

OpenStreetMap 项目研究

一、地图选择

这次研究选择了美国的芝加哥地区，因为整个课程是以芝加哥为例进行分析，而且第一次接触 OSM，避免出现无法处理的问题，保守选择了比较熟悉的数据集。

芝加哥地图链接：

<https://www.openstreetmap.org/search?query=chicago#map=12/41.8500/-87.6500>

二、地图数据中存在的问题

首先，下载芝加哥部分地区的数据集（104M）后，鉴于数据集比较大，不便于审查，因此将数据集切成 7M 大小进行审查；其次对比 104M 数据集，发现三个问题：

1. 街道名存在极个别缩写的情况，如：‘N LaSalle St, #575’、‘W. Madison St.’ 等；
2. `<k=addr:street:type>` 标签中 ‘v’ 值存在 ‘St’ 缩写；
3. 二级标签 `tag` 中存在 GPS 数据缩写的情况，如：

```
<tag k="tiger:name_type" v="Ave" />
<tag k="tiger:name_direction_prefix" v="S" />
<tag k="tiger:name_direction_prefix" v="W" />
```

鉴于将数据导入 SQL 后，数据没有标准化会对查询结果产生影响，同时导入后的数据修改会花费很大的工作量，因此在数据导入前利用了 `update_value` 方法进行了数据修改：








```
def update_value(value, mapping):
    if value in mapping:
        value=mapping[value]
    else:
        last=value.split()[-1]
        if last in mapping:
            value=value.replace(last,mapping[last])
    return value
```

这种方法会将有问题的值修改为标准化的值；如：

W. Madison St.==> West Madison Street

三、数据集概述统计

1. 文件大小

名称	修改日期	类型	大小
 chicago.osm	2017/7/9 15:33	OSM 文件	107,422 KB
 chicago.db	2017/7/11 13:24	Data Base File	74,726 KB
 nodes.csv	2017/7/10 22:16	Microsoft Office...	38,728 KB
 ways_tags.csv	2017/7/10 22:16	Microsoft Office...	14,580 KB
 ways_nodes.csv	2017/7/10 22:16	Microsoft Office...	12,781 KB
 ways.csv	2017/7/10 22:16	Microsoft Office...	4,746 KB
 nodes_tags.csv	2017/7/11 13:18	Microsoft Office...	450 KB

2. 唯一用户的数量

```
sqlite> select count( distinct f.uid) from (select uid from nodes union all select uid from ways) as f;
362
```

3. 对 openstreetmap 作出贡献最多的 10 个用户：

```
sqlite> select f.user,count(*)
...> from (select user from nodes union all select user from ways) as f
...> group by f.user
...> order by count(*) desc
...> limit 10;
chicago-buildings!420588
Chicago Park District GIS!14282
Umbugbene!13425
Steven Vance!9013
bbmiller!8413
NE2!8328
Zol87!2990
boeleman81!2332
Eliyak!1477
mappy123!1472
```

4. 节点和途径的数量

节点数量：

```
sqlite> select count(*) from nodes;
432083
```

途径数量:

```
sqlite> select count(*) from ways;
71761
```

5. 所选节点类型（如：咖啡店、商店等）的数量

选取了餐馆、披萨店和咖啡店进行了统计，具体数量如下：

（1）餐馆数量：

```
sqlite> select count(*) from nodes_tags where value='restaurant';
185
```

这些餐馆中菜肴数量、类型如下：

```
sqlite> select nodes_tags.value,count(*) as number
...> from nodes_tags ,(select distinct(id) from nodes_tags where value='resta
urant') as f
...> where nodes_tags.id =f.id
...> and nodes_tags.key='cuisine'
...> group by nodes_tags.value
...> order by number desc
...> limit 10;
mexican!20
american!11
italian!10
pizza!8
thai!8
japanese!6
sushi!6
sandwich!5
greek!4
mediterranean!3
```

（2）披萨店数量：

总数量：

```
sqlite> select count(*) from nodes_tags where value like '%Plaza';
8
```

分类数量：

```
sqlite> select value,count (*) from nodes_tags where value like '%Plaza' group b
y value order by count(*) desc;
Merchandise Mart Plaza!3
Holiday Inn Chicago Mart Plaza!1
North Riverside Plaza!1
Orleans St & Merchandise Mart Plaza!1
River West Plaza!1
South Riverside Plaza!1
sqlite>
```

（3）咖啡店数量：

```
sqlite> select count(*) from nodes_tags where value like '%cafe%' or value like
'%coffee%';
105
```

四、建议

贡献排名前十的用户中，chicago-buildings 排名第一，贡献了 420588 节点，占到了总数的 83.48%，也就是说绝大部分节点是由他来完成的；而前十名用户贡献占比是 95.73%，10 个人完成了芝加哥部分地区的节点，工作量非常大，导致数据存在很多问题，比如字符串没有标准化，存在缩写的情况；分类不明确，piazza 应该放入 restaurant 类中还是单独作为一类等。

为了调动群众的积极性去完成 openstreetmap，可以设置一个激励措施或者竞争机制，如排行榜、虚拟货币等，促使更多的人投入到这项工作中去，让 openstreetmap 逐步实现标准化、规范化、细致化。

但是也会出现一些问题：（1）操作失误，在不知情的情况下把原本正确的数据也进行了修改，那么这个脏数据很难被发现及修改；（2）丢失更新，一个数据被两个及以上的人同时修改时，会出现更新丢失的情况；（3）多个标准，每个人进行数据修改时，可能会根据自己的标准进行操作，这样会导致整个数据集中多个标准的情况，一个字形容就是‘乱’。