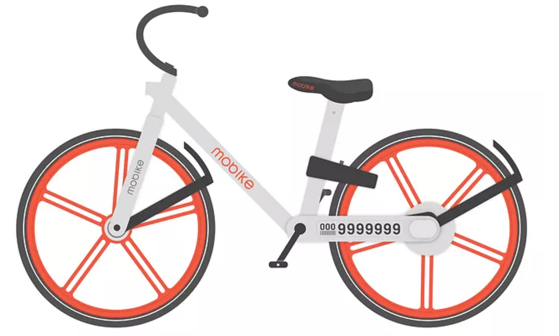
应用机器学习技术

对沪上共享单车的研究

****

****

****

**计算机科学与软件工程学院**

**2017/9/6**

**绪论**

一、研究背景

随着移动互联网的迅速发展，共享单车于2016年火爆起来。根据艾媒咨询数据显示，预计到2017年，中国共享单车用户规模将达2.09亿人。在上海，共享单车投放量已超100万，注册用户已超1300万人。共享单车市场的火爆，再次引发了人们的关注。

共享单车帮助用户解决出行服务最后一公里问题，有效减少小汽车出行次数，符合绿色出行的理念，有助于绿色城市建设。

但随着大规模单车的投放，乱停乱放现象严重，影响交通和市容；故障车辆未及时修理，成为城市垃圾……多种问题日益凸显，共享单车对城市管理提出了新的挑战。

各地政府部门与共享单车企业展开战略合作，并出台相关管理措施。2017年8月18日，上海市交通委对各共享单车企业下达告知书，明确即日起上海将暂停新增投放车辆。并对企业作出相关要求，对上海共享单车市场做出管理。

共享单车在短短几年之间迅速发展，形成巨大市场，在城市经济、城市交通、城市旅游等多个方面产生影响，但其也对城市管理提出新的挑战。虽然目前上海已经叫停投放，但这绝不是唯一的破局思路。

而目前相关管理制度尚不完善，国内尚未有较为深入的研究。自行车共享虽在国外已经发展多年，但大部分是固定桩停放。解决共享单车问题，并不能全盘借鉴国外国外经验，仍需从本国国情出发，由政府、企业、市民三方联手共同解决。

二、研究目的与意义

本研究的目的在于，通过网络爬虫技术爬取上海地区ofo小黄车与摩拜单车的相关数据，如分布地点、数量等，对典型区域进行分析研究，借助数据可视化技术，直观展现上海地区共享单车在时间、空间维度上的特征。

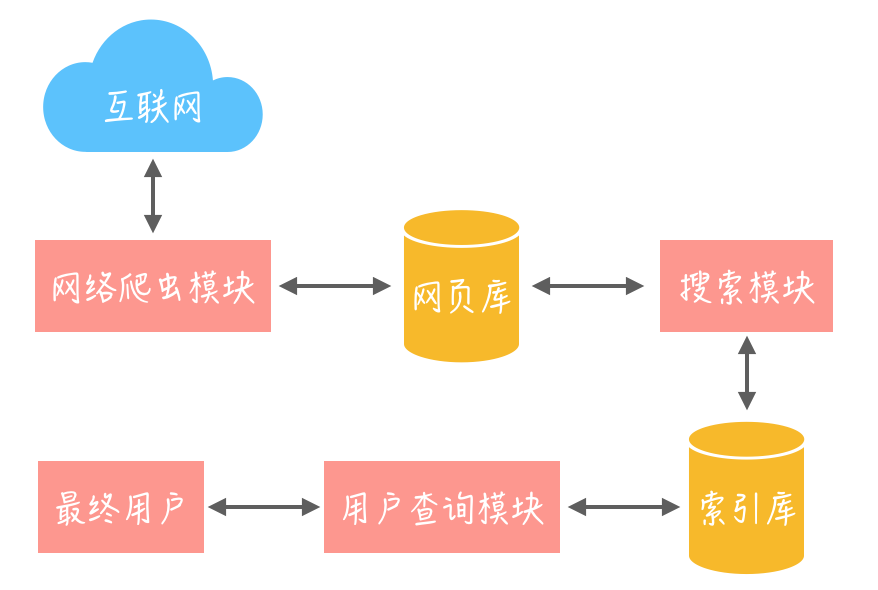
同时我们通过机器学习算法，利用已有共享单车数据中建立模型、获取规律，并通过规律对未来共享单车数据进行预测。

并以问卷调查和访谈法作为辅助手段，分析共享单车对上海地区的影响，助力共享单车行业的良性发展，帮助解决社会热点问题。

研究过程

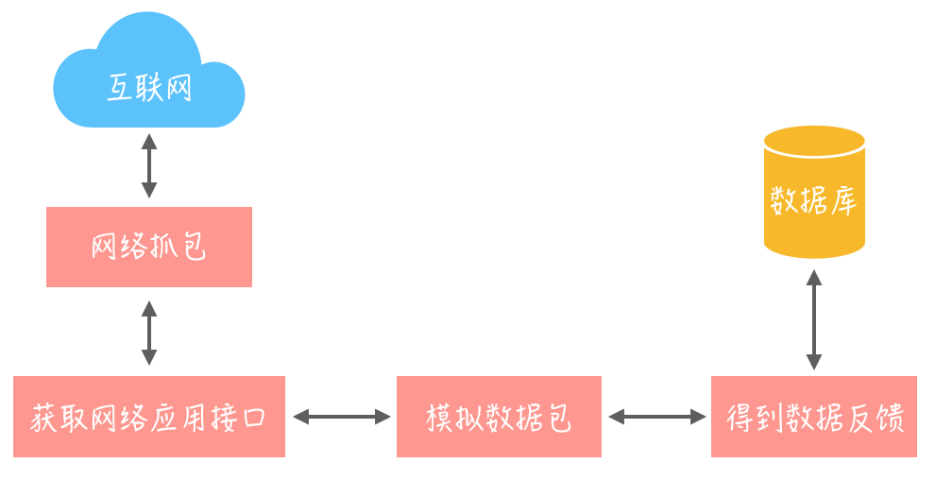
一、数据获取

（一）设计说明



链家网数据抓取原理

网络爬虫（又被称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。



ofo数据抓取原理

通过网络爬虫，可以在节省人力、物力的前提下，更轻松的获取全面、准确的第一手数据，例如，通过网络爬虫可以获取上海地区共享单车的分布、数量等，相比于传统调查方法，数据更加精准。

（二）爬取内容说明

1. 上海地区ofo与摩拜单车的分布情况

了解上海地区，ofo与摩拜单车的整体投放情况，了解投放规模、投放分布

1. 普陀区ofo不同时间点的分布情况

以普陀区为代表区域，了解ofo分布变化情况

1. 上海市部分地点500米内ofo分布情况(陆家嘴、外滩、南京东路、金沙江路地铁站、人民广场、百联中环购物广场、上海西站、上海站等)

了解部分商业区、景点、车站等区域的共享单车具体分布情况、一天内的变化情况，探究共享单车对该区域内的流动，帮助解决供需平衡问题

1. 上海部分地点500米内不同时间点的ofo分布情况（曹杨一村、真光第四小区）

了解居民区附近ofo的分布情况，探究共享单车对该区域的影响，帮助解决区域内供需平衡问题

1. 上海各区域小区户数及房价统计

通过计算各划分区域内小区户数总合，推算夜间居民人口数量，帮助了解区域内供需平衡问题

1. 普陀区ofo到最近地铁站的距离统计

探究地铁与共享单车之间关联

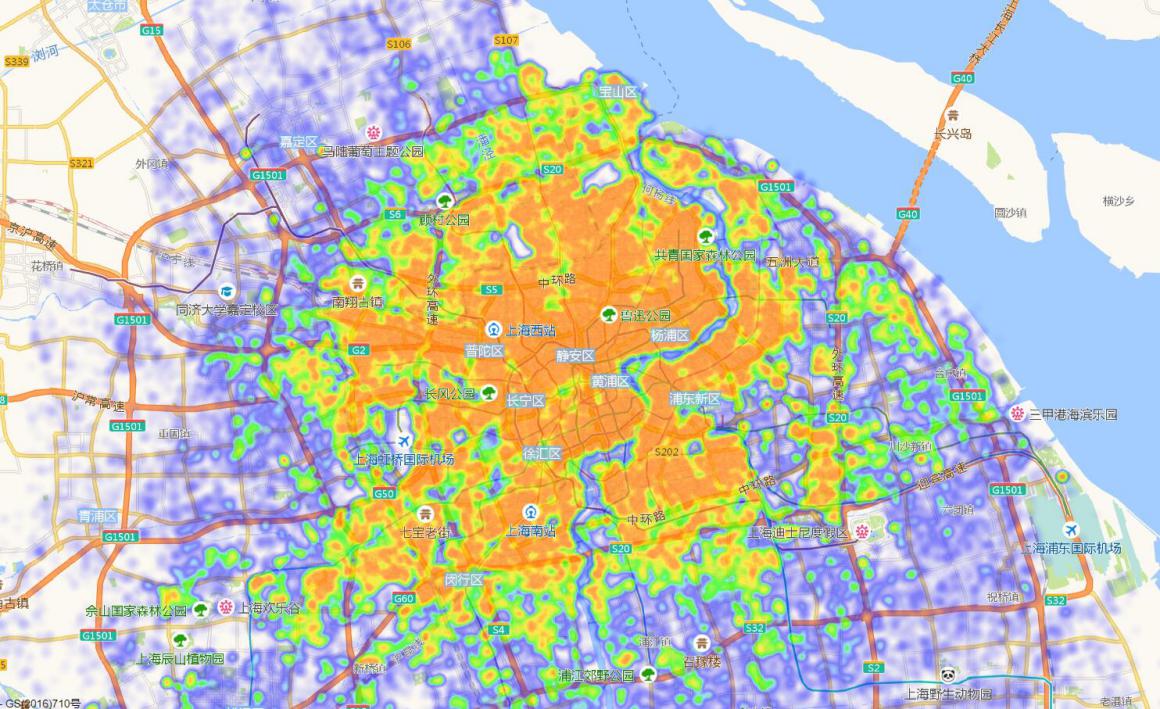
1. ofo骑行时间统计

探究共享单车对人们出行影响

二、数据初探

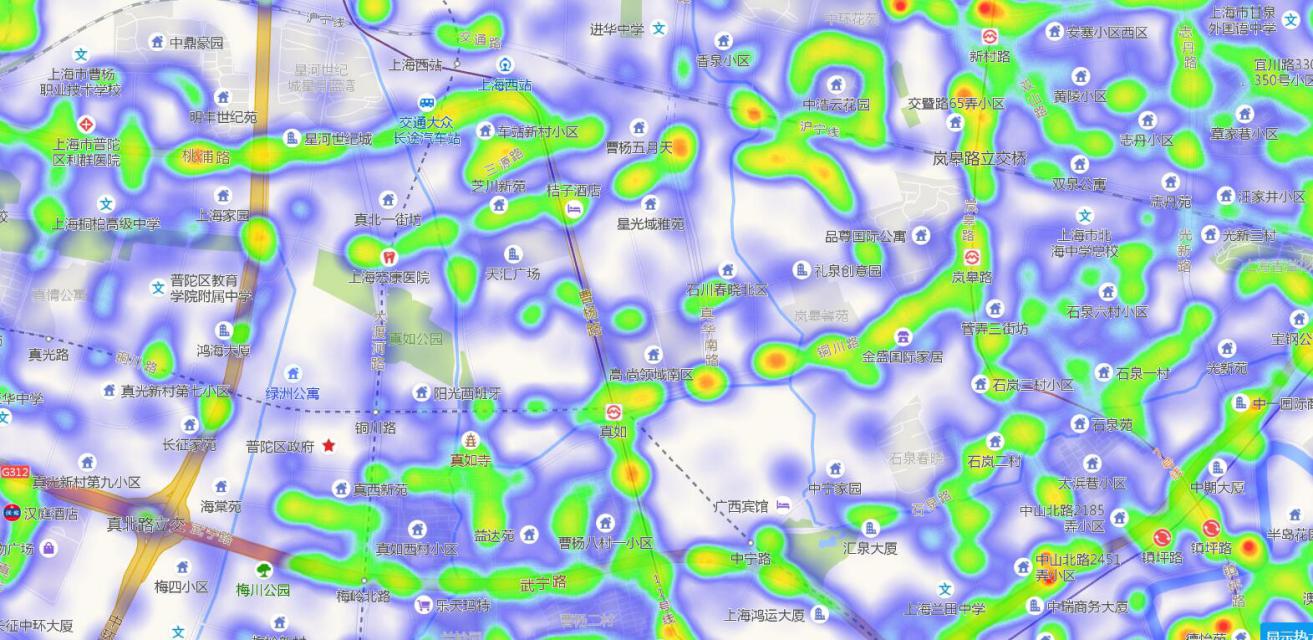
在获取了部分数据后，我们对其中的一部分数据进行了可视化，来初步探索数据的一些的特点。

1. 上海地区ofo与摩拜单车的分布情况



说明：通过不同颜色区块，表现共享单车分布密度；分布呈辐射状，主要集中在中心城区，外部明显减少，边缘呈现棋盘状；沿黄浦江有明显界限

1. 普陀区ofo不同时间点的分布情况



说明：我们将网络爬虫获取到的ofo单车位置数据，绘制在地图上，以热力图形式展示，可以看出，共享单车主要集中在地铁站、干道两侧。热力图中绿色的密集区域，能够清晰的勾勒出部分路段的轮廓。这说明了，在地铁站、路口等地区单车分布较为集中。

1. 普陀区共享单车与小区

我们统计了普陀区夜间3:00点，划分的各个区域内ofo的数量，同时结合之前从链家网抓取的上海市小区的户数、纬度位置信息，我们对各个区域内的供需进行了分析。



夜间区域内小区共享单车供应程度

（衡量标准：共享单车数／小区长驻人口）

凌晨3:00以后，共享单车的移动已经比较少，基本可以刻画第二天一早该区域内的共享单车数量。而通过链家网获取的小区户数，基本可以代表该板块内的夜间常驻人口。通过单车数量与常驻人口数，基本可以刻画出早间的单车供应需求。

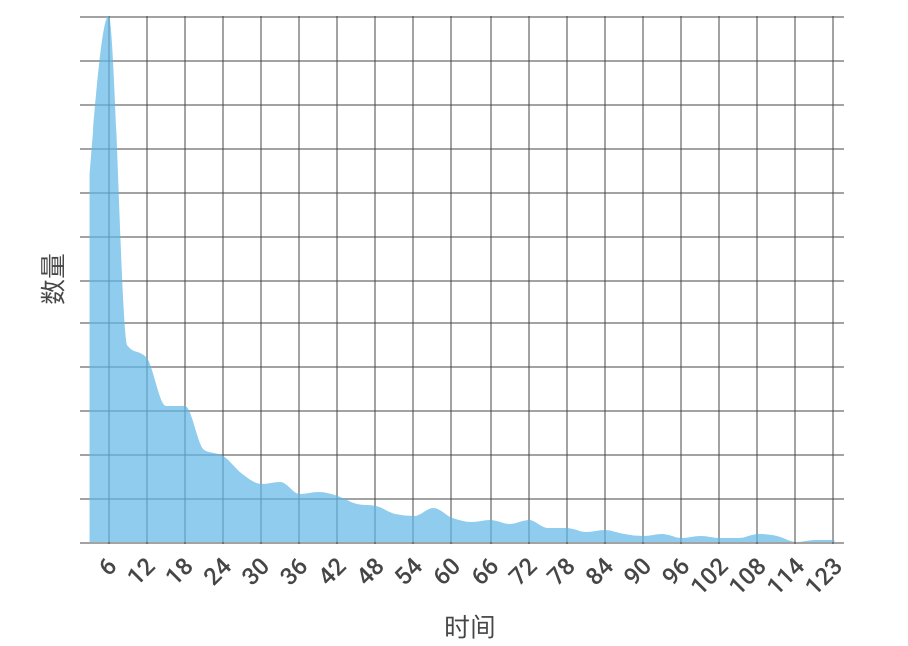
通过图片可以看出，在内环以内的长寿路、江宁路一带供应相对较多，而在外环以外的桃浦新村一带，供应相对较少。而这些区域可能就是未来共享单车企业需要去优化布局的区域。

1. 普陀区ofo到最近地铁站的距离统计

通过网络爬虫爬取到的数据，我们可以看出ofo共享单车大量集中在地铁站附近。在普陀区，59.61%的ofo小黄车分布在地铁1km以内，91.09%的ofo小黄车分布在地铁2km以内。

|  |  |  |
| --- | --- | --- |
| *ofo距离最近地铁的距离* | | |
| *距离地铁范围* | 数量 | 车辆总数百分比 |
| *0~1km* | 39626 | 59.61% |
| *0~2km* | 60551 | 91.09% |
| *0~3km* | 65525 | 98.57% |
| *0~4km* | 66476 | 100.00% |
| *最大距离3699米 平均距离978米* | | |

1. ofo骑行时间统计



根据网络爬虫所得到数据，在上海地区，用户在使用共享单车时，大多数为10分钟以内的短途出行，根据普通人骑无变速功能的自行车的骑行速度推算，距离大约在1.6km以内。

1. 上海市部分地点500米内ofo分布情况(外滩、陆家嘴、南京东路、金沙江路地铁站、人民广场、百联中环购物广场、上海西站、上海站等)

说明：6:30-10:00迅速增长，之后基本稳定，18:30后快速减少

但整体而言，全天车辆保持较大基数

1. 上海部分区点500米内不同时间点的ofo分布情况（曹杨一村、真光第四小区）

在住宅区，7:00之后，共享单车数量迅速减少，而在此时地铁站的单车数量迅速增加，结合调查数据，推断共享单车大量涌入地铁站。下午17:00后，共享单车数量开始缓慢上升，推断小区住户下班回家。但同时，结合数据，曹杨一村全天单车数量保持1000量以上，存在单车过多的现象。

1. 上海各区域小区相关信息统计

通过网络爬虫爬取链家网数据，获取上海市小区的相关信息，通过小区户数、栋数、结构等信息可以大致推测夜间常驻人口数。

可以通过夜间常驻人口数，结合凌晨单车分布，分析各区域在早间单车供需情况。

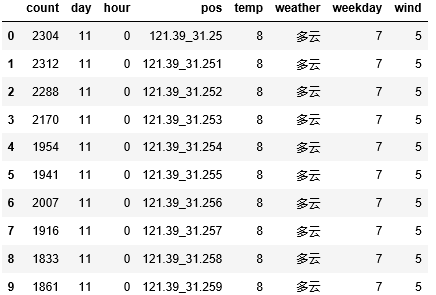


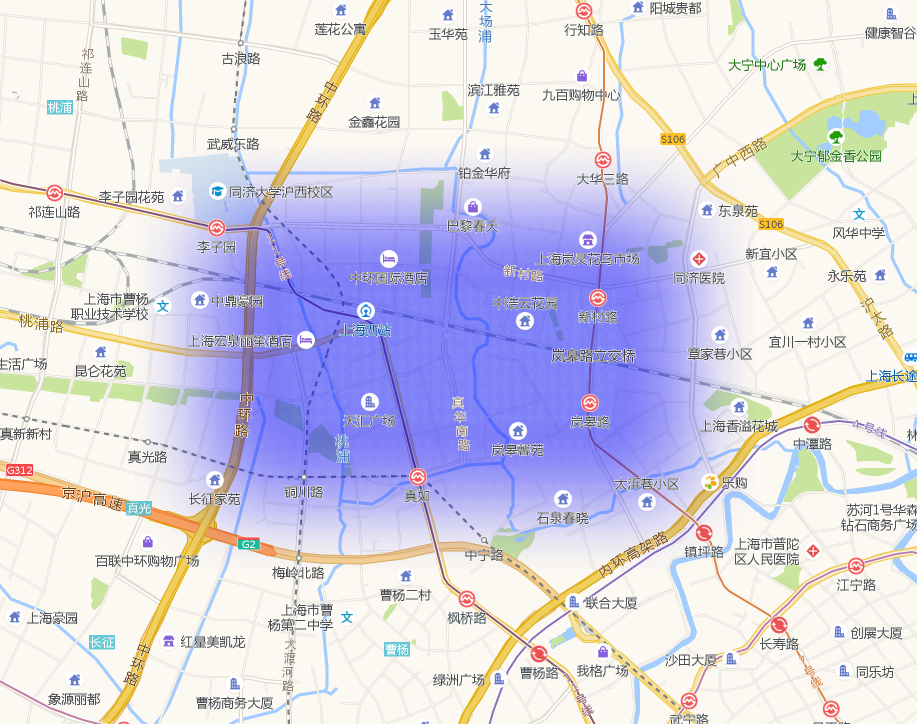
部分数据展示

1. 特征选取

首先考虑建立模型预测车辆的存储量，即预测某个区域在某个时间点的车辆数。

经过数据初探，我们首先选取了以下几个影响较为突出的特征：区域位置、时间点、天气。而时间点特征，又分成日期、整时时刻、星期几。天气也可分成温度、天气状况、风力。



在初步的特征处理中，为了方便处理，我们选择了以上特征中的部分，并且限制了区域位置的可选集合，以及区域的形状和大小，选取了891个相同大小的特定区域（总体区域见下图）。

四、机器学习

首先在初步尝试中，我们选择用线性回归模型来拟合数据。由于时间、位置等特征是离散特征，所以对整时时刻、区域位置、日期、星期几，这几个特征进行了one-hot编码（虽然这样提高了特征的维度，但在初期先不考虑这个问题），从而将数据特征的维度提高到了约900。

在特征处理过后，对两个周的数据（约30万条）进行了拟合，在测试集得到的拟合优度R^2=0.945。