# 一种利用网络爬虫获取商务楼宇和商户信息的方法

A Method of Obtaining Business Building and Merchant Information Based on Web Crawler

张雨龙，孙晓鹏，王晓东（中国联通网络技术研究院，北京 100048）
Zhang Yulong，Sun Xiaopeng，Wang Xiaodong（China Unicom Network Technology Research Institute，Beijing 100048，China）

**摘　要：**
商企市场是电信运营商在宽带接入领域的重要市场，其用户具有明显的个体性和差异性，宽带建设和营销方案均需要商务楼宇信息和商户信息作为基础信息。为了方便运营商准确实时地获取任意区域内的楼宇信息与商户信息，介绍了一种网络爬虫技术方案，该方案具有速度快、准确率高、成本低等特点。

**关键词：**
网络爬虫；商企客户；API；楼宇与商户信息
**doi：**10.12045/j.issn.1007-3043.2019.07.014
**中图分类号：**TN915
**文献标识码：**A
**文章编号：**1007-3043（2019）07-0064-03

**Abstract：**
Business enterprise market is an important market for telecom operators in the field of broadband access. Its users have obvious individuality and difference. Broadband construction and marketing schemes need commercial building and merchants information as basic information. In order to help operators obtain building and merchants information in any area accurately and timely，a network crawler technology scheme is introduced，which has the characteristics of fast speed，high accuracy and low cost.

**Keywords：**
Web crawler；Business enterprise market；API；Building and merchants information

**引用格式：**张雨龙，孙晓鹏，王晓东．一种利用网络爬虫获取商务楼宇和商户信息的方法[J]．邮电设计技术，2019(7)：64-66.

## 0 引言

固网宽带接入市场一直是电信运营商角逐的传统重要阵地。各个电信运营商都在持续加大固网宽带建设力度，投入大量资金。但是随着家庭宽带用户数量接近饱和，“二级”代理商发力占领市场，家庭宽带用户ARPU逐渐降低，电信运营商把投资重点逐步从家庭客户转变为商企客户。

商企客户一般分布在写字楼、工业园区、专业/聚类市场等区域。这类场景的网络覆盖、商业营销与家庭宽带明显不同。特别是写字楼，需要按照楼宇面积、层数、商户数量、商户属性、物业公司、已入驻企业等多个维度进行分级分类的建设和营销。银行等金融类企业、大型连锁公司、创业型小型公司对网络的需求明显不同，具有明显的个体性和差异性。同时，我国经济迅猛发展，商务楼宇信息与商户信息每时每刻都在发生变化。

因此，如何准确实时获取海量的楼宇信息与商户信息是电信运营商当前要解决的重要难题。

## 1 现状分析

目前主要通过号线系统、整理现有信息（台账）和人工摸查3种方法获取楼宇和商户信息。

号线系统：对于固网资源已经覆盖的楼宇，可以通过号线系统导出楼宇和商户信息。一般导出的数据比较准确，但是此方法仅适用于已覆盖固网资源的区域，且时效性较低。

现有信息整理（台账）：各运营商经过多年的规划

收稿日期：2019-05-18

与系统建设，积累了一定数量的楼宇信息，可以直接输出。但这种数据质量一般不高，存在楼宇条目重复、楼宇信息错误、格式不统一等问题，信息时效性差。而且处理海量数据也耗费了大量的人力物力。

人工摸查：这种方法需要相应人员逐片区域、逐个楼宇、逐层楼进行信息摸查，需要消耗大量的人力物力，效率较低。同时人工录入信息格式难以统一，后期还需要花费大量时间处理数据，后续数据更新维护也不方便。

## 2 实现方案

在“互联网+”的大数据信息时代，通过互联网手段可以获得海量的楼宇信息和商户信息数据。网络爬虫作为获取数据的一种新兴方法，具有效率高、成本低、数据时效性高等特点。

通过高德地图/百度地图可以查询到绝大多数楼宇和商户信息。同时由于商业经营等原因，商户会要求地图公司及时更新自己的地图信息。商户信息更新速度快、时效性高。因此，本方案通过高德地图/百度地图提供的接口爬取楼宇和商户信息，然后整理这些信息，利用数学算法，将商户信息匹配到特定的楼宇中，最后输出相匹配的楼宇和商户信息。

本方案中的网络爬虫通过Python语言编写，数据通过MongoDB数据库存储。整体流程如图1所示。

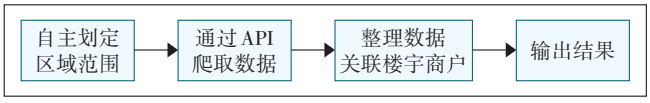


图1　整体流程图

下面结合具体案例介绍方案的实施步骤。本方案的目标是获取“上地大厦”区域内的所有楼宇和商户信息。

### 2.1 自主划定区域范围

在确定楼宇和商户的地理位置后，在地图上选取对应的矩形区域即可（见图2），其中选择的范围（矩形区域大小）没有限制，按需即可。通过高德开放平台，可以获取任一点的经纬度。如图2所示获取并记录红色矩形的左上和右下2个点的经纬坐标。这2个点的经纬度坐标会作为后续爬虫程序的输入信息。

### 2.2 通过API爬取数据

在大数据和人工智能蓬勃发展的时期，为了抢占开发市场和话语权，高德、百度等互联网企业都开放应用程序接口（API），供开发者免费使用。


图2　自主划定区域示意图

为了通过API获取数据，需要向地图公司申请大数据平台权限。首先要注册成为开发者，即用户注册，然后去控制台创建Web服务应用。经过以上步骤，得到API的唯一识别码KEY，该识别码是用户获取数据的权限标识，也是后续爬虫程序的输入信息。

按照API接口的网址要求，将获得的2个经纬度坐标、唯一识别码KEY和其他规定的信息（如商户类型等，高德API接口有分类文档，在官网查询即可）进行拼接，从而得到数据信息的网址，通过该网址，即可得到相应的信息数据。将上述操作过程编写为自动化的爬虫程序，获取数据并将返回的信息数据（即获取的楼宇和商户信息）存储进MongoDB数据库。

### 2.3 数据的整理与关联

上一个步骤输出的信息数据中，楼宇信息和商户信息是独立的，没有形成完备的数据集合，所以需要对输出的数据进行整理和归类。

整理：根据建设和营销的需求，选取有效字段（一般包括名称、类型、地址、经纬度、边框经纬度、电话、网址等信息），具体操作如图3所示。


```
⊞“0”:{…},
⊟“1”:{
    “id”:“BOFFHCVIZ”,
    “name”:“北京胶啤堂商务中心”,
    “tag”:[],
    “type”:“商务住宅;商务住宅相关;商务住宅相关”,
    “typecode”:“120000”,
    “biz_type”:[],
    “address”:“裕民大街3号院1栋1层北京方糖底商A1010号”,
    “location”:“116.557259,40.103464”,
    “tel”:“15011036331”,
    “postcode”:[],
    “website”:[],
    “email”:[],
```


图3　API输出数据示意图

归类：楼宇和商户信息均有经纬度，其中楼宇信息包含区域边框顶点经纬度。通过楼宇的边框顶点经纬度信息和商户的经纬度信息，可以判断商户的经

纬度点是否在楼宇的边框区域内。如商户的经纬度在楼宇的边框区域中，那么就把该商户匹配到对应楼宇中，以实现商户与楼宇的关联。

### 2.4 输出结果

通过上述步骤，最终输出楼宇信息表和商户信息表。表1是楼宇信息表，主要包含名称、地址、经纬度、包含的商户数量、商户名称、电话、网站、所属城市、所属区域等信息。其中“商户名称”一行应包括所有商户的名称，本表只是选取其中4个作为示例。表2是输出的楼宇信息情况，每一行为1个楼宇。

表1 楼宇信息示意表

| | |
|---|---|
| 名称 | 金隅嘉华大厦 |
| 地址 | 上地三街9号 |
| 纬度 | 40.036828 |
| 经度 | 116.307884 |
| 边框信息 | 116.308848、40.038072，116.309605、40.036255，116.309626、40.036314，116.308848、40.038072 |
| 商户数量 | 358 |
| 商户名称 | 鸿合科技、北京博创理想科技有限公司、天地阳光通信科技(北京)有限公司、北京东方正龙数字技术有限公司等 |

表2 输出数据示意表

| name | address | lat | lng | tel | pcode | pname | adname | business | Vectorgraph | ShopsName | ShopsCounts |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 金隅嘉华大厦 | 上地三街9号 | 40.03683 | 116.3079 | 010-62961××× | 110000 | 北京市 | 海淀区 | 上地 | 116.30884 | ['鸿合科技 | 242 |
| 中关村创业大厦 | 上地信息路26号 | 40.03712 | 116.3112 | 010-82898××× | 110000 | 北京市 | 海淀区 | 上地 | 116.31035 | ['北京鼎盛 | 35 |
| 上地大厦 | 信息路30号 | 40.03348 | 116.3122 | 010-62971××× | 110000 | 北京市 | 海淀区 | 上地 | 116.31259 | ['微宝科技 | 8 |

### 2.5 方案准确率分析

经过现场抽样摸查核实，抽取区域内5%的楼宇作为样本。经测算，楼宇信息准确率为100%，商户信息准确率为80%左右。商户信息出现错误的原因多为地图公司更新不及时、中小企业破产或商户变更地址后没有及时通知地图公司。

## 3 结束语

利用互联网化的技术手段来获取高质量的用户信息以拓展用户市场，是运营商互联网化运营的重要组成部分。本文所提到的方案在实际应用中可能面临以下的问题。

a）缺乏专业技术人员。掌握网络爬虫技术需要一定的专业知识，运营商的传统业务人员不能满足技术要求，需要组建专门的团队进行该工作。

b）大数据处理问题。面对海量数据，EXCEL台账等传统工具已不适用。如何从不同维度对海量数据进行分析并使其适用于运营商的业务发展，是下一步工作的关键。

c）数据共享与更新机制问题。网络爬虫获取的数据与工程核实确认的数据如何相互补充共享、如何更新是也是运营商需要解决的问题。

面对上述问题，笔者有以下几点建议。

a）推进大数据和互联网技术的应用。在信息资源时代，电信运营商应充分利用大数据和互联网技术，摸清现状，精准建设，精准发力，开拓市场，抓住战略机遇，与互联网企业合作利用其技术优势，实现资源收益最大化。

b）加快互联网化运营转型。在宽带专业运用大数据和互联网方法，为市场前端业务开展提供支撑。在后续的网络建设中用数据说话，转变思路，加快互联网化转型。

c）提升工作效率，接轨大数据。运营商应该摒弃以往人工费时费力的方法，引入网络爬虫，通过互联网的公开信息，按需获取信息数据，提升工作效率。

**参考文献：**

[1] 王振，张志敏，王伟，等. 基于百度API开源数据的居民出行研究[J]. 交通运输研究，2018(3)：18-24.
[2] 于娟，刘强. 主题网络爬虫研究综述[J]. 计算机工程与科学，2015，37(2)：231-237.
[3] 吉建培，葛娟，韩文立，等. 互联网地图服务质量模型初探[J]. 测绘通报，2016(3)：94-97.
[4] 戚利娜，刘建东. 基于Python的简单网络爬虫的实现[J]. 电脑编程技巧与维护，2017(8)：72-73.
[5] 唐琳，董依萌，何天宇. 基于Python的网络爬虫技术的关键性问题探索[J]. 电子世界，2018，548(14)：34-35.
[6] 张雨龙，王晓东，李洪栋. 宽带接入网10G PON技术发展及部署研究[J]. 邮电设计技术，2018(2)：70-73.

**作者简介：**
张雨龙，毕业于北京邮电大学，工程师，硕士，主要从事接入网规划设计工作；孙晓鹏，毕业于天津大学，工程师，硕士，主要从事投资评价工作；王晓东，毕业于北京理工大学，高级工程师，硕士，主要从事接入网规划设计工作。