城市研究之互联网数据抓取：疫情数据

李栋，博士，副主任

北京清华同衡规划设计研究院 · 技术创新中心

2020年03月16日

1

北京清华同衡规划设计研究院有限公司

大型国有独资规划设计企业

- 1993年，北京清华城市规划设计研究院成立，隶属于清华大学。
- 2012年8月，由全民所有制改为有限责任公司，整体转至清华控股有限公司，正式更名为北京清华同衡规划设计研究院有限公司。
- 2013年3月，清华同衡与多家公司进行战略重组、整合为清控人居集团。
- 清华大学产业集团中唯一专门从事城市规划设计研究与实践的公司。

全产业链和跨专业协同的综合解决方案

清华同衡提供从战略、规划、设计到落地实施、运营的全产业链综合技术服务，形成了涵盖城乡规划、风景园林、建筑设计、市政交通及科技传媒5大板块的跨专业协同服务模式，将先进的理念、科学技术支撑与地方特色相结合，为国家部委、省、市、县（区、市）、街道镇乡村各级这政府，省直、市直部门，园区管委会，大型公司（城市建设投资运营公司、产业运营公司等）提供个性化订制服务。

- 城乡规划编制甲级资质
- 国家旅游规划设计甲级资质
- 文物保护工程勘察设计甲级资质
- 风景园林工程设计专项甲级资质
- 建筑行业（建筑）专业甲级资质
- 土地规划乙级资质
- 通过ISO9001:2008质量管理体系认证

清华同衡将规划设计与科研成果转化为支撑城乡发展的源动力，持续以规划和技术回报社会，并积极投身社会公益事业，主动承担清华人“家国天下”的行业使命和社会责任。

清华同衡技术创新中心

由清华同衡规划设计研究院于2014年组建成立

- 国内规划设计单位首家成立的**大数据研发**部门
- 根植于清华大学深厚的学术积累，以及清华同衡对城市的深刻理解，利用多源数据开展城市创新型智库业务
- 提供城市规划管理领域新兴服务模式和技术产品

北京清华同衡规划设计研究院有限公司
www.thupdi.com

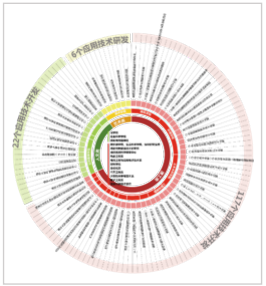
清华同衡技术创新中心
ict.thupdi.com

2

3

关于我们

优势特色

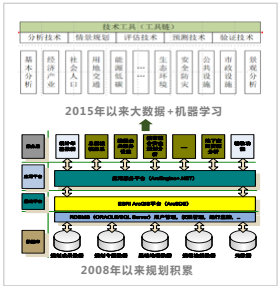


人才优势

依托清华大学学术积累，具有多年课题研究和规划实践经验的专家领衔。
团队具备规划、交通、数学、计算机等多学科交叉背景人才优势。

数据优势

多源数据渠道的融合：政府数据+商业数据，交通、产业、人口、养老、环境数据的融合与积累。
精确而动态：基于互联网和移动设备终端产生的多源用户数据对城市问题的了解精准且量化。
结合时空技术手段：数据源与专业的城市空间分析手段相融合



知识优势

问题与需求的长期积累：依托清华同衡多年面向政府对城市管理、规划和运营的咨询实践，对城市发展及政府需求有深刻与专业的理解。
丰富行业知识的沉淀：数十年中外城市发展相关专业理论研究与技术工具积累。

开发优势

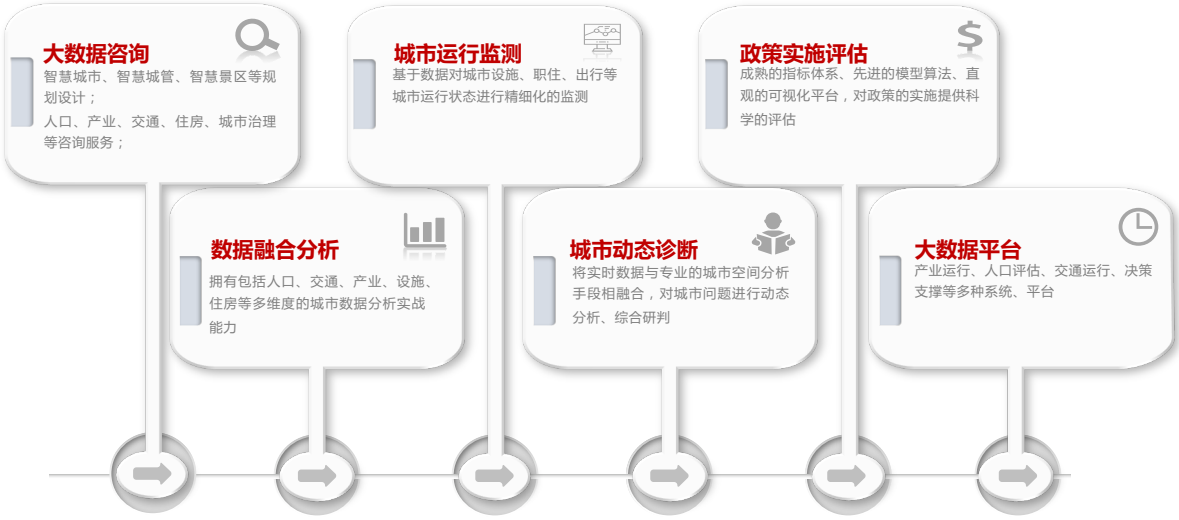
专业算法与模型的积累：时空分析模型、复杂网络分析模型、机器学习、深度学习在城市发展问题中的应用。
行业信息化系统开发经验的积累：具备跨部门跨学科综合分析能力与多种系统开发经验，可对政策实施进行模拟和评估。

清华同衡技术创新中心

4

关于我们

核心能力



提供大数据背景下城市规划、运行、管理与服务的全链条解决方案

清华同衡技术创新中心

5

开场介绍

- 破题：面向城市研究的互联网数据获取方法
  - 城市研究：关注数据的空间属性、时间属性、附加属性
  - 少量编程 = 使用某个具体软件
    - 获取、处理、分析、表达……
    - 专业软件、相关软件
  - **本次交流侧重互联网数据的获取**
    - 分析部分相对简略：如SPSS……
    - 简单涉及可视化
  - 每个软件介绍主要功能和实践经验

6

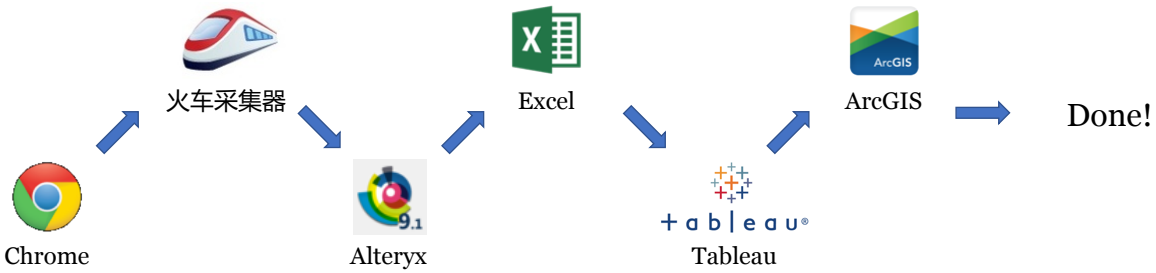
开场介绍

- 作为XXX的互联网
  - 通讯工具、社交平台、电子商务……
- 作为数据资源的互联网
  - 在线的网页、文本
  - 在线的数据库
  - 在线的程序接口
  - ……

7

开场介绍

- 工具：获取/处理 → 分析 → 表达



8

开场介绍

- Magic Quadrant for Advanced Analytics Platforms 2015
  - By Gartner
  - execute vs. vision

9

开场介绍

- Magic Quadrant for Data Science Platforms 2020
  - By Gartner
  - execute vs. vision

Figure 1. Magic Quadrant for Data Science and Machine Learning Platforms

Source: Gartner (February 2020)

10

开场介绍

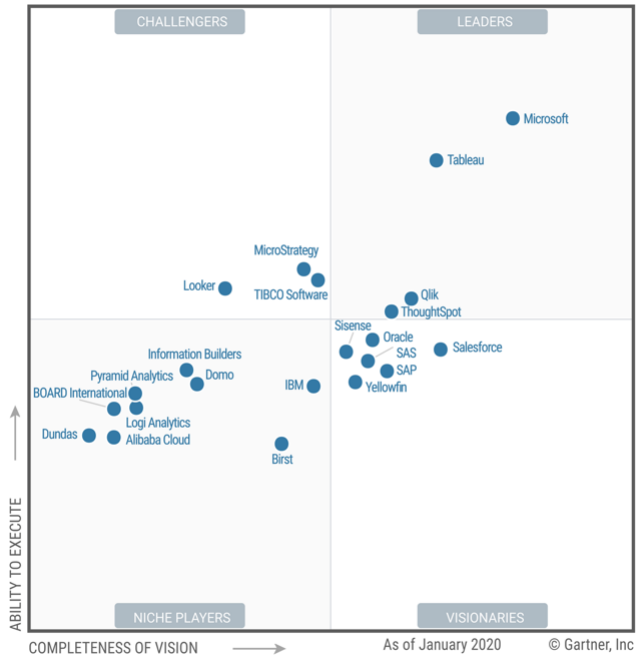
- Magic Quadrant for BI 2015
  - By Gartner
  - execute vs. vision

11

开场介绍

- Magic Quadrant for Business Intelligence and Analytics 2020
  - By Gartner
  - execute vs. vision

Figure 1. Magic Quadrant for Analytics and Business Intelligence Platforms



Source: Gartner (February 2020)

12

数据的获取与处理

- 探查、定位数据源的网络地址
- 访问该网址，并将结果保存在本地
  - 自动化：持续、反复、定时……
- 处理、清洗数据
  - 根据分析任务而定
  - 根据数据特征而定

Chrome

火车采集器

Alteryx

SAFE FME

Excel

13

数据的获取与处理

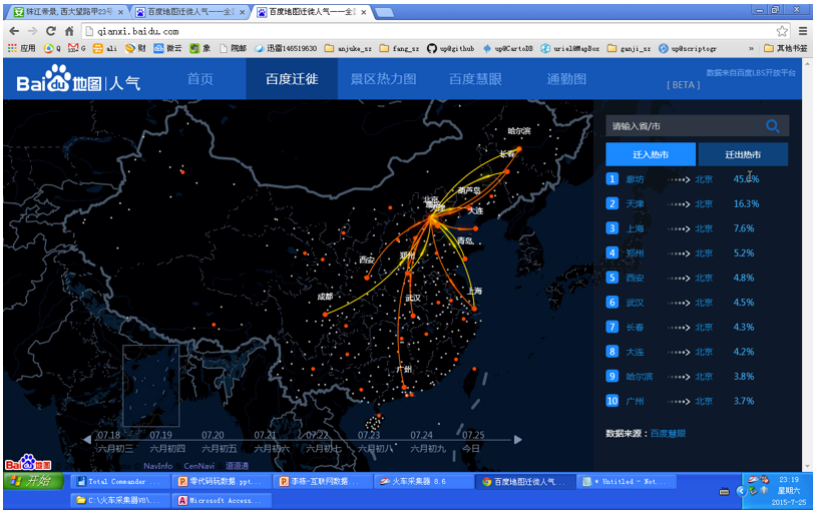
- Chrome开发者工具（Chrome DevTools）
  - 查看网页要素
  - 抓包分析
  - Javascript调试
  - ……

14

数据的获取与处理

- 抓包工具：发现隐藏的数据URL

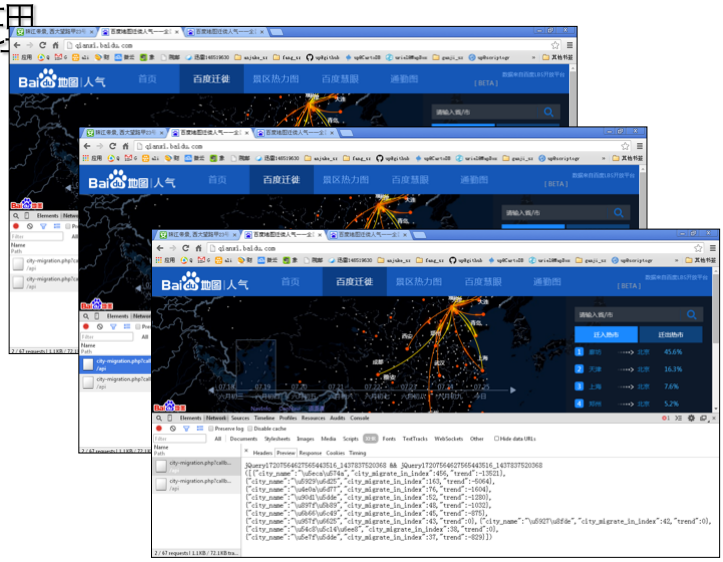
http://qianxi.baidu.com



15

数据的获取与处理

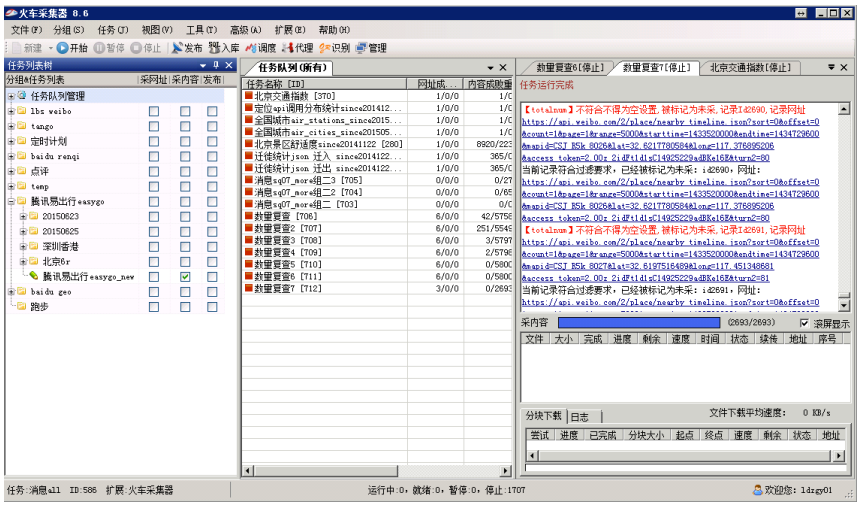
- 抓包工具：发现隐藏的数据URL
- XHR = XMLHttpRequest
  - 实现局部更新网页



16

数据的获取与处理

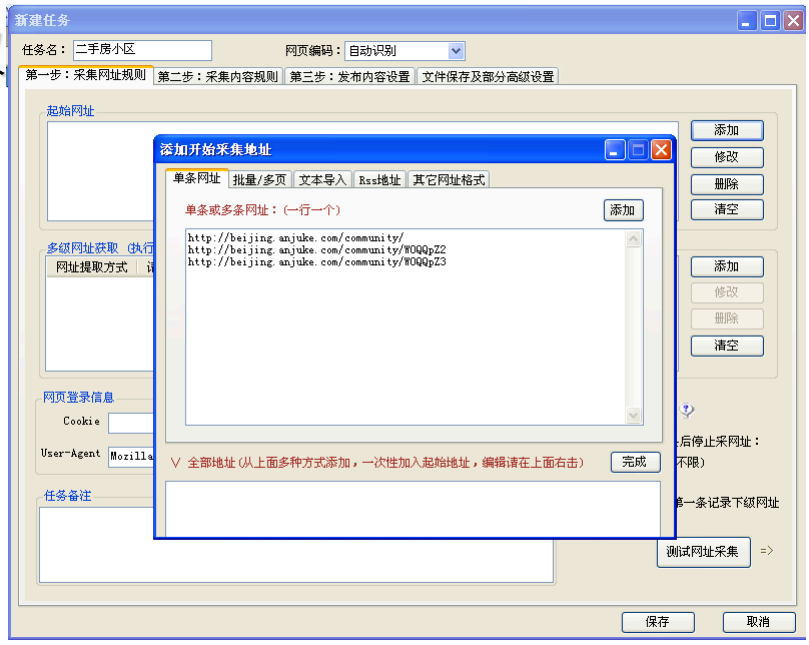
- 火车采集器
  LocoySpider



17

数据的获取

- 火车采集器
  1. 添加网址



18

数据的获取与处理

- 火车采集器
  - 抓包工具：发现隐藏的数据URL

```
http://qianxi.baidu.com/api/city-migration.php?callback=jQuery172075646275654435
16_1437837520368&type=migrate_in&sort_by=low_index&limit=10&city_name=%E5%8C%97%E4%BA%AC&date_start=20150725&date_end=20150725&_=1437837520779
```

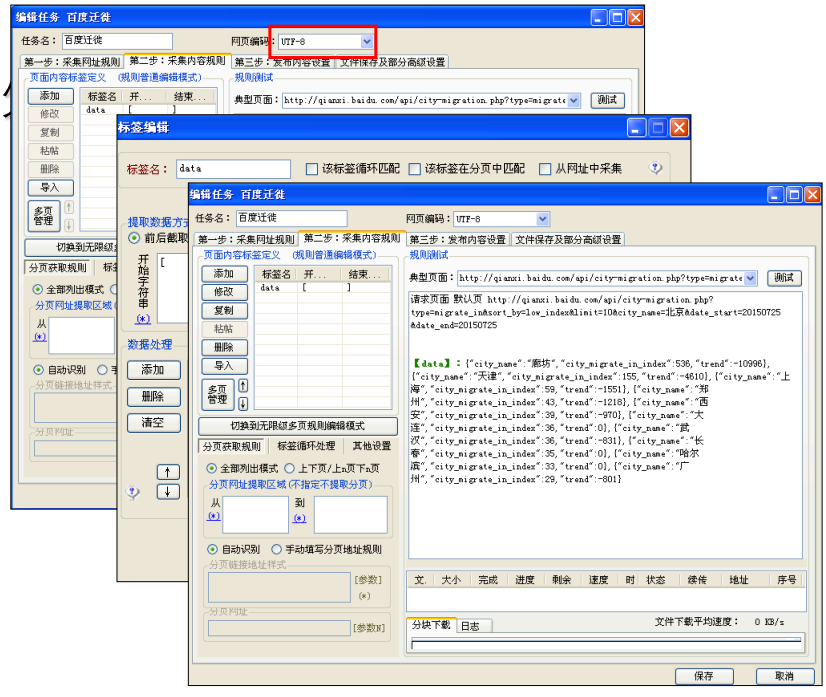
```
http://qianxi.baidu.com/api/city-migration.php?type=migrate_in&sort_by=low_index&limit=10&city_name=北京&date_start=20150725&date_end=20150725
```

```
http://qianxi.baidu.com/api/city-migration.php?type=migrate_out&sort_by=low_index&limit=10&city_name=北京&date_start=20150725&date_end=20150725
```

19

数据的获取与

- 火车采集器
  - 抓包工具：发现隐藏的数据URL

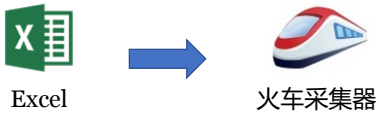


20

数据的获取与处理

- 火车采集器
  - 抓包工具：发现隐藏的数据URL
  - 按照规则生成URL列表
    - Type X 地名 X 日期

```
http://qianxi.baidu.com/api/city-migration.php?type=migrate_in&sort_by=low_index&limit=10&city_name=北京&date_start=20150725&date_end=20150725
```



21

数据的获取与处理

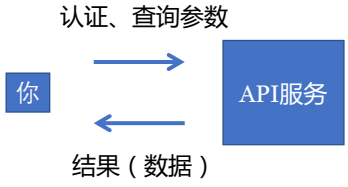
- 火车采集器
  - API类型数据的获取

API（Application Programming Interface，应用程序编程接口）是一些预先定义的函数，目的是提供应用程序开发人员基于某软件或硬件得以访问一组程序的能力，而又无需访问源码，或理解程序内部工作机制的细节。

•非官方的微博客户端：Weico、推推.....

• “XX开放平台”

•API服务商同时也在收集用户的信息：双向过程



22

数据的获取与处理

- API类型数据的获取
- **百度地图API**
  - http://developer.baidu.com/map/index.php?title=webapi

23

数据的获取与处理

```
http://api.map.baidu.com/place/v2/search?q=饭店
&region=北京
&output=json&page_size=20&page_num=1&ak=E4805d1652
0de693a3fe707cdc962045
```

- 火车采集器
  - API类型数据的获取
  - **百度地图API**
    - Place API

结果字段（以结果集合里的一条数据为例，灰色表示扩展字段）：

名称	类型	说明
name	string	poi名称
location	object	poi经纬度坐标
lat	float	纬度值
lng	float	经度值
address	string	poi地址信息
telephone	string	poi电话信息
uid	string	poi的唯一标示
detail_info	object	poi的扩展信息，仅当scope=2时，显示该字段，不同的poi类型，显示的detail_info字段不同。
distance	int32	距离中心点的距离
type	string	所属分类，如“hotel”，“cate
tag	string	标签
detail_url	string	poi的详情页
price	string	poi商户的价格
shop_hours	string	营业时间
overall_rating	string	总体评分
taste_rating	string	口味评分
service_rating	string	服务评分
environment_rating	string	环境评分
facility_rating	string	星级（设备）评分
hygiene_rating	string	卫生评分
technology_rating	string	技术评分
image_num	string	图片数
groupon_num	int	团购数
discount_num	int	优惠数
comment_num	string	评论数
favorite_num	string	收藏数
checkin_num	string	签到数

24

数据的获取与处理

- 火车采集器
  - API类型数据的获取
  - **百度地图API**
    - Place API

```
http://api.map.baidu.com/place/v2/search?q=饭店
&region=北京
&output=json&page_size=20&page_num=1&ak=E4805d1652
0de693a3fe707cdc962045
```

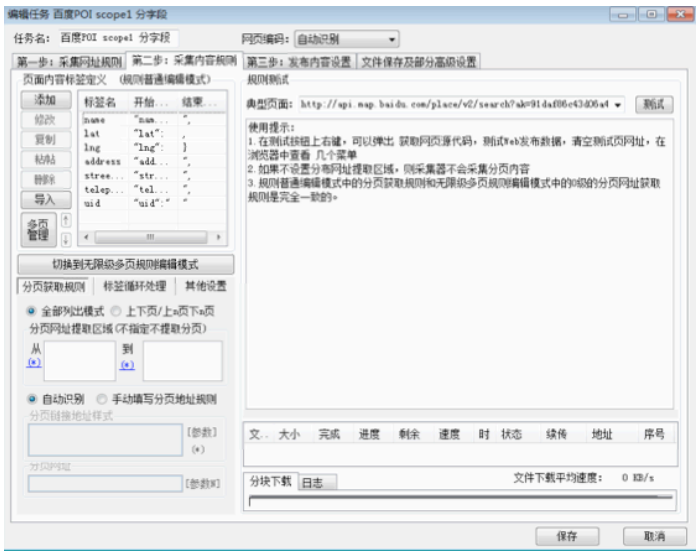

```
{
  "status": 0,
  "message": "ok",
  "total": 15610,
  "results": [
    {
      "name": "北京盘古七星酒店",
      "location": {
        "lat": 39.995074,
        "lng": 116.393207
      },
      "address": "中国北京朝阳区北四环中路27号，盘古大观",
      "street_id": "925c550f80330fab5a3ac832",
      "telephone": "13082483603",
      "detail": 1,
      "uid": "925c550f80330fab5a3ac832"
    },
    {
      "name": "泰山饭店",
      "location": {
        "lat": 40.055374,
        "lng": 116.335822
      },
      "address": "西三旗安宁里北8号(泰山饭店院内南侧)",
```


25

数据的获取与处理

- 火车采集器
  - API类型数据的获取
  - **百度地图API**
    - Place API

```
http://api.map.baidu.com/place/v2/search?q=饭店
&region=北京
&output=json&page_size=20&page_num=1&ak=E4805d1652
0de693a3fe707cdc962045
```

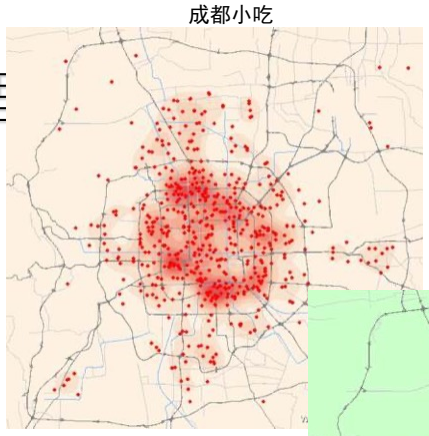



26

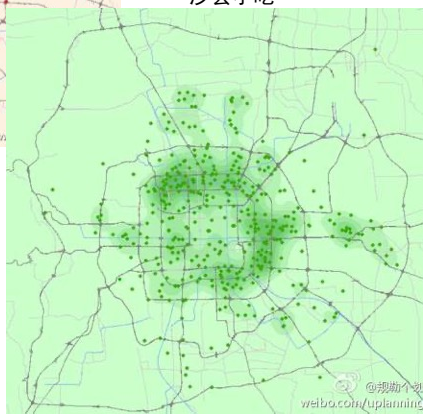
数据的获取与处理

- 火车采集器
  - API类型数据的获取
  - **百度地图API**
    - Place API

成都小吃



沙县小吃

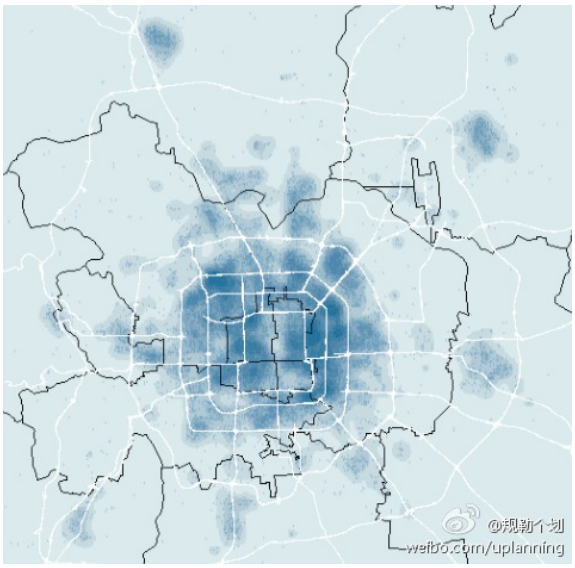


27

数据的获取与处理

- 火车采集器
  - API类型数据的获取
  - **百度地图API**
    - Place API

停车场 ≈20000



28

演示操作demo

1. URL：腾讯疫情新闻专题页
2. 数据源：确认数据接口
3. 构建抓取目标地址列表
4. 明确抓取对象和规则
5. 建立抓取任务
6. Run & pray
7. 输出
8. 清洗、分析、建模，可视化

29

对比

- 软件 vs. 命令行/脚本
  - 运行速度和效率
  - 软硬件成本
  - 软件学习的成本
    - 软件：入门相对容易
    - 命令行/脚本：复用性更高
  - 功能
    - 功能的局限：如火车采集器只支持get的方式，不支持post
    - 命令行/脚本：开源、库（功能）丰富

30

城市众包数据分析框架

1. 分布格局
2. 移动轨迹
3. 语义认知
4. 社会关系

Theme Documents
A Crowd-Sourced Data Based Analytical Framework for Urban Planning

A Crowd-Sourced Data Based Analytical Framework for Urban Planning

Li Dong, Long Ying

Abstract Aimed at the challenges faced by the current urban development and urban planning, along with the research opportunities brought by "big data," this paper proposes an analytical framework based on crowd-sourced data for urban planning by reviewing related literature and practice. The framework is mainly oriented towards three major requirements of analysis in urban planning: the physical spaces, the user communities, and the social relationships. This analytical framework can be regarded as a preliminary attempt for future data-intensive applications in urban planning and assessment.

Keywords location-based services (LBS); crowd-sourced; natural language processing; quantitative urban study

1. Introduction

1.1 Challenges for urban development and urban planning

After 30 years of rapid urban development, China currently has an urbanization rate of more than 50%. Many negative impacts of urbanization, i.e., so-called "urban diseases," are emerging, including traffic congestion, excessive population concentration, heavy consumption of resources, environmental pollution, poor safety and disaster prevention, and so on. Urban diseases are testing the capabilities of urban management and sustainability, with a tendency of spreading out from mega cities to less developed small and medium-sized cities. Since the *1996 Istanbul Declaration on Human Settlements* proposed a statement of "making human settlements safer, healthier, and more livable, equitable, sustainable and productive," the gap between the goal and the reality has become even larger, and the living quality of urban residents has faced serious challenges.

Meanwhile, as a leading player, urban planning itself also faces many threats. Since the conditions and status of cities have changed significantly and are becoming extremely complex, the effectiveness of traditional tools for urban planning such as technical standards and analytical methods has been declining. Planners' abilities of analyzing, diagnosing, and assessing the status of urban development are doubted, let alone their abilities to guide future development, to implement proactive solutions to issues in real world, or to enhance the feasibilities of targets. One of the reasons is the insufficiency of traditional data and analysis. Planners have to extract information based on a relatively small amount of data and try to seek an overall conclusion. In other words, the mode of analysis requires an effective transition from fragmented, low-frequency statistics to a complex overview picture of a city, as well as a shift from a rough aggregation of figures to a finer profiling of individuals (Zhang, 2014).

1.2 The "big data" wave

With the booming of information and communications technology (ICT), incredible amounts of data are produced and available in our cities and on our planet, via various chips, sensor networks, positioning systems, mobile communications, and high-performance computing and storing technologies. Urban daily life, such as transportation and recreation, has also been impacted by the evolving ICT. For example, Baidu.com processes 6 billion search requests every day; over 500 million people talk via WeChat app and compose over 100 billion relationships online; the bus-pass card in Beijing is used up to 20 million times per day, and so on. Human and various types of operation sensors will produce more and more data. According to the estimates from the white paper of the International Data Corporation (IDC) (Gantz and Reinsel, 2009), for every 18 months the volume of new data is equal to the sum of all data in the past, while the total amount of data generated each year will reach 40 ZB by 2020.

To respond to the "big data" wave and further reveal its impacts on cities and human society, the academic community has carried out a considerable amount of research works, represented by two special issues published respectively by *Nature* (2008) and *Science* (2011). Data deposits will be increasing gradually from the level of GB (GigaByte) to PB (PetaByte) and EB (ExaByte), while effective meta-analyses on these data with complex

49

Dong LI, Ying LONG. A Crowd-Sourcing Data Based Analysis Framework for Urban Planning[J]. China City Planning Review. 2015, 24 (1), pp. 49-57.

31

2020

THANK YOU

谢 谢 聆 听

专注于城市的大数据价值

Focusing the value of urban big-data

清华同衡规划播报

同衡创新

32