

# 交通事故位置信息与经纬度坐标转换研究

张 岚<sup>1,2</sup>, 王 睿<sup>1,2</sup>

(1.交通运输部公路科学研究院, 北京 100088; 2.北京中交华安科技有限公司, 北京 100088)

**摘要:** 交通事故记录的位置信息是事故分析研究的重要基础。中国大量的道路交通事故数据以记录发生地的地址或者明显标志物的方式描述事故位置信息, 无法批量处理并直接用于数据分析, 而人工标定过程烦琐、效率较低、成本较高。通过算法对道路交通事故记录的位置信息进行切分、匹配, 并按照描述标定到指定位置, 达到自动转换的效果。该方法可实现交通事故记录位置标定的自动化, 提高数据处理的效率, 降低成本, 为道路交通事故信息的空间分布研究提供数据支撑, 也为海量处理数据提供可能。

**关键词:** 交通安全; 转换; 算法; 位置信息; 交通事故; 经纬度

Location Information and Longitude and Latitude Coordinate Transformation for Traffic Accidents

Zhang Lan<sup>1,2</sup>, Wang Rui<sup>1,2</sup>

(1.Research Institute of Highway Ministry of Transport, Beijing 100088, China; 2.Beijing Zhongjiaohuaan Technology Co., Ltd., Beijing 100088, China)

**Abstract:** The location information of traffic accident records is an important basis for accidents analysis and research. A large number of traffic accident data describe their location information by recording the location of accident or obvious signs, which cannot be processed in batches and directly used in data analysis. However, the manual calibration process is cumbersome, inefficient and costly. Through the algorithm, the location information of road traffic accidents records is segmented and matched, and calibrated to the designated location according to the description, so as to achieve automatic conversion. The method can achieve the automatic calibration of traffic accident record location, improve data processing efficiency, reduce the cost, provide data support for the spatial distribution of road traffic accident information, and also provide powerful support for processing massive data.

**Keywords:** traffic safety; transformation; algorithm; location information; traffic accidents; latitude and longitude

收稿日期: 2019-01-04

作者简介: 张岚(1980—), 女, 河北保定人, 硕士, 高级工程师, 主要研究方向: 交通安全。

E-mail: 25192162@qq.com

## 0 引言

随着交通事业的快速发展, 道路交通事故也日益增多。据世界卫生组织(WHO)2018年12月7日发布的报告, 全球每年约有135万人死于交通事故, 平均每24 s就有1人因交通事故丧生。交通事故的伤害不仅给个人和家庭带来无法弥补的直接损失, 也给社会带来巨大的经济损失, 严重制约了可持续发展。对交通事故进行分析是人类采取措施预防事故的重要前提, 而道路交通事故记录中的位置信息是事故分析、风险评估等研究的

重要数据基础。

目前大量的交通事故记录, 尤其是历史记录, 通常是以记录地址或者明显标志物的方式描述其空间信息: “寺下村口” “美好家园桥头” “百货大楼路口” 等类似信息频繁出现在历史事故记录中, 很难批量处理, 无法直接用于数据分析, 需要熟悉区域地理环境的交警或工作人员人工逐条标定, 既耗费大量时间, 也浪费大量人力, 成本较高, 给道路交通事故研究带来诸多不便。因此, 如果能够通过计算机将这些海量信息进行自动转换, 最终实现交通事故记录位置信息标定

的自动化,就能大大提高工作效率,为道路交通事故信息的空间分布等研究提供强大的数据支撑。

道路交通安全研究旨在减少事故发生的概率以及降低事故发生时的伤害。多年的交通事故数据积累是重要基础,依据其原因、类型、空间和时间分布来研究判断各种交通管理政策的适用性,以及交通安全设施设置的合理性。交通事故研究中的位置、区域等要素与空间信息紧密相关,可根据事故数据的经纬度等空间属性,把事故数据布设到地图中,使其分布和类型得以展现,实现可视化。有了这种可视化的信息,更容易发现交通事故与道路线形及周边环境的关系,总结规律。有了准确的交通事故位置信息,研究人员还可以根据事故空间信息有针对性地对部分数据进行提取,如交叉口发生的事故、弯道发生的事故等。因此,道路交通事故记录的位置信息与经纬度坐标转换研究非常必要。

根据公安部的统计数据,仅2016年,全国共接报交通事故864.3万起<sup>[1]</sup>。虽然大部分已经数字化,但记录时采用不同的方式,导致开展交通事故研究前仍旧需要整理,而且大量历史数据难以回溯,若逐条人工添加位置信息,工作效率和准确性都较低,难以大规模开展。因此,如果能自动把事故记录的位置信息转换为空间信息,将为以交通事故为基础的研究提供巨大帮助。

## 1 研究现状综述

美国人口调查局于1990年前后首次发布了与人口数据相关的全国地理编码系统:图形整合地理编码参照系统(Topologically Integrated Geographic Encoding and Referencing, TIGER)。此系统通过地理编码将地址信息定位到区域或街道,其定位原理是将地址信息的街道名记录与数据库中的街道名进行匹配,确定街道位置,然后利用门牌号确定其具体位置。美国加州大学伯克利分校以加利福尼亚州为对象,进行了大量关于交通事故地理编码的研究,交通事故地点的描述以路名、桩号或道路交叉口为原点加对应方向与距离的方式转换为经纬度方式,并开发了加利福尼亚州全域交通编码系统(California Statewide Integrated Traffic Records System, SWITRS)。澳大利亚国家地址编码数据文件

(A Geocoded National Address File for Australia)将地址分为区域、街道和宗地(门牌号)三级,并以所在地域范围的几何中心作为基本位置进行地理编码<sup>[2]</sup>。

国外已经开展的较多地理编码方面的研究主要用于政府管理,为处理海量规则化的数据提供了巨大帮助。但由于各国之间地址描述的方式和习惯不同,国外系统只能作为参考,很难直接使用。必须对实际地址的描述规则进行研究,开发适合本国地理编码规则的算法与系统。

由于中国地址设置变化频繁,没有统一的系统性规则,所以关于地理编码的研究成果较少,主要集中在理论方面,如文献[3]对于地理编码及相关技术进行概述。文献[4]在地址匹配方面进行论述,文献[5]在中文地址名词匹配方面的研究也取得了一定成果。但是这些研究对于中文地名没有统一的规范和固定的模式等问题都没有很好的解决。文献[6]指出,中国现有的地址体系比较复杂,给地理编码应用带来了一定的难度,导致地址数据库的通用性较差,地理编码应用限定在局部地区和部门内部。总之,在国内专门针对交通事故记录位置的地理编码研究几乎没有。

## 2 道路交通事故记录的位置信息与经纬度坐标转换

针对现有道路交通事故记录中位置信息多用描述性语句记录、人工逐条定位工作量大的问题,根据地理编码的原理,设计位置描述与经纬度转换算法,将道路交通事故记录中的位置信息自动转换为经纬度坐标。

### 2.1 事故地点描述分析

根据原始事故数据可以看出,事故数据的地点描述与一般地点描述不同,普通的地点描述多用于描述地标,如一幢大厦,一个小区,格式具体到门牌号。而对于事故地点的描述,除使用地标描述外,通常使用道路名称描述。

事故数据记录主要有以下几种方式:

1) 使用桩号描述。桩号是描述公路位置的最好方式。此类描述使用道路桩号,如“北京市通州区永乐店镇孔兴路1.4 km处”,由于交警描述习惯不同,可能为k1+400、1公里400米处、1公里+400米处等,但大多

特征比较明显,有关键字“公里”“+”等。由于具体道路可能有国家、地市级等多套桩号描述方式,需要根据事故数据所属路段桩号范围区分。

2) 使用电线杆编号描述。接近市区的道路,其桩号标记较少,交警有时会使用电线杆号作为编码标准,如“北京市延庆区刁五路许家营村850号电线杆处”。通过电线杆或者灯杆位置可以找到事故发生的位置。此类地标是沿路排列的,有一定规律性,仅需要少量信息就可以排列。

3) 使用道路交叉口描述。当事故发生于道路交叉口或附近路段时,通常用两条道路的路名描述,如“密云区密兴路丰各庄路交叉口”。此时通过两条道路的位置可以获得事故位置。

4) 使用明显地标描述,如“房山区闫东路中国原子能科学研究院生活区门前”。通过地标位置及地标与道路的关系可以计算出事故位置,但有时地标描述有歧义或者有多个位置符合此地标描述,所以需要结合道路所属路段,否则将会得到多个结果。

5) 使用地标+道路交叉口描述,如“梅兰芳大剧院交叉口”。使用地标加附近的道路交叉口位置时,应通过地标位置计算出附近最近的道路交叉口。

6) 使用道路+通向地描述,如“密云区密兴路丰各庄路口南”。

7) 只有路名。只使用一条路名记录,此时事故位置难以判断,需要提出问题,以待进一步处理。

## 2.2 事故地点描述规范化调整

### 2.2.1 位置描述规范

一条描述记录通常包括几个部分:所属地市、所属区县、道路名称和详细位置。但在实际位置描述中,往往会出现信息缺省、信息冗余和口语化描述:所属地市与所属区县的缺省,如“右堤路东大桥环岛北侧”;详细信息与道路名称的口语化描述,如“门头沟区城子路口以南处”。因此,对于位置描述,通常应做以下处理:

1) 根据资料来源补全缺省的信息,如“右堤路东大桥环岛北侧”是来自北京市顺义区的事故资料,应补全为“北京市顺义区右堤路东大桥环岛北侧”;

2) 找出位置描述的关键信息,通过关

键字“市”“区”“县”“镇”等区分所属地市,通过关键字“路”区分道路名称;

3) 尤其应注意将道路名称与详细信息切分成一个一个的词语,如“门头沟区城子路口以南处”,切分后的位置信息应为“北京市”“门头沟区”“城子路”“路口”“路口以南”。

### 2.2.2 逆地理编码

逆地理编码的作用是通过位置描述获得经纬度,常用的地理编码包 geopy 和各地图 API 都有类似的功能,可以通过路名得到经纬度坐标。

由于实际道路通常为线状结构,每个路名对应一条道路,即对应很多个经纬度坐标。因此,本文的逆地理编码部分没有使用常见的地理编码包,而是专门搭建了道路信息数据库及相应逆地理编码模块。

### 2.2.3 详细位置解析

获得道路信息后,需要分类对详细位置进行解析:1)使用桩号描述的地址最为简单,在道路信息数据库中找到道路的方向和起点桩号即可进行计算;2)使用电线杆号进行描述的地址需要数据库中至少包括2个电线杆号的位置,才能根据号码进行计算;3)对于使用道路交叉口位置进行描述的,应根据道路信息计算交叉口位置,但对于只使用一条路名标记的道路交叉口(如“长安街交叉口”),因为描述比较模糊,从起点到终点都可能为事故发生的位置,此时将输出多个位置,必须补充其他信息;4)对使用明显地标进行描述的,通过路名获得道路信息,使用常用地理编码包 geopy 获取地标位置,然后投影到道路上进行计算。

详细位置在上述四种信息及其组合外的,还会附加方位信息,如“某某路口以南”“某某小区以东”等。这种情况需要根据道路走向,将计算出的具体位置进行移动。一般情况下,这种补充的方位描述距离具体位置并不太远,默认为100~200 m的位置移动即可。

根据上述理论,使用python编程实现事故数据处理自动化。通过计算机语言对提供的事故描述信息进行挖掘分析,获取有效的位置相关描述关键词和关键句。关键信息种类包括:省、市、县、道路名称、道路编号、方向、事故发生桩号等。

### 2.3 建立道路信息数据库

目前能获取的地理位置信息数据源有两种：高德POI数据和道路信息采集数据。

#### 2.3.1 高德POI数据

##### 1) 数据源分析。

高德POI又名高德兴趣点，来源于高德地图。数据源结构为分类结构，按照省市县分类。每个POI信息点包含以下信息：所属省份，所属城市，所属区县，详细地址，电话号码，所属分类(一个集合)，大地坐标，火星坐标，百度坐标。

##### 2) 数据源获取。

高德地图没有提供直接的POI数据，网上购买的数据也不能保证是最新的。为了获取最新的全国POI数据，采用直接从高德地图发布的POI数据列表里爬取。数据源展示网页为静态列表分层数据，可借用第三方数据采集工具。本文采用了比较流行的火车头采集器。

##### 3) 数据库建立。

通过火车头采集器采集到数据，经过配置存储到MySQL5.0数据库。POI位置信息表数据结构如表1所示。

#### 2.3.2 道路信息采集数据

##### 1) 数据源分析。

为了满足数据分析的需要，研究人员有时需要自行采集道路信息数据，此种方式得到的数据一般具有比较准确的位置信息，采集的原始数据包含：道路经纬度数据、道路设施数据、道路桩号数据、道路视频数据，采集的数据通过专业化软件处理成标准的数据结构，数据项有：省、市、县、道路名称、道路编号、方向、桩号、经度、纬度、采集时间、采集图片。例如，在道路安全风险评估项目实施过程中，会对评估对象进行调研，获得全部道路的位置信息，通过计算获得道路信息数据并写入道路信息数据库。

##### 2) 数据库建立。

采集的数据通过程序导入到数据库，表结构如表2所示。

### 2.4 实现交通事故位置匹配算法

建立道路信息数据库后，即可通过计算产生地区字典和道路字典，从而更容易获得较准确的地址关键信息部分结果。

地址分词库jieba为python中常用的中文分词模块，但在针对地址的解析中，对路

名的解析效果并不尽如人意。所以，必须专门针对地址进行解析，将“路”“大街”“路口”切分出来，如“密云区密兴路丰各庄路口”被切分成“密云区，密兴路，丰各庄，路口”，然后通过词典被正确切分为“北京市，密云区，密兴路，丰各庄路，路口”。

获得地址切分结果后即可进行逆地理编码。可以通过查找密兴路与丰各庄路相关的

表1 POI位置信息

Tab.1 POI location information

字段名	主键	数据类型	长度	允许空	字段描述
ID	是	bigint		否	序列号
Gengxsj		varchar	50		更新时间
Sheng		varchar	50		省
Shi		varchar	50		市
Xian		varchar	50		县
Xiangxdz		varchar	200		详细地址
Dianhnm		varchar	50		电话号码
Suosfl		varchar	100		所属分类
Dadzb		varchar	50		大地坐标
huoxzb		varchar	50		火星坐标
Baidzb		varchar	50		百度坐标
Gengxbz		varchar	2		更新标志
Shifsc		varchar	2		是否删除
shujydz		varchar	100		数据源地址

表2 采集数据的位置信息

Tab.2 Location information of collected data

字段名	主键	数据类型	长度	允许空	字段描述
ID	是	bigint		否	序列号
Gengxsj		varchar	50		更新时间
Sheng		varchar	50		省
Shi		varchar	50		市
Xian		varchar	50		县
Daolbh		varchar	200		道路编号
Daolmc		varchar	50		道路名称
Fangx		varchar	100		方向
zhuangh		varchar	50		桩号
Jingd		varchar	50		经度
Weid		varchar	50		纬度
Gengxbz		varchar	2		更新标志
Shifsc		varchar	2		是否删除

交叉口, 获得该交叉口的 ID 及经纬度, 即事故发生的准确位置。

实现交通事故位置信息与经纬度坐标转换的算法分为 4 步: 1) 对事故地点描述进行规范化调整, 提取交通事故描述的位置信息, 建立交通事故位置描述关键信息提取数据库; 2) 研究高德 POI 数据和道路信息采集数据, 建立道路信息数据库; 3) 根据事故位置描述关键词在道路信息数据库中查找位置, 实现位置匹配; 4) 输出最终计算结果。

部分地址无法获得准确位置, 有如下几种情况: 1) 当发现道路信息在数据库中不存在时, 程序会输出此类型错误, 并显示查找不到的路名, 可能是因为事故记录时存在笔误或者使用了道路别名; 2) 当发现位置信息不足时, 程序会输出此类型错误, 并显示位置信息不足, 可能是因为当时事故记录中位置信息模糊。

## 2.5 关键问题解决

1) 针对位置信息的词语切分和映射算法的普适性和准确性。

由于位置信息使用自然语言描述, 所以词语切分方式的不同可能会产生歧义。通过分析历年的交通事故数据, 使算法尽量全面, 并且具有基本的学习能力, 能够通过训练样本更好地完成关键信息切分, 尽量减少词语切分时产生的歧义。

在获得关键词后, 由于事故记录有可能使用该地点的旧称、别称或简称, 又或者在地图中有相似的地名, 可能会产生歧义。因此在算法中, 对于地点的不同称呼进行了检查, 对于会产生歧义的地名, 需要考虑设定权重, 并按照权重列出所有可能的选择, 方便用户进行选择。

2) 位置标定算法的准确性。

在事故位置信息中存在一些定性描述位置的词语, 如“门口”“附近”。对于使用有非正向道路(如正南正北)的交叉口作为关键词的位置, 方向词的描述也可能产生歧义。通过分析历年的交通事故数据, 对于这些定性的词语如何定位, 在算法设计中也进行了合适的定义, 例如“附近”默认为 100 m 的位置移动。

## 3 结语

道路交通事故位置信息与经纬度坐标转

换的研究成果可以改变人工逐条处理道路交通事故记录位置信息的现状, 提高数据处理的效率, 降低成本, 为道路安全空间分析等相关研究提供数据处理服务。

参考文献:

References:

- [1] 公安部交通管理局. 中华人民共和国道路交通事故统计年报(2016年度)[M]. 无锡: 公安部交通管理科学研究所, 2017.
- [2] Bigham J M, Husby H. A Review of the Spatial Accuracy of FARS Coordinate Locations in California[C]//Poster Presentation for the American Public Health Association (APHA) 2010 Annual Meeting, Denver, Nov 6-10, 2010: Apha Meeting & Exposition.
- [3] 张晓光. 地理编码和比对技术[J]. 知识经济, 2010(13): 112-113.
- [4] 胡青, 徐建华, 王志海. GIS 数据库中地址自动匹配方法研究[J]. 测绘与空间地理信息, 2008, 31(6): 50-52.  
Hu Qing, Xu Jianhua, Wang Zhihai. Research on Automatic Address Matching Method in GIS Database[J]. Surveying and Spatial Geography Information, 2008, 31(6): 50-52.
- [5] 程昌秀, 于滨. 一种基于规则的模糊中文地址分词匹配方法[J]. 地理与地理信息科学, 2011, 27(3): 26-29.  
Cheng Changxiu, Yu Bin. A Rule-Based Fuzzy Chinese Address Segmentation Matching Method[J]. Geography and Geo-Information Science, 2011, 27(3): 26-29.
- [6] 高小平, 兀伟. 地理编码技术探讨[J]. 测绘标准化, 2012, 28(4): 1-3.  
Gao Xiaoping, Wu Wei. Discussion on Geocoding Technology[J]. Standardization of Surveying and Mapping, 2012, 28(4): 1-3.

## 更正

本刊 2020 年第 1 期 36-43 页《市域(郊)铁路车辆选型研究》一文, 表 3 中  $200 \text{ km} \cdot \text{h}^{-1}$  速度等级对应的最小减速距离应为 1.71 km, 特此更正。