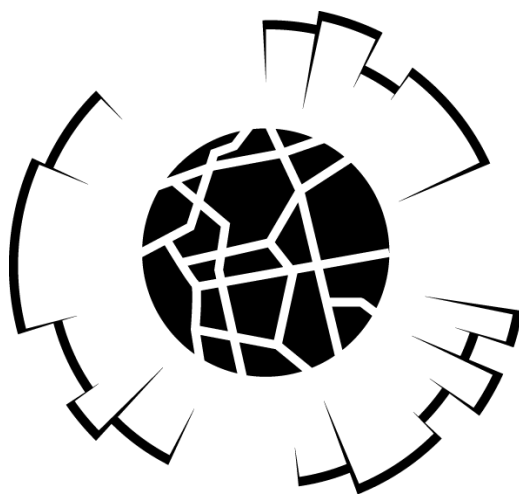


OSpider 用户手册

v1.0.0



二〇一八年七月

目录

1 版本与功能.....	1
1.1 当前版本功能	1
1.2 升级说明.....	1
2 使用教程.....	1
2.1 文件结构说明	1
2.2 操作示范.....	1
2.3 FQA.....	5
3 关于.....	5
3.1 开发者	5
3.2 Bug 报告与意见反馈	6
附录 1 爬取原理与 OSpider 爬取偏差问题.....	7
附录 2 WGS84 与 BD09 坐标问题.....	8

1 版本与功能

1.1 当前版本功能

- 爬取指定区域的百度 POI 数据，并将坐标转化为 WGS84

1.2 升级说明

- 版本号 0.8.0->1.0.0，OSpider 正式对外发布；
- 突破指定区域爬取 poi 数量小于 400 个限制，支持单区域 1w+POI 爬取；
- 绕过了区域中存在多个行政区时只返回一个行政区 POI 的潜在反爬虫机制；
- 爬取 poi 的同时，实现 bd09 坐标到 wgs84 坐标的转换；
- 使用配置文件 property.ini 控制爬取参数；

2 使用教程

2.1 文件结构说明






	OSpider_v1.0.0.exe	2018/7/23 17:15	应用程序	6,258 KB
	OSpider用户手册_v1.0.0.pdf	2018/7/23 18:21	Chrome HTML D...	140 KB
	OSpider_map.html	2018/7/23 16:45	Chrome HTML D...	4 KB
	property.ini	2018/7/23 17:13	配置设置	1 KB
	results	2018/7/23 18:00	文件夹	

图 1 OSpider 相关文件结构图

- **OSpider.exe** 为 OSpider 的主程序，双击后弹出执行窗口，执行 POI 爬取程序。如果杀毒软件报警，请忽略，这是杀毒软件对加壳程序的正常反应，OSpider 很安全。
- **OSpider 用户手册.pdf** 为当前版本的使用说明，帮助用户掌握 OSpider 操作技巧，并对软件使用过程中的常见问题给出解决方案。
- **OSpider_map.htm** 为 OSpider 的辅助 web 工具，用于查询指定行政区域的边界坐标信息。
- **property.ini** 为 OSpider 配置文件，内含爬取参数设置。直接双击打开，根据需求调整参数即可。
- **results 文件夹** 是 POI 爬取结果的储存目录，运行 OSpider.exe 后，爬取结果以.txt 格式存储在该目录中。

2.2 操作示范

需求：爬取武汉市内所有的烟酒店(liquor store)，并转化为 shp 文件，为后续探究武汉市烟酒店空间格局打下数据基础。

操作流程：确定参数->执行程序->业务使用

STEP 1 (定义参数)：

打开百度地图 (<https://map.baidu.com/>) 寻找合适的关键词获得 liquor store，并考量 liquor store 的量级。尝试‘烟酒店’、‘烟酒超市’、‘烟酒副食’、‘烟酒’作为关键词查询后，认为‘烟酒’更为合适，此时，根据百度地图反馈武汉市有 4813 个相关 POI。

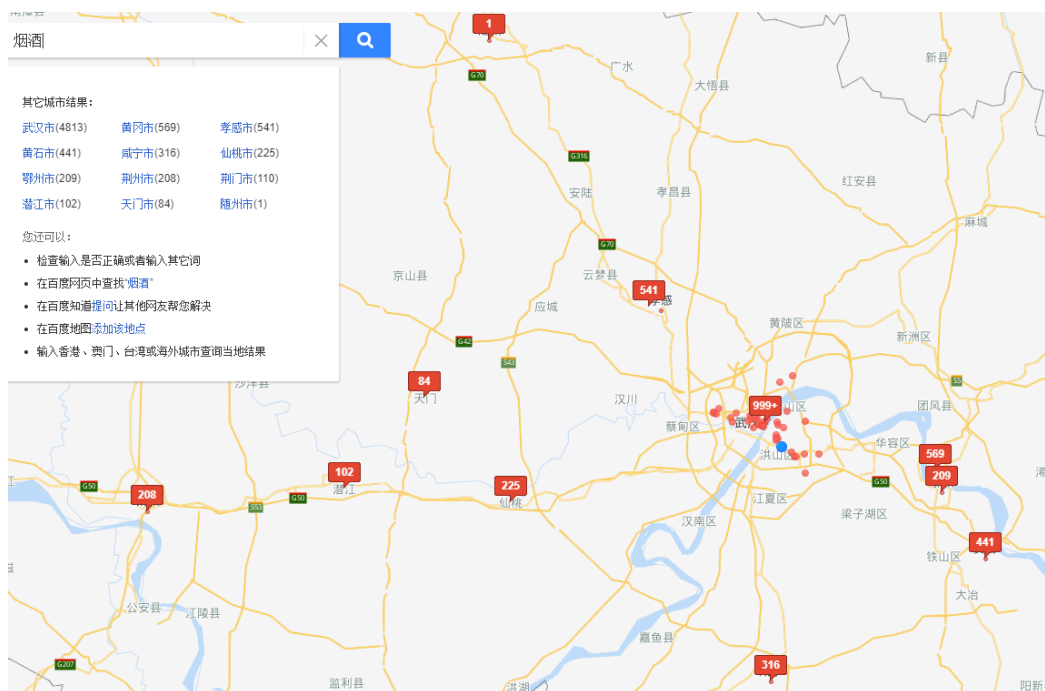


图 2 确认关键词及 POI 量级

打开 OSpider_map.html, 在查询框中输入“武汉市”, 查询得到武汉市行政区的边界信息。其中图 3 中红框标明部分是需要参数:“经纬度:左下角, 右上角: 113.707695,29.972898;115.085775,31.367052”。

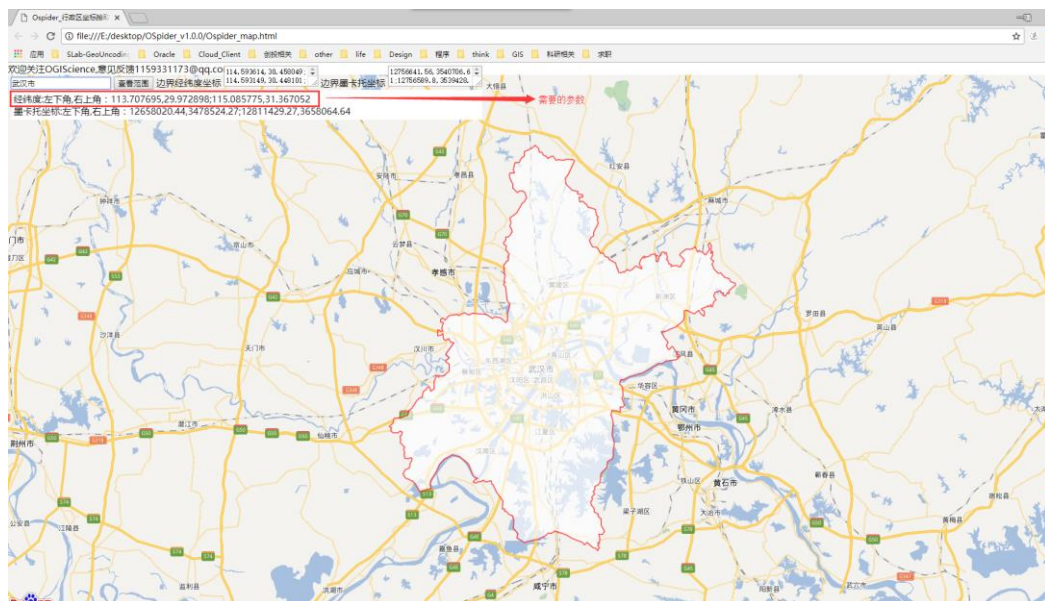
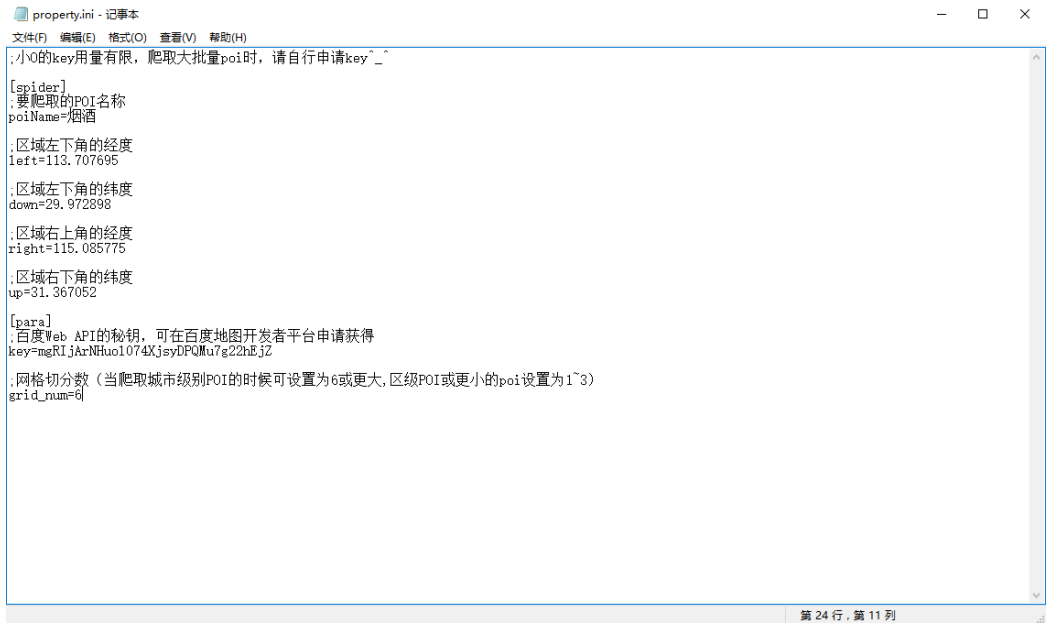


图 3 获得爬取区域的坐标

STEP 2 (执行程序):

双击打开 property.ini (在多数电脑上, 可以直接双击用记事本打开), 填写参数。参数主要包括两个部分, 第一部分为[spider], 包括要爬取的 POI 名称 (poiName)、爬取区域左下角和右上角经纬度坐标 (left/down/right/up), 该部分根据 STEP 1 结果填写即可; 第二部分为[para], 该部分包括百度 WebAPI 秘钥 (key) 和初始网格切分数 (grid_num)。原始的秘钥是小 O 提供的, 仅供测试和小规模爬取解燃眉之需, 如果需要大规模爬取 POI, 还请自行登录百度地图开发者平台, 申请 key。初始网格切分数(grid_num)一般设置为 6 就足够了,

可不改动，关于初始网格切分数的详细讨论请参见“附录 1 爬取原理与 OSpider 爬取偏差问题”。



```
property.ini - 记事本
文件(F)  编辑(E)  格式(O)  查看(V)  帮助(H)

;小V的key用量有限，爬取大批量poi时，请自行申请key~^

[spider]
;要爬取的POI名称
poiName=烟酒

;区域左下角的经度
left=113.707695

;区域左下角的纬度
down=29.972898

;区域右上角的经度
right=115.085775

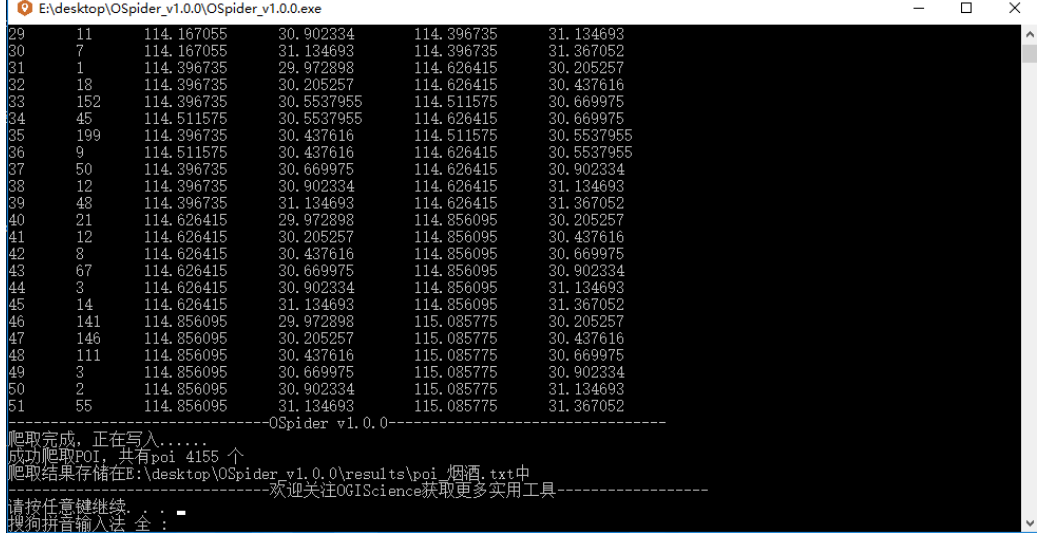
;区域右下角的纬度
up=31.367052

[para]
;百度Web API的秘钥，可在百度地图开发者平台申请获得
key=ngRiJArNhuo1074XjsyDPQMuIg22hEjZ

;网格切分数（当爬取城市级别POI的时候可设置为6或更大，区级POI或更小的poi设置为1~3）
grid_num=6
```

图 4 配置文件

完成参数设置后，双击 OSpider.exe 执行爬取操作。此时会弹出黑色命令行，不断反馈爬取进展，一般上万的 POI 几分钟就爬完了，几千的分分钟，几百的就更快了。如果长时间没有爬完，可能是关键词选择问题，请关闭程序，重新选择关键词。程序执行完毕后，按任意键可结束程序。



```
E:\desktop\OSpider_v1.0.0\OSpider_v1.0.0.exe

29 11 114.167055 30.902334 114.396735 31.134693
30 7 114.167055 31.134693 114.396735 31.367052
31 1 114.396735 29.972898 114.626415 30.205257
32 18 114.396735 30.205257 114.626415 30.437616
33 152 114.396735 30.5537955 114.511575 30.669975
34 45 114.511575 30.5537955 114.626415 30.669975
35 199 114.396735 30.437616 114.511575 30.5537955
36 9 114.511575 30.437616 114.626415 30.5537955
37 50 114.396735 30.669975 114.626415 30.902334
38 12 114.396735 30.902334 114.626415 31.134693
39 48 114.396735 31.134693 114.626415 31.367052
40 21 114.626415 29.972898 114.856095 30.205257
41 12 114.626415 30.205257 114.856095 30.437616
42 8 114.626415 30.437616 114.856095 30.669975
43 67 114.626415 30.669975 114.856095 30.902334
44 3 114.626415 30.902334 114.856095 31.134693
45 14 114.626415 31.134693 114.856095 31.367052
46 141 114.856095 29.972898 115.085775 30.205257
47 146 114.856095 30.205257 115.085775 30.437616
48 111 114.856095 30.437616 115.085775 30.669975
49 3 114.856095 30.669975 115.085775 30.902334
50 2 114.856095 30.902334 115.085775 31.134693
51 55 114.856095 31.134693 115.085775 31.367052

-----OSpider v1.0.0-----
爬取完成，正在写入.....
成功爬取POI，共有poi 4155 个
爬取结果存储在E:\desktop\OSpider_v1.0.0\results\poi_烟酒.txt中
-----欢迎关注OGIScience获取更多实用工具-----
请按任意键继续.
搜狗拼音输入法 全：
```

图 5 OSpider 执行窗口

爬取结果以文本文件格式存储在 results 文件夹中，文件包括爬取信息头和结果两个部分。数据与数据之间采用的是\t 间隔，这使得把数据复制到 Excel 表格中时，能够自动分列。

poi_烟酒.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

于 2018-07-25 10:23:26 提取指定范围内的：烟酒，总计 4155 个

区域左上坐标为：113.707695, 30.672998

区域右上坐标为：115.085775, 31.367052

选择的基础网格切分方式为：6*6网格

results

id	name	address	province	city	area	wgs84_lng	wgs84_lat	bd09_lng	bd09_lat
1	春源酒庄	湖北省荆州市洪湖市文卫路大沙田园农庄西北100米	湖北省	荆州市	洪湖市	113.857304675	30.0094669885	113.86923	30.012975
2	盛源烟酒批发	湖北省武汉市蔡甸区318国道附近	湖北省	武汉市	蔡甸区	113.714416911	30.4090404707	113.726735	30.412335
3	正九纯粮酒专卖(汉南旗舰店)	湖北省孝感市汉川市康泰路40号城发福源	湖北省	孝感市	汉川市	113.849431996	30.2653912967	113.861442	30.268797
4	便民烟酒副食	湖北省武汉市蔡甸区香涛村育才小学西300米	湖北省	武汉市	蔡甸区	113.802750392	30.32985534	113.814957	30.332972
5	聚源烟酒副食批发	湖北省武汉市蔡甸区318国道附近	湖北省	武汉市	蔡甸区	113.713020747	30.4067742208	113.725345	30.410046
6	桂柱中心烟酒副食批发	成功现代都市农业发展有限公司成功街西北路74号	湖北省	武汉市	蔡甸区	113.713261841	30.4072773513	113.725585	30.410553
7	鑫源酒庄	湖北省孝感市汉川市家洪粮油副食(龙腰路东)	湖北省	孝感市	汉川市	113.811652916	30.5696030722	113.823872	30.572707
8	元泰烟酒批发	马口镇广福街123号	湖北省	孝感市	汉川市	113.819244997	30.5647384315	113.831441	30.567835
9	经纬名烟名酒	仙玄山街道人民大道同仁堂药店对面	湖北省	孝感市	汉川市	113.825835133	30.6614598692	113.838027	30.664637
10	兴持酒业	孝感市汉川市三台一路61-2号	湖北省	孝感市	汉川市	113.829817714	30.6504713895	113.841995	30.653685
11	威凯烟酒副食	仙玄山街道体育馆路报于城市酒店	湖北省	孝感市	汉川市	113.825625631	30.6588909997	113.837821	30.662066
12	收售名烟名酒	湖北省孝感市汉川市康城大道汉川汽车客运中心站南100米	湖北省	孝感市	汉川市	113.823716506	30.6666856258	113.835915	30.669848
13	少川名烟名酒批发	湖北省孝感市汉川市中港花园(体育馆路北)	湖北省	孝感市	汉川市	113.825705537	30.659403268	113.8379	30.662579
14	仁和烟酒	湘军街城北水陆派出所旁里潭建筑商	湖北省	孝感市	汉川市	113.828786059	30.6631411669	113.840959	30.666346
15	钟氏烟酒副食批发	湖北省孝感市汉川市农商银行蚌湖支行东南(106省道)	湖北省	孝感市	汉川市	113.701788189	30.6169880224	113.714154	30.620183
16	杨城烟酒	湖北省孝感市汉川市嘉善街会所南(柳车街)	湖北省	孝感市	汉川市	113.829154943	30.6659868361	113.841323	30.669196
17	立成烟酒	湖北省孝感市汉川市体育路147号-7	湖北省	孝感市	汉川市	113.829591901	30.658551372	113.841764	30.661764
18	至上海烟酒	湖北省孝感市汉川市西湖大道54号	湖北省	孝感市	汉川市	113.816399307	30.6494056281	113.828644	30.65254
19	意和祥烟酒	湖北省孝感市汉川市仙玄大道98号	湖北省	孝感市	汉川市	113.826441389	30.6491270321	113.83864	30.652307
20	富丽华名烟名酒商行	湖北省孝感市汉川市西湖大道73-1号	湖北省	孝感市	汉川市	113.814383345	30.6499258938	113.826634	30.653062
21	195名烟名酒	仙玄大道213号附近	湖北省	孝感市	汉川市	113.824334307	30.6526948945	113.838542	30.655769
22	御清香烟酒	湖北省孝感市汉川市文化路27-4	湖北省	孝感市	汉川市	113.831768427	30.6544099513	113.84393	30.657647
23	金鑫烟酒	人民大道250-16	湖北省	孝感市	汉川市	113.825603889	30.6615097392	113.837797	30.664685
24	鼎成烟酒	湖北省孝感市汉川市湘军路34号	湖北省	孝感市	汉川市	113.828576042	30.6645310932	113.840749	30.667734
25	江江烟酒批发	湖北省孝感市汉川市北正街城中路	湖北省	孝感市	汉川市	113.835546333	30.6499332211	113.847696	30.652224
26	财利烟酒茶批发	湖北省孝感市汉川市人民大道27号	湖北省	孝感市	汉川市	113.826176966	30.661471024	113.838366	30.664651
27	荣华烟酒城	湖北省孝感市汉川市中国农业银行(人民大道)	湖北省	孝感市	汉川市	113.827058092	30.661422274	113.839243	30.6646
28	高荣烟酒批发部	湖北省孝感市汉川市山后一路5号	湖北省	孝感市	汉川市	113.834045003	30.6506814903	113.846194	30.653947
29	名人烟酒	湖北省孝感市汉川市北桥路枫梓北苑	湖北省	孝感市	汉川市	113.835086498	30.663736024	113.847217	30.667018

第 3 行, 第 23 列

图 6 结果文件

STEP 3 (业务使用)：

该部分根据实际需求来定。本范例中，小 O 先把 txt 中的数据复制到.xlsx 文件中，再加载到 Arcgis 中，确认无误后导出为.shp 文件。

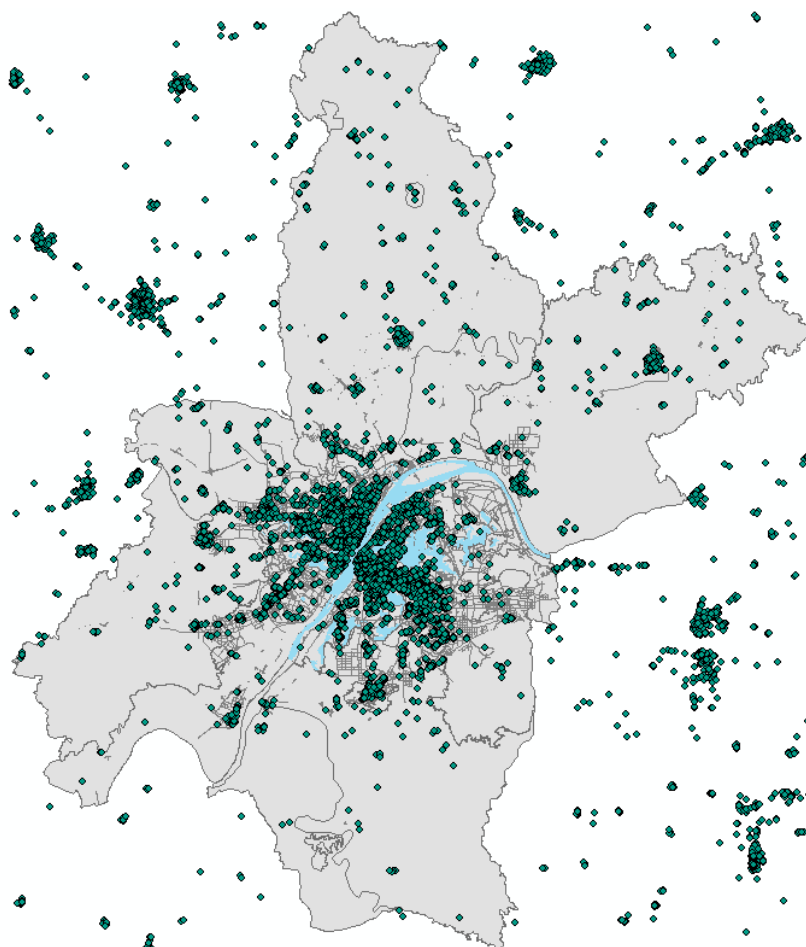


图 7 在 Arcgis 中加载爬取的 POI 数据

2.3 FQA

- **如何获取我想要爬取的区域的坐标？**

如果您希望爬取某一行政区的 POI 数据，请打开 Ospider_map.html 查询行政区边界坐标；如果您希望其他区域的相关坐标，请使用百度坐标拾取器 (<http://api.map.baidu.com/lbsapi/getpoint/index.html>)。

- **为什么程序会闪退？**

请检查 OSpider.exe 文件路径中是否存在中文，如果路径中含有中文，经常会出现闪退，改为英文路径即可；请检查 OSpider 可执行程序目录下是否有 results 目录，若没有请手动添加后再次运行程序；确认网络状况是否良好，保持 ip 稳定（使用 vpn 切换网络状态时会造成程序中断）；确认指定 key 对应的 Place api 是否达到最大使用次数，可更换 key 值重新尝试。

- **为什么 grid_name 设置不同，爬取的同区域同种 POI 数量不同？为什么爬取的 POI 数量与百度地图（map.baidu.com）上显示的不一致？**

参见“附录 1 爬取原理与 OSpider 爬取偏差问题”。

- **小 O 帅不帅？应不应该关注 OGIScience？**

帅 !!! 100%应该。

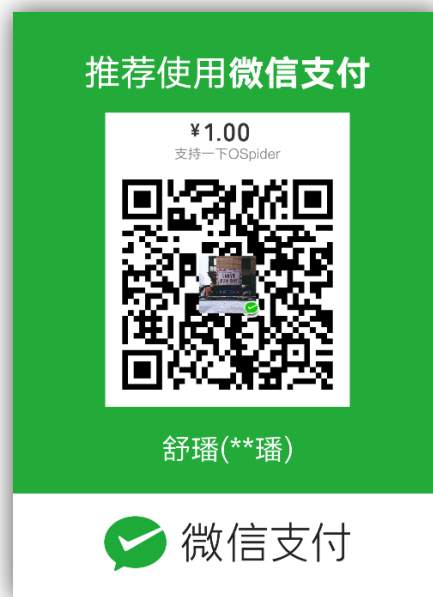
3 关于

3.1 开发者

武大城市化研究室 小 O

“如果 OSpider 切实帮助了您，欢迎向小 O 赞赏支持哦（1 块也 odk 哟）。以及，记得关注小 O 的个人公众号 OGIScience 获取更多资讯和实用工具~”





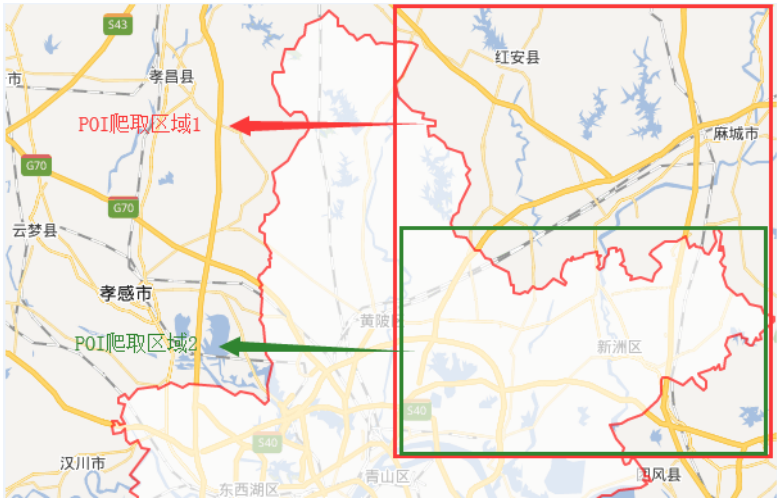
3.2 Bug 报告与意见反馈

如果您在使用 OSpider 的过程中发现 Bug 或对 OSpider 的使用有什么意见或建议，
请向 1159331173@qq.com 发送邮件，小 O 收到邮件后会尽快回复。

附录 1 爬取原理与 OSpider 爬取偏差问题

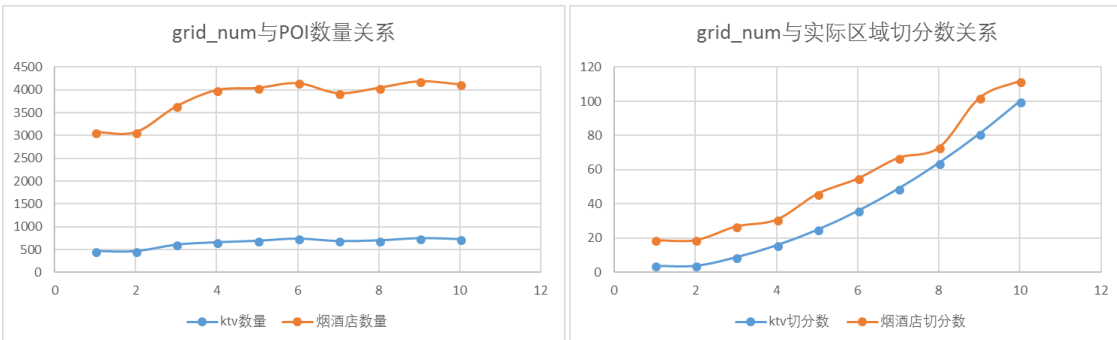
OSpider v1.0.0 使用百度的 Place API 进行爬取，该 API 能够提供多种场景的 POI 检索功能，具体包括城市检索、周边检索、矩形区域检索和地点详情检索。其中行政区划区域检索可以检索某一行政区划内（目前最细到城市级别）的地点信息；圆形区域检索允许开发者设置圆心和半径，检索圆形区域内的地点信息；矩形区域检索可通过设置检索区域左下角和右上角坐标，检索坐标对应矩形内的地点信息。

使用 Place API 进行爬取的时候主要存在两个问题。第一个是，不论是行政区还是区域，结果列表最多只返回 400 个 POI；第二个问题是当检索区域存在多个行政区时，结果只返回其中一个行政区的检索结果（下图中，区域 1 返回 POI_ktv 127 个，区域 2 返回 POI_ktv 145 个）。



本文采用手动网格切分+自动四分递归的方法来进行 POI 爬取，完美突破第一个反爬虫机制，并部分绕过第二个反爬虫机制。程序运行时，读取配置文件中的 grid_num，将爬取区域切分成 grid_num*grid_num 的小区域，然后对每一个小区域进行四分递归爬取。具体而言为，判断该区域的 POI 数量是否在 400 及以上，若在 400 及以上就将该区域切分成四块，爬取其子区域的 POI；如果该区域的 POI 数量在 400 以下，便直接爬取 POI。

一方面由于百度地图（map.baidu.com）主页上的检索模式不仅仅是关键词检索，另一方面百度地图 Place API 的第二种反爬虫机制，当前 OSpider 爬取大范围 POI 的结果与百度地图查询显示的结果有一定出入，且爬取结果随着 grid_num 的不同而不同。虽然还存在一定的不足，但是通过设置相对合理的 grid_num，爬取的城市尺度上的 POI 已经足够普通业务、科研和学习使用了。



小 O 测试了 ktv 和烟酒店两种 POI 爬取与 grid_num 的关系，结果如上图所示。建议爬取城市级 POI（以武汉为例）的时候，设置 grid_num 为 6 或 9 较好。

附录 2 WGS84 与 BD09 坐标问题

wgs84 坐标是世界通用、科研常用的地理坐标系，google 采用的坐标系也是 wgs84，国内从事科研时，很多时候也用 GCGS2000 坐标系，这个坐标系和 wgs84 基本一致，除非特别高精度要求，一般可以把 wgs84 的数据当 GCGS2000 的或把 GCGS2000 的当 wgs84 的来用。

国内测绘地理信息行业有保密条例，要求对外使用的至少是经过国测局加密一次的 GCJ-02 坐标系（火星坐标），而不同互联网地图服务商又往往会在 GCJ-02 的基础上进行二次加密，百度二次加密后的坐标系是 BD-09。

OSpider 内置了坐标转换程序，输出结果中同时保留了 wgs84 和 bd09 坐标。下图中武汉市的底图是 GCGS 2000 地理坐标系加投影的，红色的点是把 wgs84 坐标定义为 GCGS 2000 地理坐标系后加载入 Arcgis 的结果，而绿色的点是把 bd-09 坐标定义为 GCGS2000 地理坐标系后加载入 Arcgis 的结果。可以看出，绿色点偏差明显很多点掉入了江湖，而红色点与底图相匹配。

