传入参数为服务器返回值 response。直接返回的 response 需要程序提取所需数据。首先需要生成一个 selector 实例 sel,对其可使用 XPath 等方法解析。以安居客在售二手房(<http://shanghai.anjuke.com/sale/>)信息为例,说明解析步骤。观察当前在售房屋信息,在 view-source 中寻找对应源码。面对嵌套标签,可利用类似于 list 的提取方式提取嵌套标签的同胞标签。

```
HTML 解析函数代码如下:

def parse(self,response): #返回的 response 传入 parse 解析
    sel=Selector(response) #声明选择器对象 sel
    datalist=sel.xpath('//div[@class="house-details"]/div[2]')
    #抓取属性为 class="house-details" 的 div 标签下嵌套的第二个 div 标签
    items=[]
    item=ConItem() #声明 Items
    item['address']=sel.xpath('//span[@class="comm-address"]/@title')
    for data in datalist:
        item['area']=data.xpath('.//span[1]')
        #抓取的面积数据存放在 area 中
        item['area']=item['area'].xpath('string(.)')
        #读取 span 标签下面的文本
        item['price']=data.xpath('.//span[3]')
        item['price']=item['price'].xpath('string(.)')
        item['btime']=data.xpath('.//span[5]')
        item['btime']=item['btime'].xpath('string(.)')
        items.append(item)
    return items
```

2.4 存储数据

从网页中抓取的文本数据一般可以选择存储为 CSV、JSON 等形式。为避开编码问题,将网页数据存储为 JSON 格式。打开 ajkspider 所在文件夹路径的命令提示符,直接输入:scrapy crawl ajkspider -o ajkdata.json -t json,即可在根目录中找到数据文件。Python 中自带 JSON 模块解析,将原始数据存储规范格式后存储为 Excel 文件。

3 房产数据的空间定位

在链家网中,网页中含有百度地图的经纬度坐标,可以直接获取;而在安居客网站中,可直接抽取地理信息为地址,因此需要进一步地理编码,从而得到地理坐标。根据上述两种地理信息类型来介绍房产数据的空间坐标获取方法。

3.1 坐标纠偏

一般移动设备上安装 GPS 芯片或北斗芯片,获取

WGS84 坐标系下的经纬度,谷歌地图应用 WGS84 坐标系。根据国家测绘地理信息局国土测绘司在 2006 年发布的文件《导航电子地图安全处理技术基本要求》,导航电子地图在公开出版时必须进行空间位置技术处理,该技术由指定机构采用国家规定办法统一实现,因此开放给大众的地理坐标是经过首次技术处理的。高德地图、谷歌中国地图和搜搜地图应用该坐标系,而百度地图使用百度坐标系。针对不同的地理坐标数据,需要统一坐标系统,才可进行下一步研究。

3.2 地理编码

链家小区数据中含有地理坐标,因此只需要坐标转换即可,而安居客房屋只有地址,因此需要使用地图 API 得到地理坐标。选用百度 API,一方面是因为目前较为主流的地图 API 为高德和百度,而百度 API 当日可返回数据量高于高德;另一方面是考虑到抓取链家小区经纬度坐标时,该网站采用的是百度 API 接口,以保持地理坐标的一致性。使用 API 类似于爬虫的原理,区别是百度的坐标返回只有单一数据,在浏览器中打开也是一组文本数据,因此读写方便。

首先在百度官网(<http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-geocoding>)申请密钥 AK,在对百度服务器的请求中输入 AK 和要返回坐标的地址。而后对服务器返回的 JSON 结果进行函数解析。实现上述功能的代码示例如下:

```
def parseaddress(url): #对返回结果解析函数
    response=urllib2.urlopen(url)
    s=response.read()
    dic=json.loads(s) #声明一个 JSON 实例
    status=dic.get('status',11) #读取 JSON 中 status 的值,0 为成功返回结果

    if status==0: #如果返回结果成功,那么依次读取参数值
        bd_lng=dic["result"]["location"]["lng"] #规范 JSON 数据格式
        bd_lat=dic["result"]["location"]["lat"]
        precise=dic["result"]["precise"]
        confidence=dic["result"]["confidence"]
    else:
        bd_lng, bd_lat, precise, confidence=0,0,0,0
    data.extend([status, precise, confidence])
    return data #返回 "longitude", "latitude", "status", "precise", "confidence"

urlAddress='http://api.map.baidu.com/geocoder/v2/?address=%s&output=json&ak=申请的'
AK='%('上海市'+address) #向服务器请求数据,输入参数为 AK 和地址 address

dataAddress=parseaddress(urlAddress) #使用函数解析,得到结果
```


4 房产信息在线分析工具开发

根据上述方法,开发可内嵌在浏览器中的房产信息在线分析工具,Web 中数据量膨胀的同时也要求数据分析结果的同步。该工具集成房产数据获取与空间定位功能,用户只需输入要爬取的房产网站,即可获得网站实时数据。同时,工具按照用户选择的统计方法,以地图和图表方式即时显示房产信息的特征。在线工具界面如图 4 所示。

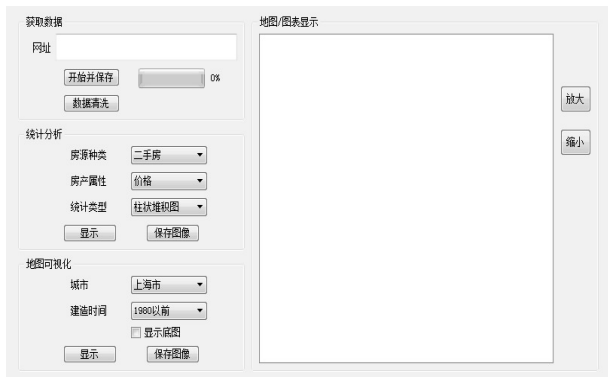


图 4 房产信息在线分析工具使用界面

该工具基于 PyQt 等模块开发,轻量级,用户可以直接下载并运行 .exe 文件即可,不需要配置 Python 环境。在线分析工具主要针对普通用户,使用基本的科学计算方法与直观的地图可视化实现对网站数据的进一步分析。由于其开源的设计,在数据分析方面用户可以自行加入一些计算方法,来设计更符合自身需求的在线分析工具。面对 Web 中大量的房产数据,如何在短时间内高效处理大量数据才是工具面对的关键问题。同时,由于网速等的限制,浏览器加载工具时更希望工具是轻量级的,而不是几十兆、几百兆的应用程序。房产信息开发工具考虑了上述两点,旨在开发面向大众的轻量级工具。以安居客网站和链家网为例,以上海市为研究区域演示工具运行。

4.1 房产信息获取

选择网站:

在“网址”处输入需要抓取的房产网站网址。单击“开始并保存”按钮,获取全网数据。抓取数据存入 Excel 中。考虑到两点因素:一是大量的数据会占用系统内存空间,拖慢计算机运行速度;二是网站中的数据抓取下来可能存在格式错误或无效数据,需要进一步的数据清洗才可以进行分析。由于每次爬虫抓取是向服务器请求当前全部界面,遇到错误数据不得不重新请求相同的网页,对资源造成浪费。基于上述两点,抓取的数据存入 Excel。

以安居客、链家网作为输入 URL,上海市作为抓取城市,截止 2016 年 4 月 20 日,共抓取上海市 16 个行政区(不包括崇明县)小区数据 21 371 条,二手房在

售数据 41 849 条,在售信息中,获取房产属性包括小区名称,地址,所售楼层与该楼房总层数,建造时间,该房屋售价,房屋面积等。小区信息中,“房产属性”包括小区名称地址,建成时间,总楼栋数,总户数,绿化率,容积率等。

4.2 房产的空间分布

(1) 转化为 Shapefile 文件。

在爬虫抓取数据的同时,会相应返回房产经纬度坐标。利用 ArcGIS 中自带的 Python 接口,导入 ArcPy 模块,将点对经纬度坐标转化为点要素并存储在 ESRI Shapefile 文件中。

(2) 地图可视化。

在“地图可视化”组合框中设置各项参数,“城市”选择“上海市”,“建造时间”根据用户需要,单个年份显示,例如 1981 年,1982 年,或者选择周期性年份,如 1980 年以前,1981 年-1990 年等。同时勾选“显示底图”复选框,可将点要素与地图叠合在一起。在图框右侧有“放大”和“缩小”按钮实现对地图缩放操作。

4.3 不同类型房屋数据分析

选择统计分析指标:

在“房源种类”中包括二手房、新房、租房、商业写字楼、海外房产等,工具按照不同类型分析数据。数据分析的指标包括“房产属性”中的价格、面积等。可在“统计类型”中选择不同的统计方式显示结果。

环线二手房、新房房价与出售数量统计图如图 5 所示。

(1) 折线图统计分析。

在“房源种类”中选择二手房,“房产属性”选择价格,“统计类型”选择折线图。以上海市三条交通环线为分割线,将区域划分为四部分,分别是内环以内区域,内环—中环区域,中环—外环区域,外环以外区域。(截止 2016 年 4 月 20 日)内环—中环区域二手房出售数量占 16%,中环—外环区域占 12%,与同一时期新房相比,内环—中环区域,中环—外环区域的二手房,新房均出售较少,反映出该区域是住宅密度较大的城市生活区。内环以内区域,外环以外区域出现了二手房出售的高峰,相对的外环以外区域出现了新房出售的高峰,体现出人们生活水平的提高。从前在郊区住的人会选择靠近市区买房,因此出现郊区售楼高峰;而从前在市中心居住的人们希望在郊区买到环境质量更好的房屋,因此出现了市中心二手房出售和郊区新房出售的高峰。

(2) 饼状图统计分析。

在“房源种类”中选择二手房,“房产属性”选择建造年份,“统计类型”选择饼状图,生成图表表示上海市不同建造年份二手房占比,数据标签在饼图中标出。

其中 2001 年-2010 年的二手房数量占在售二手房的 50%,其次是 1991 年-2000 年,占比为 25%,2011 年至今的二手房占 18%,而这一数字将持续上涨。用户在自己选择二手房时,可参考该统计数据判断房屋新旧和价格的合理性,纠正决策。

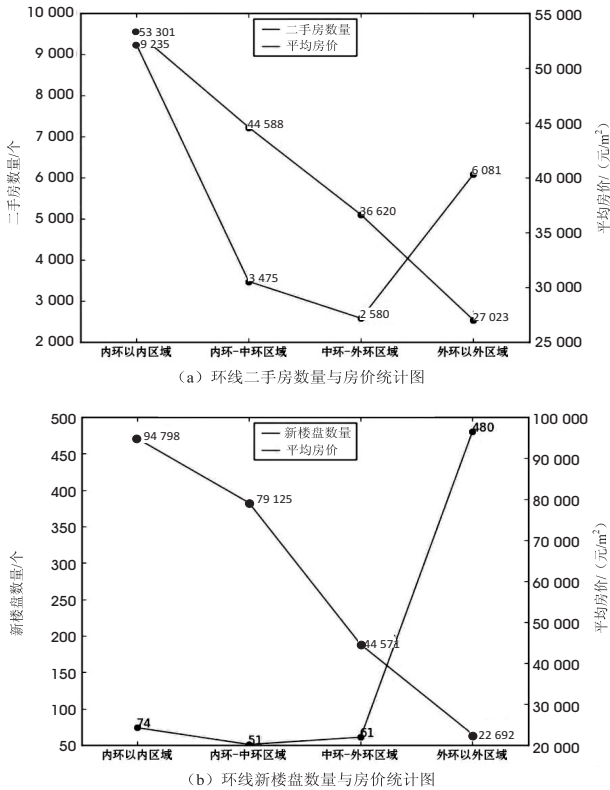


图5 环线二手房、新房房价与出售数量统计图

5 结束语

房产在线分析工具包括抓取数据、统计分析、地图可视化三大功能,旨在面向大众对 Web 数据做进一步的数据分析处理。工具抓取数据方法使用 Scrapy 爬虫,根据 Python 的多线程特点,可快速抓取全网数据,但同时向服务器发送请求也给网站服务器带来了负担。用户可选择所需数据类型进行抓取并保存,减轻服务器压力,也可以设置定时抓取,例如每个月抓取一次数据,存储在本机中,这同样也是一种历史数据保存方式,便于对房产数据做长期数据分析时使用。其统计分析功能利用 Python 科学计算方法,在工具中即时计算结果,并直观显示在图表中。虽然 Python 相较于 C 语言没有计算速度优势,但是万级单位数据还是影响不大的,另外 Python 数据处理方面有很多第三方包,数据可视化同样不亚于 C 语言,其代码比 C 语言要简洁得多,因此将 Python 作为在线工具开发的编程语言。工具的地图可视化借助 ArcPy 函数自动生成地图结果,实时显示在工具界面,使用户对房产数据有一定的空间认识。如今网页已经不仅仅是向人们展示数

据,同时也可以帮助人们计算指标,量化分析数据,更加便于人的决策。在线工具实际上是对网页功能的弥补,轻量级和开源性是在线工具的特点,其优势在于弥补了数据获取与分析的不同步情况,以使用户在短期内掌握房产变化动态,预测趋势等。在线工具经过不断的完善和发展,将不仅局限于对特定网站的数据获取以及分析,还可应用于各个领域。但其局限性也是显而易见的,并不能具备应用程序完整的数据分析功能。就像电子书和纸质书一样,两者永远不会相互替代,反而是相辅相成共同进步。

参考文献:

[1] Liu Bing, Crossman R, Zhai Yanhong. Mining data records in web pages[C]//KDD2003. [s. l.]:[s. n.],2003:601-606.

[2] Zhai Y, Liu B. Web data extraction based on partial tree alignment[C]//International conference on world wide web. [s. l.]:[s. n.],2005:76-85.

[3] 梅雪,程学旗,郭岩,等.一种全自动生成网页信息抽取 Wrapper 的方法[J].中文信息学报,2008,22(1):22-29.

[4] 欧健文,董守斌,蔡斌.模板化网页主题信息的提取方法[J].清华大学学报:自然科学版,2005,45(S1):1743-1747.

[5] 王曙,吉雷静,张雪英,等.面向网页文本的地理要素变化检测[J].地球信息科学学报,2013,15(5):625-634.

[6] 廖邦固,徐建刚,梅安新.1947~2007 年上海中心城区居住空间分异变化—基于居住用地类型视角[J].地理研究,2012,31(6):1089-1102.

[7] 邹高禄,渠文晋,邓沛,等.二手房价格对于住房特征和区位变化敏感性分析[J].西南师范大学学报:自然科学版,2005,30(3):552-555.

[8] 李晓文,方精朴,朴世龙.上海城市土地利用形成、变化及其空间作用机制[J].长江流域资源与环境,2006,15(1):34-40.

[9] 郭太飞,何洁月.归纳学习 XPath Web 信息提取规则[J].计算机技术与发展,2007,17(3):98-101.

[10] 赫特兰. Python 基础教程[M].第2版.北京:人民邮电出版社,2010.

[11] Shaw Z A. 笨办法学 Python[M].王巍巍,译.北京:人民邮电出版社,2014.

[12] Beazley D, Jones B K. Python cookbook[M].南京:东南大学出版社,2014.

[13] Bird S, Klein E, Loper E. Python 自然语言处理[M].陈涛,译.北京:人民邮电出版社,2014.

[14] McKinney W. 利用 Python 进行数据分析[M].唐学韬,译.北京:机械工业出版社,2013.

[15] 高军,杨冬青,唐世渭,等.基于树自动机的 XPath 在 XML 数据流上的高效执行[J].软件学报,2005,16(2):223-232.