

众包时空数据驱动的城市地理信息推测综述

阮思捷¹, 熊可钦¹, 王树良¹, 耿 晶¹, 鲍 捷^{2,3}, 郑 宇^{2,3}

(1. 北京理工大学计算机学院, 北京 100081; 2. 京东商城(北京)数字科技有限公司, 北京 100176;
3. 京东智能城市研究院, 北京 100176)

摘要: 对地理信息的准确掌握是城市中各种智能决策得以实现的基础. 传统地理信息收集主要靠人工测绘、人工巡检或固定传感器感知, 设备、人力成本高昂. 近年来, 随着移动互联网的发展, 泛在的移动群体在城市中产生了海量的时空数据, 他们有意或无意间成为城市的传感器, 使研究人员有机会利用众包的思路基于此类数据推测城市地理信息. 基于众包时空数据推测城市地理信息具有成本低、空间覆盖广、更新及时等优点. 但其具有严重的数据质量问题, 对城市地理信息推测带来了巨大挑战. 本文综述了根据轨迹、基于位置的社交网络、街景等众包时空数据, 推测城市中以路段、兴趣点、兴趣面为代表的地理实体的位置和属性的方法. 本文给出了众包时空数据和地理实体的定义, 详细比较了众包时空数据驱动的推测方法与传统方法的优劣, 说明了研究问题和挑战, 然后讨论了地图匹配、名称提取、位置发现和统计属性推测四个研究问题的研究进展, 最后展望了该领域未来的研究方向.

关键词: 自发地理信息; 众包时空数据挖掘; 低质量数据处理; 机会式感知; 城市计算

基金项目: 国家自然科学基金(No.62076027, No.42201461); 北京理工大学青年教师学术启动计划(No.6120220113)

中图分类号: TP391

文献标识码: A

文章编号: 0372-2112(2023)08-2238-22

电子学报 URL: <http://www.ejournal.org.cn>

DOI: 10.12263/DZXB.20230131

A Survey of Urban Geographic Information Inference Driven by Crowd-Sourced Spatio-Temporal Data

RUAN Si-jie¹, XIONG Ke-qin¹, WANG Shu-liang¹, GENG Jing¹, BAO Jie^{2,3}, ZHENG Yu^{2,3}

(1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China;

2. JD iCity, JD Technology, Beijing 100176, China;

3. JD Intelligent Cities Research, Beijing 100176, China)

Abstract: Knowing the accurate geographic information is the basis to achieve the intelligent decisions in cities. Traditional geographic information collecting mainly relies on manual mapping, patrolling or sensing by static geographical sensors, which are expensive given specialized equipment and labors. Recently, with the development of the mobile Internet, the ubiquitous moving objects has generated massive spatio-temporal data in the urban spaces, who act as sensors of the city consciously or unconsciously, and make it possible to infer the geographic information based on those data in a crowd-sourced manner. The geographic information inference based on the crowd-sourced spatio-temporal data enjoys the advantages of low cost, high spatial coverage, and timely updates. However, it also has the data quality issues, which introduce great challenges to the urban geographic information inference. In this paper, we survey the location and attribute inference of geospatial entities, including the road network, point of interest and area of interest, based on crowd-sourced spatio-temporal data, e.g., trajectories, location-based social network, and street views. We give the definitions of crowd-sourced spatio-temporal data and geospatial entities, compare the pros and cons of utilizing the crowd-sourced spatio-temporal data to infer the geographic information against traditional methods, and elaborate the research problems and challenges. After that, we review four research problems, i.e., map matching, name extraction, location discovery and statistical attribute inference. Finally, we present the future research directions and conclude the paper.

Key words: volunteered geographic information; crowd-sourced spatio-temporal data mining; low-quality data processing; opportunistic sensing; urban computing

Foundation Item(s): National Natural Science Foundation of China (No.62076027, No.42201461); Beijing Institute of Technology Research Fund Program for Young Scholar (No.6120220113)

1 引言

城市中存在着各种与位置相关的决策问题。比如城市管理者希望利用给定的预算合理地修建设施^[1,2],提前预判人流量的激增以杜绝安全隐患^[3,4];企业希望实现效益最大的商业选址^[5,6]、高效的资源部署^[7];市民希望规划耗时最短的出行路线,被推荐感兴趣的景点或餐厅^[8]等等。随着社会经济的发展,城市中各种决策的科学化和精细化的要求在不断提高,而支撑科学、精细决策的基础是准确掌握城市地理信息。

地理信息包含相对静态的、通用的基础信息以及用于监测交通状况等城市运行的动态信息^[9]。在传统地理信息收集方法上,前者主要通过人工测绘^[10,11]的方式收集,后者以人工巡检或者固定地理传感器感知^[12]收集为主。这些手段通常面临设备、人力两方面成本较高的问题。比如,Google维护地图数据的成本高达10亿美金^[13],城市管理者同样需要花费高昂的成本才能部署足量地理传感器覆盖城市重要区域^[14]。诞生于2004年的OpenStreetMap(OSM)(<https://www.openstreetmap.org/>)项目利用众包(Crowd-sourcing)的思想,由广大的互联网用户共同贡献地理信息数据,显著降低了地理信息的收集成本。但是这些地理信息仍然需要依靠用户主动提供,数据的准确性和及时性受制于用户的活跃程度。

近年来,随着定位技术的普及以及移动互联网的发展,人群在日常的活动中产生了海量的移动性数据。这些带有时空属性的移动性数据,能够反映城市中部分重要的地理信息。例如,人群的轨迹反映了未收录的路网位置和交通状态,基于位置的社交网络(Location-based Social Network, LBSN)中用户在某地的贴文提及了该地的名称或别名,司机的行车记录仪等街景数据中记录了交通标识符的位置和交通事故等等。这种基于非专职群体贡献的时空数据推测城市地理信息的理念,是把移动群智感知(mobile crowdsensing)^[15,16]的思想应用在城市地理信息推测中。特别地,因为人群并非是为了收集地理信息而产生数据,这种没有平台施加控制的感知方式在移动群智感知中属于机会式感知(opportunistic sensing)^[14]范畴。Goodchild^[17]将这一理念称为自发地理信息(Volunteered Geographic Information, VGI);Campbell等^[14]将其看作城市计算(urban computing)^[18]中城市感知的一种方式,称其为以人为中心的城市感知(people-centric urban sensing)。前者是从基础地理信息采集角度对这一理念进行阐述,而后者更多是从地理传感器感知动态信息的角度阐述。但两

者传达的含义本质相同:利用城市中泛在的大量非专职人群的时空数据推测城市地理信息是一种成本低、空间覆盖广、时效性强的替代或补充方式。由于用来推测城市地理信息的众包时空数据由移动众包的模式获得,且数据均具有时空属性,故此类数据又被称为众包时空数据^[13]。

虽然利用众包时空数据得到城市地理信息的理念很美好,但是众包时空数据却因其源自非专职人员而面临一个非常巨大的挑战——数据质量低。具体体现在由非专业采集设备导致的定位精度不高、采样频率低的问题,以及由非特定目的产生导致的时空分布倾斜、注解行为不及时、注解或感知内容随意的问题。这些问题导致了基于众包时空数据推测得到的地理信息不准确、时空覆盖不全面,因此,引起了国内外研究人员的广泛关注。比如Zhao等^[19]在利用具有定位噪声的轨迹数据推测道路位置方面做了许多工作,Kravi等^[20]基于LBSN数据中杂乱的文本内容推测兴趣点类别,Song等^[21]基于配送员对轨迹数据的不规范文本注解识别兴趣点名称,Xu等^[22]基于时空倾斜的轨迹数据推测各时段各路段的平均速度。

值得一提的是,与上述基于给定众包时空数据推测城市地理信息不同,还有一类研究旨在通过众包平台施加控制从而提升众包时空数据质量^[13]。施加控制的方式包括:众包平台有选择性地招募众包工人(根据众包工人自身出行、专业程度等信息)^[23];众包平台为众包工人分配重复度较低的时空众包任务^[24];众包平台为不同重要程度的任务设置不同激励措施^[25]等。但此类研究需要众包平台能够控制对象的移动行为或者任务发布,对于很大一部分通过机会式感知获取的众包时空数据不适用。在众包平台可进行控制时,该类工作可看作众包时空数据驱动的地理信息推测的前置步骤,为地理信息推测提供了更高质量的数据来源。但由于众包时空数据的数据质量问题只能被缓解,而不能被完全消除,因此仍然需要后续地理信息推测工作。

目前众包时空数据驱动的地理信息推测相关综述大多数据来源单一、地理实体类型单一且推测信息的维度较为有限,如Chao等^[26]综述了基于众包轨迹数据推测路段位置的方法,Silva等^[27]综述了基于LBSN数据的兴趣面位置发现与语义推测方法。虽然唐炉亮等^[28]的综述介绍了利用多种众包时空数据推测各类地理实体的位置和动态属性的方法,但是该综述一方面对地理信息推测的维度仍然不够全面(缺少对地理实体名称等重要静态属性推测方法的总结),另一方面对

除了定位精度不高以外的其他数据质量问题的讨论较少(忽略了时空分布倾斜等其他数据质量问题对地理信息推测的影响)。

鉴于此,本文以城市中的道路、兴趣点、兴趣面等典型的地理实体为信息推测的对象,针对众包时空数据定位精度不高、采样频率低、时空分布倾斜、注解行为不及时、注解或感知内容随意等数据质量问题,总结了对这些地理实体的位置和各类属性的推测方法。

2 问题与挑战

本节首先介绍了如何利用众包时空数据推测地理信息相关的概念,然后介绍了多种典型的众包时空数据,接着比较了众包时空数据与其他数据源相比收集地理信息的优劣,最后阐明了该领域的核心研究问题和主要挑战。

2.1 概念

定义1 时空活动记录

时空活动记录 $p=(oid, l, t, readings, anno)$ 由移动对象的唯一标识符 oid , 活动发生的地理空间坐标 l , 活动发生的时间 t , 移动对象在该时刻通过携带的传感器被动感知的信息 $readings$, 以及移动对象在该时刻主动产生的注解 $anno$ 五元组构成。根据众包时空数据类型不同, $readings$ 和 $anno$ 可能为空。

定义2 时空活动序列

时空活动序列为一个移动对象按时间递增排序的时空活动记录序列, 记为 $T_{oid}=\langle p_1, p_2, \dots, p_N \rangle$ 。其中, N 为时空活动记录个数。

定义3 众包时空数据

众包时空数据是海量移动对象的时空活动序列集合, 记为 $T=\{T_1, T_2, \dots, T_M\}$, 其中, M 为移动对象的个数。

定义4 地理实体

地理实体是占据了空间中一定位置的实体, 具有地理位置 shp 和属性信息 $F^{[29]}$, 记为 $e_g=(shp, F)$, 其中 shp 为各地理实体 e_g 必须具备的信息, F 为可选信息。地理实体 e_g 的集合记为 GE 。当 shp 为一个地理空间坐标 l 时, 该地理实体为点状实体, 如餐馆、酒店、住宅楼等兴趣点 (Point Of Interest, POI); 当 shp 为一个地理空间坐标构成的序列 $\langle l_1, l_2, \dots, l_n \rangle$ 时, 该地理实体为线状实体, 如路段; 当 shp 为一个封闭的地理空间多边形 $\langle l_1, l_2, \dots, l_n, l_1 \rangle$ 时, 该地理实体为面状实体, 如小区、公园等兴趣面 (Area Of Interest, AOI)。属性 F 根据值类型的差异可分为文本属性和统计属性^[30]。文本属性的值是描述性文本, 没有固定的取值范围, 如地理实体的名称就属于文本属性; 统计属性的属性值有固定的取值范围, 可以是类别属性也可以是数值属性, 如兴趣点类

别、道路限速等静态统计属性, 以及区域人流量、路段平均速度等动态统计属性。

特别地, 如果地理实体表示的是路段, 则地理空间坐标构成的序列中第一个坐标通常表示路段起点 l_1 , 最后一个坐标 l_n 通常表示路段终点。多个路段可能共享同一个起点或者终点, 这些点即为交叉路口。路网是路段的集合。

定义5 众包时空数据驱动的地理信息推测

基于众包时空数据 T , 生成 $e_g.shp (ge \in GE_{new})$, 或推测 $ge.F_i (ge \in GE_{exist})$, 其中, GE_{new} 是新生成的地理实体集合, GE_{exist} 是现有地理实体集合, F_i 为地理实体的某个属性。

2.2 典型的众包时空数据类型

众包时空数据有多种形式, 本文介绍3种典型的众包时空数据即轨迹数据、LBSN数据、众包街景数据, 并在表1中对这些数据进行比较, 使读者有较为直观的概念。

表1 典型的众包时空数据类型

类型	数据来源	采样率	多媒体	注解
轨迹	配送员、快递员、网约车等	高	×	部分
LBSN	Flickr、微博、FourSquare等	较低	部分	√
街景	KartaView、Mapillary等	高	√	×

轨迹数据是城市中常见的众包时空数据。轨迹数据中的每条时空活动记录是全球导航卫星系统 (Global Navigation Satellite System, GNSS) 装置被动采集的轨迹点, 通常轨迹点间采样间隔较短 (一般在数秒到数分钟) 且采样时间间隔均匀。由于监管的需要, 城市中的出租车、网约车、共享单车、物流快递员、外卖配送员等在业务经营过程中都会产生海量的轨迹数据。此外, 因为不同业务自身的特点, 移动对象有时还会产生一些活动日志。通过时间将轨迹与活动日志关联后, 这些活动日志可以看作移动对象对轨迹的注解。比如网约车司机在载客时会确认发起打车请求的用户的上下客状态, 快递员、配送员在配送时会确认各地址的包裹的送达状态等等。

LBSN数据是一种普遍存在的众包时空数据源。随着移动互联网的发展, 社交媒体用户常在平台上发布贴文时可能携带位置信息, 故一条带有位置的社交媒体文档也可以看作一个时空活动记录。因为贴文的发布是一种主动行为, 因此用户贴文发布的时间间隔不固定, 而且一般发布频率较低。从分享内容上看, 微博、Twitter (<https://twitter.com>) 用户发布的信息涵盖了地名、标签、包含情绪的文字等, Flickr (<http://www.flickr.com>) 用户以分享照片为主。

众包街景数据是一类具有前景的众包时空数据, 其从地面视角记录了海量的地理实景。Mapillary

(<https://www.mapillary.com>)和 KartaView (<https://karta-view.org>)是目前两个较大的众包街景数据平台. 平台应用程序会自动以固定的时间间隔(一般为数秒)拍摄照片,同时记录下照片拍摄时的位置和方向等信息,并以图像序列的形式上传到平台. 因此,众包街景数据也可以看作一种以照片、传感器读数为被动感知信息的轨迹数据.

2.3 基于众包时空数据的地理信息推测与其他地理信息收集方法的对比

地理信息的收集有两大主要目的,一是为了测绘

采集基础地理信息,二是为了监测城市的运行. 在传统上,前者的手段以人工测绘^[10,11]为主,后者以人工巡检和地理传感器^[12]为主. 由于众包时空数据可被用来推测地理信息,因此其可以被看作一种新型的地理信息收集方法,且一定程度上能够同时完成测绘和监测两大任务. 虽然从卫星图像推测地理信息也具有类似的优点,但基于众包时空数据得到的地理信息维度更丰富的基础上,在成本、空间覆盖、时效性上也具有一定优势. 表2从各维度比较了基于众包时空数据的地理信息推测与其余地理信息收集方法.

表2 基于众包时空数据的地理信息推测与其他地理信息收集方法的对比

收集目的	收集方式	成本	空间覆盖	时效性	地理信息感知能力				数据质量
					位置	名称	统计属性(静)	统计属性(动)	
测绘	人工测绘	高	有限	季度级	√	√	√	×	高
监测	人工巡检	高	有限	分钟级	×	×	×	√	高
	地理传感器	高	有限	分钟级	×	×	×	√	高
测绘+监测	基于卫星图像推测	高	广	周级	√	×	√	√	中
	基于众包时空数据推测	低	广	秒级	√	部分	√	√	差

从成本角度,传统地理信息收集的方式以及通过卫星收集的方式不论是在设备安装部署还是专业人员聘用方面都具有较高的成本. 而对众包时空数据驱动的收集方法而言,一方面数据是由海量移动对象在自身活动时被动完成积累,没有主动收集成本,另一方面位置等地理信息是通过移动设备中的价格低廉的GNSS模块实现收集,因此成本极低.

从空间覆盖角度,因为专职人员有限,因此不论是人工测绘和人工巡检,都仅能覆盖有限地理空间. 以道路监控、地感线圈为代表的地理传感器通常也仅部署在重点区域,如济南某区域内18 727条路段中仅有354个点位部署了道路监控^[31]. 与上述方法相比,基于卫星图像的方法具有广泛的空间覆盖,但是俯视时云层和树木的遮挡会对地理信息的感知产生一定影响. 而众包时空数据驱动的收集方法,综合了市域范围内非专职人员提供的信息,具有广泛的空间覆盖且不存在视觉遮挡的问题.

从时效性角度,人工测绘由于高昂的成本仅能以季度为周期进行. 公众可以接触到的卫星图像的更新周期一般在周和月级别. 人工巡检涉及人的移动,因此信息收集的频率大概在分钟级和小时级. 地理传感器信息收集的时间间隔与传感器设置有关,秒级和分钟级是常见的信息收集间隔. 而对于众包时空数据来说,由于其基于群体上报的数据进行更新,因此及时性非常高,时效性在秒级.

从地理信息感知能力角度,本文主要从位置、名称和统计属性三个维度展开比较,主要是因为文本属性具有任意性,而名称属性是较为通用的地理实体文本

属性. 人工测绘的方法自主性高、目的明确,因此位置、静态统计属性等信息都能完成采集,但因其采集频率较低,一般不用于监测快速变化的动态统计属性. 与之相反,监测类的手段聚焦在地理实体的动态统计属性感知上,如人工巡检常用于发现违章停车等实时事件^[32],地理传感器能够监测车流量^[33]等信息. 卫星图像中包含的信息相对丰富,除了专有的名称信息无法感知外,可以实现位置信息的推测^[34],以及完成土地用途分类^[35]、车道数识别^[36]等一些统计属性的推断. 由于其数据更新周期相对较短,因此一定程度上也能实现动态统计属性的感知^[37]. 而众包时空数据包含更为丰富的信息. 首先,由于人群的移动性特点,其天然能够用于推测位置信息. 其次,不同区域数据量的多少本身也反映了人群活动的热度,照片、文本等多媒体信息反映了土地使用、情绪变化等信息,轨迹等采样频率较高的众包时空数据还能实现速度等信息的感知,故其能够实现多维度的,包括静态和动态的地理实体统计属性推测. 最后,移动对象注解的本文中可能还蕴含了名称,因此基于众包时空数据提取名称同样可行.

综上所述,与其他方法相比,利用众包时空数据推测地理实体信息具有收集成本低、空间覆盖广、数据鲜度高、信息维度多的优点,因此其成为一种理想的收集城市地理实体信息的补充方案. 但由于该数据并非由专职人员收集,故其在收集的数据质量上与其他方式相比处于劣势. 该劣势即是利用众包时空数据推测地理信息面临的主要挑战和研究人员着力解决的问题.

2.4 主要研究内容和框架

如第 2.1 节所述,地理实体包含了位置以及属性两种类型的知识. 由于名称信息和统计属性信息在推测方法上存在较大不同,因此本文进一步将属性知识细分为名称知识和统计属性知识. 利用众包时空数据完成地理信息的收集本质上是要实现这些知识的发现与关联.

除了名称、位置、统计属性三类信息的推测以外,在位置信息和统计属性推测前,地图匹配技术常常需要被用到. 故本文将基于众包时空数据的地理信息推测从实现路径上分为地图匹配、名称提取、位置发现和统计属性推测四部分,各块研究内容之间的关联关系如图 1 所示. 挖掘得到的知识可以直接支持设施选址、流量预测、路线规划、调度部署等上层基于位置的城市应用,或者与现有地理信息进行地图融合^[38,39]后支持上层应用. 四块研究内容的具体含义如下.

(1)地图匹配旨在将众包时空数据与现有地理信息关联,用于推测地理实体统计属性,以及基于未匹配上的众包时空数据发现地理实体位置.

(2)名称提取旨在基于众包时空数据中的海量注解信息提取名称.

(3)位置发现旨在基于众包时空数据中的空间坐标信息推测地理实体的位置. 根据待发现位置的地理实体的名称是否已知,分为命名地理实体发现和匿名地理实体发现两种方式. 两者的主要区别为是否对用于推测位置信息的众包时空数据做出限定. 前者需要基于“名称提取”提取得到的名称,在众包时空数据中筛选提及该名称的众包时空数据子集推测名称对应的位置;而后者一般直接采用全量的或者通过经过“地图匹配”而匹配失败的众包时空数据推测地理实体位置.

(4)统计属性推测旨在基于关联在地理实体上的众包时空数据推测地理实体的统计属性. 根据地理实

体位置是否源自现有地理信息,数据的来源有两方面. 第一,如果地理实体的位置信息通过“位置发现”推测,则众包时空数据与地理实体位置已经天然完成关联,可直接用于统计属性推测;第二,如果地理实体的位置来自现有地理信息,则众包时空数据需要先通过“地图匹配”将众包时空数据关联到现有地理实体上,然后基于匹配成功的数据推测统计属性.

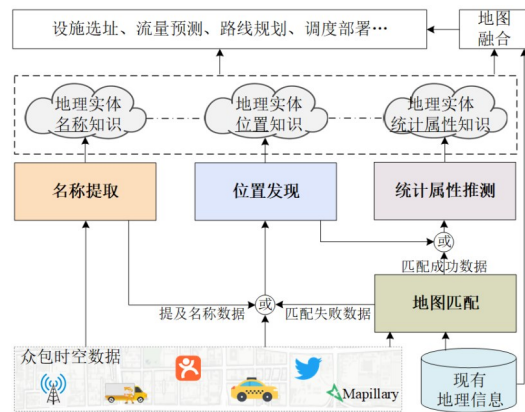


图 1 众包时空数据驱动的地理信息推测框架

2.5 数据质量问题对地理信息推测的影响

如第 2.3 节所述,由于众包时空数据质量较低的问题,其对于地图匹配、名称提取、位置发现以及统计属性推测四方面都带来了巨大的挑战. 图 2 对众包时空数据中的各种问题对地理信息推测带来的挑战进行了总结. 众包时空数据由非专职人员产生,这一方面意味着人群采用非专业设备进行数据采集,另一方面意味着人群产生数据的行为并非是测绘或监测驱使. 前者存在定位精度不高、采样频率低的设备质量问题,后者导致了数据时空分布高度倾斜、注解时间存在噪声、注解或感知的内容随意的人为问题. 下面对这五方面问题分别进行介绍,并讨论这些问题对地理信息推测带来的挑战.

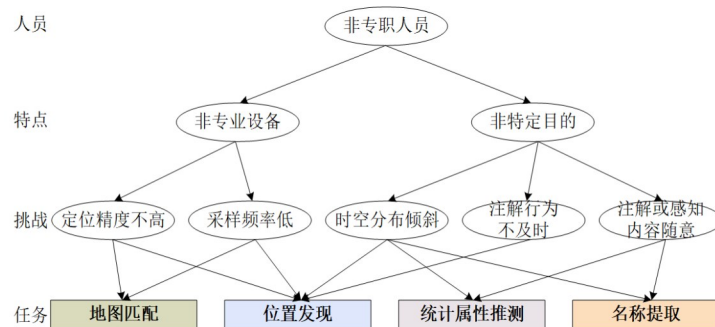


图 2 众包时空数据驱动的地理信息推测框架

2.5.1 定位精度不高

由于众包时空数据是随着移动互联网的兴起而广

泛出现,因此一般众包时空数据中位置信息来自手机中的定位模块. 这类 GNSS 定位装置的定位误差一般为

5~15 m,尤其在一些城市中楼宇密集的区域,因为峡谷效应会产生较大的定位误差.这类定位误差对于位置发现和地图匹配均会产生不利影响.从位置发现角度,较大的定位误差会导致空间相近的平行路段难以区分、空间相近的楼宇识别困难等问题出现.从地图匹配角度,由于定位误差的存在,在地理实体分布较密的城市区域,最近邻匹配往往会关联出错.

2.5.2 采样频率低

由于LBSN等因为移动对象主动发布的行为而产生的众包时空数据本身频率极低,故这里主要讨论轨迹等通过被动方式采集的众包时空数据的采样频率对地理信息收集的影响.通常情况下,移动设备为了省电等因素的考虑,不会将采样频率设置过高.根据Yuan等^[40]的统计,在北京的一份出租车数据集中仅有34%的轨迹数据采样间隔在1 min以内.这些低采样率的轨迹数据同样会给位置发现和地图匹配带来困难.比如对于位置发现的典型应用路网生成而言,为克服空间上数据稀疏的问题,通常会将一条轨迹中时间上前后连续的两个点画直线连接,基于这些连线生成路段.如果移动对象行驶在非直线上移动或者进行了转弯,在采样频率较低时,两点的连线与真实路网情况相差甚远.对于地图匹配来说,低采样频率导致两点间移动对象的位置存在极大的不确定性,进而影响移动对象真实移动路段序列的推测.

2.5.3 时空分布倾斜

测绘、巡逻这些地理信息收集方法一般都是有计划、有组织的,因此数据的收集在时空上分布较为均匀.而对于众包时空数据而言,移动对象的移动由其主观意愿驱动(以LBSN为代表)或者是业务经营需要产生(以网约车、快递员为代表).人群活动在时空上服从长尾分布的现象导致了收集的数据在时空上的分布具有高度倾斜的特点^[41].时空上高度倾斜的数据分布造成了在空间上某些热门区域数据重复收集,而另一些冷门区域仅有很少或者几乎没有数据,该问题在时间域上同样存在.在数据缺失的区域,由于位置和名称的唯一性,无法完成位置发现和名称提取任务,但是对统计属性的推测通过相关性建模依然有可能实现.然而,对缺失属性进行准确推测本身也是难点.

2.5.4 注解行为不及时

不同于传感器被动感知,感知的信息一定对应了感知的时间,注解是人群的主动行为.因此,注解内容发布的时间可能对应了稍早之前已经完成了的活动.比如,配送员在配送完某地的包裹后过了一段时间才确认包裹送达,社交媒体用户晚上回家后才提及白天游玩的景点等等.在这种情况下,注解中提及名称时用户的空间坐标就不是名称对应的真正地点.该问题会

对位置发现中的命名位置发现产生干扰.

2.5.5 注解或感知内容随意

由于人群的活动并非是为了测绘或者监测,因此他们注解或者携带的传感器感知的内容具有很强的随意性.这种随意性对名称提取和属性推测都会产生影响.从名称提取角度,首先注解的内容不一定是名称,其次即使注解的内容包含了名称,如何准确地识别名称、发现别名,以及纠正自然语言中的错别字都是难点.从统计属性推测角度,人群提供的信息可能并不能直接得到地理实体的统计属性,但可能注解或感知的信息和地理实体的统计属性具有相关性.比如众包时空数据可能并没有收集地理实体的类别,但是可根据移动对象的平均通行速度、驻留时长等信息反映,因此需要进一步建模.

3 研究现状分析

本节对众包时空数据驱动的地理信息推测的研究现状进行总结与对比分析.首先介绍利用众包时空数据推测城市地理信息的预处理步骤地图匹配,然后分别介绍对地理实体名称、位置、统计属性三类信息推测的现有工作.

3.1 地图匹配

地图匹配是指将众包时空数据与现有地理信息数据的关联,其是利用众包时空数据推测地理信息的基础.地图匹配有两个作用:第一,通过关联在现有地理实体上的众包时空数据,实现对后续多种地理实体统计属性(比如速度、流量等)的推测;第二,匹配失败的众包时空数据,意味着某位置有活动的发生,而该位置缺失地理实体信息,进而可以利用众包时空数据补充和更新缺失地理实体位置信息.

对于轨迹等采样频率较高的众包时空数据,因为其序列上下文信息丰富,故在地图匹配之前,不仅能够通过噪声过滤消除一些定位错误,而且还可以通过驻留点检测识别移动对象运动和静止的语义,以实现后续更高质量、目的更明确的地图匹配^[42].噪声过滤方法大致分为三大类:基于均值滤波的方法^[43],基于卡尔曼或粒子滤波的方法^[43],以及基于启发式的异常点检测方法^[44].前两者将噪声点替换为一个估计的值,而后者直接将异常点从轨迹中移除.基于启发式的异常点检测方法基于当前轨迹点及其前驱轨迹点,计算每个点的瞬时速度.如果瞬时速度大于一个给定的阈值,当前轨迹点会从原有数据中移除.在驻留点识别方面,Li等^[45]提出了一种基于距离阈值和最短停留时长的驻留点检测方法.首先从一个锚点出发,定位一段由其和其后继构成的最长子时空活动序列,在该序列中锚点与后继点的距离均小于距离阈值 D_{max} .如果该锚点与时

间上间隔最远的一个 D_{\max} 内的后继点的时间跨度大于最短停留时间 T_{\min} , 则该段子序列看作一个驻留点, 锚点会被移动到该段子序列的直接后继。否则, 锚点设置为当前锚点的直接后继, 继续检测。该过程不断重复, 直到锚点移动到时空活动序列的末端, 该时空活动序列中所有驻留点检测完毕。Yuan 等^[46]用密度聚类的思想对其进行改进, 使在检测候选驻留点时锚点可以根据密度可达移动, 使其能够更好地识别出租车的上下客地点。

在完成众包时空数据预处理后, 处理干净后的众包时空数据即可与地图匹配。由于 AOI 在地理空间上占据一定较大区域面积, 因此众包时空数据与 AOI 的匹配方法较为简单, 只须判断空间包含关系^[47]即可。而对众包时空数据与 POI 或路网的匹配则较为复杂。根据众包时空数据在匹配时是基于单条时空活动记录还是整条时空活动序列, 地图匹配可分为时空活动记录与地理实体的匹配和时空活动序列与地理实体的匹配。

3.1.1 基于时空活动记录的匹配

时空活动记录与地理实体的匹配一般用于时间序列上非常稀疏的、几乎不存在序列依赖关系的众包时空数据匹配问题上。如 LBSN 数据与 POI 匹配时, 由于每条数据基本都是独立产生, 一般采用最近邻匹配或者基于距离阈值的匹配^[20]。在轨迹中的驻留点与 POI 匹配时, 可使用基于空间相交关系或者最近邻的方法^[48-50]、采用贝叶斯网络的方法^[51]以及基于梯度提升树的方法^[52,53]等。此外, 在打车等场景中, 请求方仅有单个坐标点已知, Wu 等^[54]利用生成模型建模了在路段上的移动对象产生一个坐标点的过程, 并基于历史数据估计了生成模型的参数, 然后将单坐标点匹配建模为最大后验估计问题 (Maximum-A-Posteriori, MAP)。

3.1.2 基于时空活动序列的匹配

时空活动序列与地理实体的匹配可以考虑更多因素, 原因是时空活动序列前后存在一定程度的依赖关系, 关联时考虑上下文能够提升整体匹配的准确性。

由于轨迹数据的普遍性和通用性, 轨迹数据和路网的匹配吸引了研究人员的广泛关注, 该工作通常被称为路网匹配^[40,55-57]。经典的算法通常将该问题建模为求解隐马尔可夫模型 (Hidden Markov Model, HMM) 的隐状态序列, 动态规划算法是常用解法^[40,55,56]。随着深度学习的发展, 现在也有一些研究采用递归神经网络等深度学习模型完成这一目的^[57,58]。如 Ren 等^[58]将轨迹与路网的匹配看作一个“序列到序列”问题, 提出了细粒度路网匹配框架 MTrajRec, 其不仅能实现轨迹与路网的匹配, 还能减少低采样频率轨迹的不确定性。MTrajRec 的结构如图 3 所示, 其遵循编码器-解码器结

构。编码器以网格化的原始轨迹作为输入, 解码器融合天气等外部因素后, 通过注意力机制和前一时间步的信息, 基于多任务模块同时输出匹配的路段编号以及距离路段起始的偏移率。由于目标采样频率可能高于原始输入, 因此解码器根据当前时间步观测是否缺失, 分别执行插值匹配或者直接匹配。在直接匹配时, 当前轨迹点距离附近路段的距离会作为一个软掩码施加在路段预测之前。该软掩码作为先验知识以提升匹配准确性。

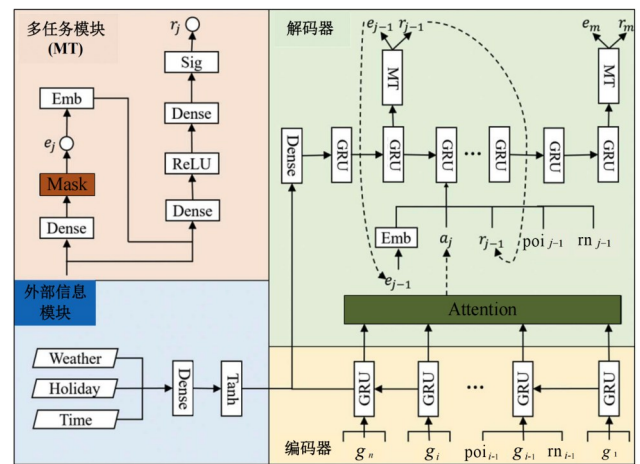


图3 基于深度学习的细粒度路网匹配模型 MTrajRec 的结构

在将驻留点与 POI 匹配时, 由于一个移动对象的驻留点间虽然时间间隔较远, 但有时语义上前后依然存在着某些依赖关系, 故有些工作也考虑了上下文信息^[59-61]。Yan 等^[59]利用 HMM 推测驻留点序列对应的 POI 类别访问序列。Nishida 等^[60]将驻留点和地点的关联问题建模为半监督学习, 利用分层贝叶斯模型进行匹配。Suzuki 等^[61]利用整数规划的方法进行匹配。上述工作主要聚焦在研究移动对象行动意图未知的情况下, 将轨迹数据与空间相近的地理实体进行关联。而如果已知移动对象要访问的地理实体, 现有方法可能产生不合理的匹配。Ruan 等^[62]以平台知晓配送员配送旅程中必须要访问的投递地点集合为例, 研究了已知驻留意图下, 考虑上下文的驻留点与地点的匹配问题。

3.2 名称提取

名称是对某一地理实体的唯一性指代 (指位性), 比空间地理坐标更容易让人理解和记忆。人们通常很少知道他们经常访问的场所、道路的经纬度, 但是当提及这些场所或道路的名称时, 他们能够很容易将其与具体地理实体联系起来。此外, 名称对于地理信息的检索等场景而言也不可或缺。

有别于地理实体的统计属性, 由于名称的唯一性, 该信息无法基于别的因素推测, 只能依靠人的采集。从众包数据角度, 早期有一些工作利用海量网页文档提

取名称^[63]。随着移动互联网的发展,基于众包时空数据中移动对象的主动注解提取名称变得更为常见,这些方法主要利用两种带注解的众包时空数据:LBSN数据和物流配送数据。

3.2.1 基于LBSN数据的名称提取

在LBSN数据方面,研究人员主要聚焦在判断LBSN数据中用户打的标签(tag)是否是名称。一些早期的工作^[64]从GeoNames等外部数据源中获得名称,通过和标签比较来确定其是否为名称。但这类方法仅能完成标签是否为已有名称,无法从LBSN数据中提取新名称。Hollenstein等^[65]采用一些简单规则进行判断,如标签中是否包含了“市”“区”“路”等关键词。但这类方法只能识别有限的符合一定命名模式的名称。Rattenbury等^[66]提出了两种基于统计学中显著性检验的名称识别方法。这两种方法都基于一个非常直观的洞察:一个在一定区域内经常被提及,而在其他区域很少被提及的标签,更有可能是名称。这一想法与信息检索领域中的著名算法词频-逆文档频率法(Term Frequency-Inverse Document Frequency, TF-IDF)^[67]思想类似。故Rattenbury等^[66]的第一种方法是基于TF-IDF的算法。该方法首先通过k-Means^[68]将LBSN数据聚类为多个地理区域,然后将当前区域看作文档,区域内所有的标签看作词,全体区域看作文档语料库,然后计算当前区域上每个标签的TF-IDF值,如果一个标签的TF-IDF值高于一定阈值,其被认为是名称。Rattenbury等^[66]还提出了一种尺度结构识别方法,该方法的思想是如果一个标签是名称,提及该标签的LBSN数据的空间分布应该在任意空间尺度上都仅以单个簇的形式出现。但是上述两个方法也存在一定问题:由于其没有考虑时间维度的信息,统计意义上显著的标签可能是一些事件词汇,比如“某马拉松比赛”等等。

3.2.2 基于物流配送数据的名称提取

在物流配送中,名称信息隐含在用户填写的收货地址中。从收货地址中提取名称可以建模为自然语言处理中经典的命名实体识别问题(Named Entity Recognition, NER),并通过序列标注模型^[69]进行提取。如Abid等^[70]利用双向长短期记忆神经网络(Bi-directional Long Short Term Memory neural network, BiLSTM)实现名称提取。得益于最近自注意力结构在序列数据上相较于长短期记忆神经网络(Long Short Term Memory neural network, LSTM)等递归神经网络的建模优势,Song等^[21]提出了如图4所示的在预训练语言表征模型(Bidirectional Encoder Representation from Transformers, BERT)^[71]上叠加条件随机场(Conditional Random Field, CRF)的序列标注方法,实现了更为精准的名称提取。此外,He等^[72]在名称提取的基础上,以填写某名

称的用户群体移动性画像为依据,利用深度孪生网络实现了地理实体别名识别,进一步完善了名称知识库。

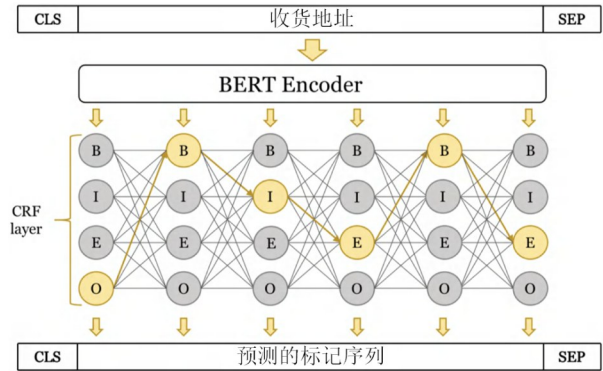


图4 Song等^[21]提出的名称提取模型

3.3 位置发现

地理实体的位置是地理实体最重要的信息。在地理实体位置发现中,根据是否利用名称对众包时空数据进行筛选,可以将问题分为命名地理实体位置发现和匿名地理实体位置发现两类。在命名地理实体位置发现中,需要基于众包时空数据发现对应该名称的单一地理实体位置;而在匿名地理实体位置发现中,需要基于众包时空数据生成包含位置信息的地理实体集合。由于这两类问题发现的地理实体可以是个体也可以是集合,故下面分别进行讨论。

3.3.1 命名位置发现

利用众包时空数据进行命名地理实体发现,不仅能够发现地理实体的位置,本质上还完成了名称与位置的链接。命名地理实体位置发现的一般流程是基于3.2节提取的名称,筛选出注解中提及该名称的众包时空活动记录子集,基于这部分子集的空间坐标集合,生成不同地理实体的位置表示。根据生成的地理实体的类型不同,下面分别介绍命名路段位置发现和命名POI, AOI位置发现两类问题。

3.3.1.1 命名路段位置发现

由于用户在路上较少产生注解行为,因此命名路段位置发现的工作较少。Gao等^[73]为了解决路段上数据较为稀疏的问题,基于路段一般为线状的观点,将LBSN数据按经度或纬度值大小顺序依次连线构成路名对应的路段表示。但在数据非常稀疏的路段,基于连线的方法描述路段准确性不高。

3.3.1.2 命名POI、AOI位置发现

除道路之外,POI和AOI是另两类重要的地理实体。与命名路段位置发现鲜有研究相反,命名POI, AOI位置发现的问题有较多的研究。原因是LBSN数据和物流配送数据中的注解行为更多发生在地点或区域上。

在LBSN数据方面,在利用名称对标签进行筛选之

后,筛选得到的文档的经纬度点集合后续通过聚类^[74]、密度估计^[64,65]、Delaunay 三角化^[75]以及标准差椭圆^[73]等方法得到了 POI 或 AOI 的点表示和面表示. 较为遗憾的是,由于 LBSN 数据的产生方式,往往用户提及的名称较为集中,都是一些较为知名的景点、商铺等场所,而且这些名称往往是对一个较大的区域范围的指代. 因此,得到的信息对现有 POI, AOI 数据起到补充或更新的作用有限.

在物流配送数据方面,基于确认某名称的包裹收货时配送员所在的位置集合,也能够进行命名 POI, AOI 推测. 在 POI 位置推测方面, Ruan 等^[62,76]最先研究了该问题,并通过计算空间质心的方式得到了名称对应的地点. 意识到配送员确认收货的时间可能存在滞后,从而导致了配送员注解的位置并非是名称对应的地点的问题, Srivastava 等^[77]提出了基于密度的含噪数据空间聚类算法 (Density-Based Spatial Clustering of Applications with Noise, DBSCAN), 其选取数据点个数最多的簇作为名称对应的地点表示. 除了上述无监督的位置推测方法之外, Forman 等^[78]以名称对应的所有位置作为地点候选, 利用成对学习排序的模型, 选择赢得比较次数最多的位置作为名称对应的地点. 由于成对学习排序模型每次仅考虑两个地点候选, 缺少全局的视角, Song 等^[21]提出了一个端到端的位置推测模型, 直接推测名称对应的地点位置. 该方法以位置数据最多的区域为中心, 将其附近一定范围内的地理空间网格化, 并为每个格子从物流注解数据以及客户位置数据中提取多种特征, 然后采用图像中的语义分割模型挑选出名称最有可能对应的地点所在格子. 上述方法中对位置集合的选定仅局限在配送员确认收货时的位置附近, 在配送员存在普遍的确收货滞后的情况下难以准确推测地点. 与每次配送仅考虑配送员确认收货时所在位置不同, Ruan 等^[79]意识到配送员在配送过程中会在地点产生驻留行为, 因此在每次配送某个地名的包裹时, 以配送员出发配送的时间为下界, 以确认收货的时间为上界, 提取这一段时间内的配送员轨迹中的驻留点作为该次配送的地点候选, 使地点候选更有可能包括地名对应的地点. 基于框定的地点候选, Ruan 等^[79]提出了一个基于 Transformer 编码器^[80]的端到端地点推测模型 LocMatcher 以适应不同地名对应的地点候选个数动态变化的情况. LocMatcher 的结构如图 5 所示. 每个地点候选自身的画像以及地点候选和名称的相关性特征构成每个地点候选的特征向量, 所有地点候选的特征向量会一同输入到 Transformer 编码器中用于感知全局地点候选之间各维度之间的关系. 在全局关联性建模后, 每个地点的嵌入向量会与名称特征向量以注意力的形式计算相关性, 得分最高的地点候选为模型

推测的名称对应地点. 与 Song 等^[21]的方法会产生冗余地点候选不同 (一定范围内所有格子都会被看作地点候选), 该方法仅考虑移动对象有访问的位置作为地点候选, 并且地点候选的位置直接估计自轨迹, 相较于格子中心更为准确. 除命名 POI 位置推测之外, 在命名 AOI 位置推测方面, Dahiya 等^[81]采用核密度估计 (Kernel Density Estimation, KDE) 的方法识别了社区名的区域范围.

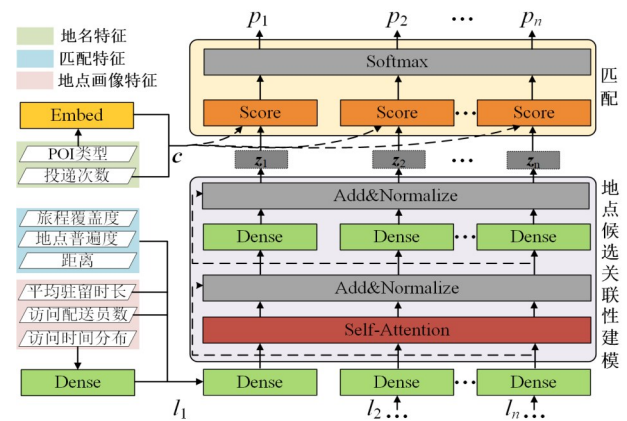


图5 Ruan 等^[79]提出的命名地点发现模型

表 3 对现有的众包时空数据驱动的命名地理实体位置发现方法进行了对比. 可以发现, 虽然 LBSN 数据、物流配送数据产生方式不同且两个研究领域交集较少, 但两类数据本质上都隐含了人对于位置的名称注解, 因此具有类似的地理实体位置发现方法可供一起讨论. 所有方法都可以大致分为位置候选生成和位置候选选择两阶段, 具体可分为三类: (1) 仅生成唯一的位置候选的方法, (2) 生成多个位置候选并启发式选择位置候选的方法, (3) 同样生成多个位置候选, 但采用监督学习选择位置候选的方法. 从位置候选表示能力的角度, 由于 GNSS 定位噪声的存在, 通过多个注解位置聚合得到的位置候选具有较准的位置候选表示能力; 其次是空间网格划分的方法, 其位置候选表示能力取决于网格划分的空间分辨率 (空间分辨率是位置候选表示能力和位置候选选择准确性的权衡); 最后是直接使用原始注解位置来表示位置候选的方法. 从对注解不及时鲁棒性的角度, 全量注解位置共同作用于地理实体位置表示的方法没有抵御注解不及时的能力, 仅采用部分注解位置表示地理实体位置的方法对注解不及时具有一定的鲁棒性, 在前者基础之上, 融入额外信息的方法能够进一步提升对注解不及时的鲁棒性.

3.3.2 匿名位置发现

命名地理实体位置发现需要依靠移动对象的主动注解, 因此仅限于通过 LBSN、物流配送等有限的众包时空数据实现. 众包时空数据中更常见的一类数据是

表 3 众包时空数据驱动的命名地理实体位置发现方法对比

地理实体	额外信息	监督	文献	位置候选生成方法	位置候选个数	位置候选选择方法	能力比较	
							位置候选表示能力	对注解不及时鲁棒性
路段	无	无	文献[73]	经纬度顺序连线	唯一	—	弱	弱
POI 或 AOI	无	无	文献[62,76]	质心	唯一	—	强	弱
			文献[74,77]	聚类	多个	点数最多	强	中
			文献[64,65,81]	KDE	多个	密度最大	强	中
			文献[75]	Delaunay 三角化	唯一	—	强	中
			文献[73]	标准差椭圆	唯一	—	强	中
	有	文献[78]	所有注解位置	多个	F 成对排序	弱	中	
	客户位置	有	文献[21]	空间网格划分	多个	UNet	中	强
注解前 驻留位置	有	文献[79]	聚类	多个	Transformer	强	强	

仅由移动对象被动产生的、无名称注解的轨迹数据。用这类数据发现的地理实体虽然是匿名的,但是对于路径规划、嫌疑地点发现等一些不依赖名称的应用场景,同样很有价值。与介绍命名地理实体位置发现时相同,根据发现地理实体的类型不同,下面分别讨论匿名路网位置发现和更新,以及匿名兴趣点、兴趣面发现和更新。

3.3.2.1 匿名路网位置发现

基于众包轨迹数据的路网推测在近 20 年被广泛研究,至今仍是一个较为热门的研究问题^[82,83]。该方法相较于基于卫星图像的路网推测,一方面树木的遮挡不会造成道路信息收集失败的问题,另一方面轨迹数据天然反映了道路的通行方向,生成的路网具有更强的实用价值。

早期的路网推测算法以无监督学习为主, Biagioni 等^[84]将这些方法分为三大类。(1) 基于聚类的方法^[82,85-87]。这类方法首先通过空间聚类算法从原始轨迹点中得到顶点(或者小路段),然后利用轨迹数据将这些顶点(或者小路段)连接起来形成路网。如 Edelkamp 等^[85]首先基于距离和方向角初始化一些种子地点,然后采用 k-Means 聚类算法调整簇的中心,再利用轨迹数据将这些地点连接,构成路段; Chen 等^[86]首先采样一些轨迹点,然后基于聚类的方法生成局部的小路段,最后利用轨迹数据将这些小路段连接; Stanojevic 等^[87]通过聚类得到顶点,通过轨迹得到路段,最后利用图稀疏化操作简化路网; Fang 等^[82]基于多种移动对象的轨迹数据,设计了车辆类型感知的聚类算法找到局部有代表性的路网顶点,然后通过路网匹配算法将路网顶点连接。(2) 基于轨迹合并的方法^[88,89]。这类方法直接合并不同轨迹中由任意连续的两个轨迹点构成的边,从而得到路网。如 Cao 等^[88]基于粒子模拟减少轨迹点的噪声,用一种强度较大但是影响范围较近的力将

附近的轨迹拉到一起,用一种强度较小但是影响范围较远的力防止轨迹偏离过远。基于去噪后的轨迹,通过共享轨迹段生成路网。(3) 基于 KDE 的方法^[83,90-92]。这类方法对原始轨迹点进行核密度估计,并用多种图像处理技术(如膨胀、闭运算、骨骼化等)得到道路中心线。由于这类方法基于图像处理技术,故还需要通过一些后处理算法从图片中提取可路由的路网。如 Biagioni 等^[90]利用 KDE,将区域内轨迹点的分布提取为不同置信度的道路中心线,然后利用骨骼化技术从图片中抽取路网,最后采用一种密度感知的路网匹配算法移除较少车辆通行的道路中心线。与大部分路网推测算法关心全局路网推测准确性不同, RoadRunner^[93]侧重在轨迹数据量多的区域生成更高精度的路网,其也可以被看作一种基于轨迹合并的算法。RoadRunner 从一个顶点出发,沿轨迹流出方向创建小路段,并以轨迹集合后续移动的空间分布相似性为依据,尝试将小路段与已生成路网进行融合。由于 RoadRunner 生成路段需要依靠较多的轨迹,因此其后续又可与现有路网推测算法的结果融合,以进一步提高路网推测的召回率。

上述无监督的路网推测算法通常需要设置大量经验参数,且一致的参数难以确保在各区域都能获得较好的效果,若参数设置不当则可能导致空间相近的平行路段过度合并,或者一条路段识别为多条的情况出现。此外,基于连接一条轨迹中前后两个轨迹点从而生成路段的方法^[85,88,90]会导致产生很多“快捷路径”(shortcut)的问题。因此,最近也出现了一些基于监督学习的路网推测算法^[94,95],以学习轨迹到路网的映射关系。这类方法主要受基于卫星图像的路网推测启发,将要推测路网的区域划分为多个空间瓦片,然后从每个瓦片内的轨迹数据中抽取多种特征,将瓦片内轨迹到路网的映射建模为图像翻译问题。这类方法一方面基于数据自动学习映射关系,减少了需要手工设置的

参数量,另一方面可以看作基于KDE的方法的拓展,能够较好地应对低采样率的轨迹数据. DeepMG^[95]在一个瓦片内,从空间视角和转移视角提取了轨迹中的多种特征,然后通过多任务学习提升道路中心线推测的准确性. Wu等^[94]和Yang等^[96]具有类似的思想,但其意识到路网推测在轨迹稀疏区域较难实现,进而通过同时以轨迹数据和卫星图像数据作为输入,从而推测更为准确的路网. 以Wu等^[94]提出的DeepDualMapper为例,

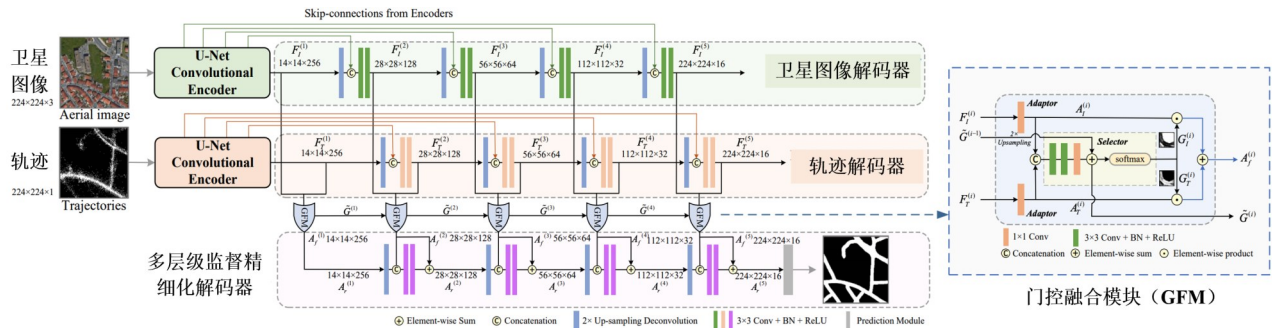


图6 Wu等^[94]提出的融合轨迹数据和卫星图像的路网发现方法

表4对比总结了具有代表性的众包轨迹数据驱动的匿名路网位置发现方法. 由于路网不仅具有位置信息,还包含路段与路段之间的拓扑关系,现有方法大致可分为位置推测和拓扑构建两阶段. 在位置推测阶段,除了基于图像的方法直接端到端推测所有路网位置外,其余方法均可分为道路点识别和路段生成两步. 道路点识别方法各异,但核心是从海量轨迹点中寻找局部较为密集、朝向等性质相似的代表性位置;路段生成用于连接上一步识别得到的道路点构成路段,具体可分为基于轨迹连接的生成以及基于朝向和距离的聚类两种. 在拓扑构建阶段,由于轨迹天然代表了路段之间的拓扑连通性,因此现有方法基本都通过轨迹构建路网拓扑. 在具体方法上,轨迹连接、轨迹合并属于在构建路网位置时同时保证拓扑正确性的方法,而路网匹配则是属于在位置推测结束后再执行的后处理方法. 同时,表4从“反向相近路(识别)”“同向相近路(识别)”“低采样率轨迹(处理)”“稀疏轨迹生成(路段)”四个维度比较了不同方法的处理能力. 可以发现,在路网生成中考虑对象移动方向信息的方法均具有反向相近路识别能力;监督学习方法以及融入额外信息的方法具有识别同向相近路的能力;在路段生成中,只要非基于轨迹连接生成路段的方法,均能较好地处理低采样率轨迹带来的“快捷路径”问题,此外如果采用低采样率轨迹连接生成路段,图稀疏化也一定程度上能够缓和该问题;在基于稀疏轨迹生成路段方面,基于朝向、距离的聚类生成路段的方法会面临数据不够无法生成的问题,考虑了轨迹前后点连线的路段生成方法一定程度上能够解决该问题,而通过融合卫星图像侧检测得到

所提方法如图6所示. 其通过UNet网络分别单独实现了基于轨迹数据和基于卫星图像的路网推测,然后在解码最终融合后的路网时,不同层级的特征图(feature map)会通过门控机制进行融合. 该模型在训练时,不同层级的特征图都会得到训练,使模型能够更好地融合两部分信息,结合两种信息各自优势完成预测. 与基于KDE的方法类似,这类基于图像的方法也需要后处理算法^[95]来保证路网的拓扑连通性和正确性.

的道路,具有更优的路段生成能力.

除了从无到有的路网推测外,利用众包轨迹数据进行现有路网的更新同样吸引了学界的关注. 目前大部分有人居住的区域多少有积累一些路网数据,这些路网数据有可能缺乏维护,从而较为过时或者不全. 因此,对这部分现有路网的更新也更具有现实意义. 利用众包轨迹数据进行路网更新一般遵循三步模式:(1)将轨迹数据在现有路网上进行路网匹配,并筛选得到匹配失败的轨迹;(2)利用路网推测算法基于匹配失败的轨迹生成路网;(3)将生成的路网与现有路网进行融合^[97-100]. 因此,上述路网位置发现算法也均可以看作路网更新算法的第二步. Wang等^[97]首先提出了路网更新的问题,其对未匹配成功的轨迹采用以轨迹Hausdorff距离为相似性度量的单联接聚类的流式变种算法^[101]进行聚类,然后利用多边形主曲线(polygonal principal curve)算法^[102]提取每个簇的代表性路段,最后通过一些启发式规则将生成的路段与现有路网连接. Cao等^[100]首次研究了基于众包行人轨迹的步行道路补全问题. 其首先将所有行人轨迹中最后路网匹配成功的轨迹点进行聚类,得到步行道路的出入口,然后对每个出入口对之间的轨迹用相似方法聚类,之后每个簇内轨迹对应的空间区域转换为地点网格为顶点、地点间相邻关系为边、地点轨迹出现概率为权重的图,并将代表性步行道提取问题建模为图上的最短路发现问题,最后利用A*算法提取每个簇的代表性步行道. 但如介绍路网推测时提到的,基于轨迹聚类的方法无法基于低采样率轨迹生成准确的路段表示,故Wu等^[98]和Shan等^[99]提出了基于轨迹点的路网更新算法以解决这一问

表 4 众包轨迹数据驱动的匿名路网位置发现代表性方法对比

额外信息	监督	文献	位置推测		拓扑构建	能力比较			
			道路点识别	路段生成		反向相近路	同向相近路	低采样率轨迹	稀疏轨迹生成
无	无	文献[85]	聚类	基于轨迹连接的生成	轨迹连接	√	×	×	中
		文献[86]	轨迹点采样	基于朝向、距离的聚类	路网匹配	√	×	√	弱
		文献[87]	聚类	基于轨迹连接的生成	轨迹连接 + 图稀疏化	√	×	√	中
		文献[88]	粒子模拟	基于轨迹连接的生成	轨迹合并	√	×	×	中
		文献[90]	轨迹线图 KDE		路网匹配	×	×	×	中
	有	文献[95]	语义分割网络		路网匹配	√	√	√	中
车辆类型	无	文献[82]	EM 算法	基于朝向、距离的聚类	路网匹配	√	√	√	弱
卫星图像	有	文献[94]	语义分割网络		—	√	√	√	强
		文献[96]	轨迹密度估计+遥感语义分割		规则	√	√	√	强

题. Wu 等^[98]提出了一种考虑方向和距离的聚类算法对原始轨迹点进行聚类,然后通过轨迹将生成的路段和现有路网相连. Shan 等^[99]基于全量轨迹点构建无向图,然后通过对其稀疏化以及进一步精细化得到缺失路段,最后以匹配失败轨迹前后匹配成功路段作为接口路段,将缺失路段补全至现有路网中. 意识到基于轨迹点的路网更新算法在轨迹数据量少的区域无法生成道路,同样存在不足,故 Wang 等^[103]提出了一种同时考虑轨迹点和轨迹段的路网更新算法 HyMU,他们分别对轨迹点和轨迹段采用聚类算法,然后利用扫描线(sweep line)算法^[104]得到每个簇的代表性路段候选,接着将相似的路段候选继续通过扫描线算法合并,最后基于空间临近性将推测的路段与现有路网连接. 上述算法均遵循传统的三步路网更新模式,但第一步路网匹配本身也可能带来误差,从而导致路网更新的误差. 针对该问题,Chao 等^[105]提出了一个路网匹配和路网更新联合优化的通用框架,以互相增强路网匹配和路网更新的质量. 该框架是一个迭代式的优化方法,在每轮迭代中,会依次进行路网更新、在更新后的路网上进行路网匹配、评估效果增益,该过程不断迭代直到不再有效果增益,即得到最终更新后的路网. 此外,路网中最需要确保的是路口的拓扑准确性. Zhao 等^[19]专门从校准路口拓扑角度出发对路网更新问题进行了研究,他们首先通过基于方向信息的聚类算法定位路口及其影响范围,然后再对影响范围内的轨迹进行聚类,并从每个簇中提取道路中心线作为路口的子路网,最后通过将子路网与对应区域内的现有路网进行路网匹配以发现不一致的拓扑,从而进行路口拓扑更新.

3.3.2.2 匿名 POI、AOI 位置发现

与匿名路网位置发现相同,匿名 POI、AOI 位置发现同样具有价值. 其一般有两个应用场景:首先,可以发现一些现有地理数据库中不存在的地点,可能是新的地点,也可能是一些嫌疑地点,比如可以从危化品车辆的驻留中发现违规小化工厂^[106];其次,与线状的路段不同,地点通常在空间上占有一定面积,且大小不一,但地理数据中鲜有地点的轮廓信息,而对匿名 POI、AOI 的推测过程中通常可以得到地点较为准确的轮廓,有助于后续实现众包时空数据和地点的准确关联.

匿名 POI、AOI 位置发现主要通过对轨迹中的驻留点聚类实现,簇的质心是 POI 的地理坐标点表示,簇的边界是 POI、AOI 的地理多边形表示. Ashbrook 等^[107]较早研究了该问题,并应用 k -Means 对众包时空数据聚类从而生成地点. 但是在地理空间上,实际的地点个数未知,难以设置. 故 DBSCAN^[108]和用于识别聚类结构的排序点聚类算法(Ordering Points To Identify the Clustering Structure, OPTICS)^[109]等基于密度聚类的方法后续更常被用于众包时空数据的地点生成任务中^[110-113]. 然而,上述算法无法约束生成的簇的大小,而在地点推荐中,用户希望得到的地点较为具体,故 Zheng 等^[114]提出了一种基于网格合并的方法生成地点. 该方法首先将空间划分为均匀的网格,然后统计每个格子内数据点的个数,接着每次贪心地选择驻留点个数最多的格子,将其与周围 8 个邻接的包含驻留点的格子合并,构成一个簇,该过程不断迭代,直到所有包含数据点的格子被分配到某一簇中时停止. 同样为了控制生成地点的大小,Ruan 等^[79]利用以最大距离阈值为停止条件的层次聚类算法^[115]得到包裹的投递地点候选集合. 该算法避免了基于网格合并的方法对空间的离散化,并在

实验中取得了更优的效果。

上述方法生成的地点均是空间二维信息。近年来也出现了利用众包 GNSS 数据还原周围环境中的三维楼宇的相关工作^[116]。其原理和上述方法基于移动对象实际驻留感知地点不同,该工作利用移动对象定位信号被干扰的情况,构建周围楼宇的三维信息。

3.4 统计属性推测

在第 3.3 节中,众包时空数据不仅用于发现地理实体的位置,而且还隐式地将众包时空数据与发现的地理实体关联起来。除此之外,众包时空数据还能够通过第 3.1 节介绍的地图匹配方法与现有地理实体进行关联。基于关联得到的众包时空数据可以推测地理实体的统计属性,以便对地理实体理解和监测,支撑上层设施选址、路段规划、流量预测等各种城市应用。

推测地理实体的统计属性有两方面挑战:其一,虽然地理实体的一些属性通过对众包时空数据的简单统计即可获得,而另一些属性无法直接获得,但是与众包时空数据存在相关性,需要利用模型进一步推断;其二,由于众包时空数据在时空上分布高度倾斜的特点,地理实体关联的众包时空数据也存在时空稀疏的问题,推断缺失数据关联的地理实体的静态属性或其某时间片的动态属性同样存在挑战。由此引出地理实体属性推测的两个研究方向:相关性建模和缺失值填补。

3.4.1 相关性建模

在完成众包时空数据和地理实体的关联后,已有一些统计属性可以通过简单统计众包时空数据获得,而另外一些属性需要通过建模进一步推测。表 5 对利用众包时空数据建模地理实体统计属性的方法进行了总结。

表 5 众包时空数据驱动的地理实体统计属性相关性建模方法总结

数据类型	地理实体	属性类别	属性名称	方法
轨迹	路网	静态	热度	统计 ^[1,2]
			可通行类型	统计 ^[117]
			道路限速	统计 ^[117]
			车道数	高斯模型 ^[118] 、高斯混合模型 ^[119]
		路段类别	层叠泛化 ^[120] 、SVM ^[82]	
		动态	平均车速	统计 ^[22,31,121]
	流量		贝叶斯网络 ^[122]	
	违章停车		KS 检验 ^[32]	
	POI	静态	热度	统计 ^[5,6]
			有趣程度	HITS ^[111]
			地点类别	XGBoost ^[106]
		动态	加油等待时长期望	统计 ^[52,53]
	AOI	静态	风险区域	CNN ^[123] 、图模式挖掘 ^[124]
			土地用途	SVM ^[125]
动态		流量	统计 ^[3,4]	
LBSN	POI	静态	地点类别	逻辑回归 ^[20]
		静态	土地用途	SVM ^[47]
	AOI	动态	情绪	CNN ^[126] 、关键词匹配 ^[113]
众包街景	路网	静态	可通行类型	改进的 YOLOv3 模型 ^[83]
	POI	静态	交通标识类别	RetinaNet ^[127]

对众包时空数据的简单统计揭示了如下信息:轨迹数量、移动对象数量可以反映地理实体的热度^[1,2,5,6]和各时段的流量^[3,4];轨迹中包含的移动对象的类型可以得出路段可通行移动对象类型^[117];轨迹的平均速度可反映路段的限速^[117]和各时段的车速^[22,31,121];车辆在加油站驻留时长的均值可以反映各时段的加油等待时长期望^[52,53]等等。Wu 等^[25]在共享单车上装备了便携式空气质量检测装置感知不同时段不同区域的空气质量。在多个观测值融合时,对于有污染源的区域采用高

斯平均估计区域的空气质量,对于没有污染源的区域采用平均值估计空气质量。值得一提的是,通过众包轨迹数据(如出租车轨迹)得到的流量通常是该路段真实路段的有偏采样,因此也有一些工作也聚焦在基于众包轨迹得到的流量去推测路段的真实流量。如 Shang 等^[122]提出了一个贝叶斯网络模型,刻画了真实流量和观测流量的依赖关系,用以推测真实流量。

除了这些直接获得的统计属性,众包时空数据还能够间接反映地理实体的一些统计属性,研究人员一

般通过建立统计模型推测这些属性。

地理实体类别是一个较为典型的需要通过建模得到的地理实体统计属性。如Ding等^[120]从出租车轨迹和路网拓扑结构两方面提取了多种特征,分别输入到逻辑回归和朴素贝叶斯两个分类器,再通过层叠泛化机制推测了道路类别。Fang等^[82]基于多种车辆的轨迹数据中的移动性特征,训练SVM分类器推测道路类别。Ding等^[83]利用众包街景数据提出了通过提取照片中的交通标识的方法补全道路可通行类型。由于交通标识一般较小,Ding等^[83]移除了YOLOv3模型中13×13大目标预测分支用于识别交通标识。Zhu等^[106]基于危化品车辆的驻留频次、时长、上下游关系等信息,利用XG-Boost推测POI是否为违规小化工厂。Kravi等^[20]将关联到同一POI上的Twitter拼接并抽取多种文本特征,然后利用逻辑回归对POI进行分类。Pedersen等^[127]利用RetinaNet目标检测网络检测众包街景图片中交通标识的类型。AOI的类别更常被称为土地用途。如Pan等^[125]基于出租车轨迹的上下客、周内周末的差异等信息,训练了SVM模型用于土地用途分类。Leung等^[47]从关联在地理实体上的Flickr文档中抽取了视觉特征和文本特征,同样训练了SVM模型用于土地用途分类。

除了地理实体类别之外,从路网角度,为得到路段的车道数信息,Zhang等^[118]利用高斯分布拟合轨迹在道路中心线附近的分布,并基于高斯分布的方差推测车道数;Chen等^[119]假设轨迹在道路中心线附近的分布服从高斯混合模型,并认为能让数据拟合误差最小的高斯分量个数为车道数。同样利用轨迹在道路中心线附近的分布信息,He等^[32]基于共享单车的轨迹距离道路中心线的偏移,利用KS检验判断道路上某时段是否出现了违章停车现象。从POI角度,Zheng等^[111]利用行人和地点的交互信息,通过基于超链接的主题搜索(Hyperlink-Induced Topic Search, HITS)算法,得到地点的有趣程度。从AOI角度,Wang等^[124]基于危化品车辆的到访和区域人群密度,利用图模式挖掘的方法发现潜在风险区域;He等^[123]基于轨迹数据、卫星图像和现有路网,利用卷积神经网络(Convolutional Neural Networks, CNN)推测交通事故高风险区域。此外,Zhu等^[126]和Doytsher等^[113]等还分别基于Flickr照片和Twitter文档分析了区域不同时段的情绪变化。

3.4.2 缺失值填补

由于众包时空数据在时空上高度倾斜的特点,某时段内某地理实体数据缺失的现象非常常见。在进一步提升众包时空数据感知地理实体的能力方面,缺失值填补起到了关键作用。

对于道路速度的填补是一个典型的案例^[22,31,121,122]。Shang等^[122]提出了上下文感知的基于矩阵分解的协同

过滤算法对路段速度进行填补,该算法对时段-路段构成的速度矩阵、时间-网格构成的流量矩阵、路段-特征矩阵以及历史均值矩阵进行联合矩阵分解,从而填补速度缺失值。Yi等^[31]利用同时考虑时间临近性、周期性的自注意力机制填补速度缺失值。Qin等^[121]提出了自利联合学习的方法,利用重建器填补速度,并利用判别器保证推测结果的可解释性。不同于Yi等^[31]仅考虑历史时段当前道路的速度用于插值,Xu等^[22]提出了一个基于时空注意力机制的插值网络,并且通过图像中循环一致性的思想对模型进行训练,确保生成结果的鲁棒性。该方法中最核心的基于时空注意力机制的插值模块如图7所示该方法能够综合考虑当前时段不同道路、历史时段当前道路、历史时段其余道路的速度读数来填补当前路段当前时段缺失的速度信息。

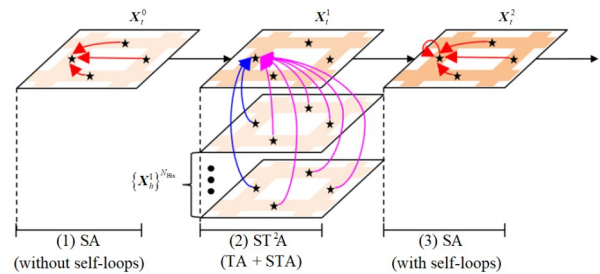


图7 Xu等^[22]提出的基于时空注意力机制的插值模块

相似地,为了填补时空上缺失的加油等待时长期望,Zhang等^[52,53]构建了加油站-天-小时构成的加油等待时长期望张量,利用上下文感知的高阶张量分解方法进行缺失值填补,在张量分解时考虑了交通、地理等特征提升填补的准确性;为了填补时空上缺失的空气质量数据,Wu等^[25]利用贝叶斯压缩感知技术推测缺失的空气质量读数。

4 未来研究方向

虽然利用众包时空数据收集地理信息的思想最早可以追溯到近20年前^[85],但因数据稀缺、基础设施建设不完备的原因,直到移动互联网和工业数字化浪潮出现后才真正开始吸引研究人员的广泛关注。随着传感器设备的进一步升级完善、自动驾驶领域蓬勃的发展,基于众包时空数据推测城市地理信息的方式逐渐获得业界的认可^[128,129]。基于当前研究现状,本文从数据可用性、地理实体建模、地理信息整合三个角度展望未来的研究方向。

4.1 数据可用性

从数据可用性角度,多源数据融合、激励机制设计、隐私保护为三个值得深入探索的内容。

(1)多源数据融合。如第2.5节提到的,众包时空数据具有时空分布高度倾斜的特点,如果某区域某时段

无任何众包时空数据,很多地理信息收集的工作无法开展.为缓解数据时空稀疏性对地理信息收集带来的挑战,除了第3.4.2节提到利用缺失值填补模型对属性进行填补外,多源数据融合也是一种途径.Wu等^[94]和Yang等^[96]已就卫星图像和轨迹数据融合生成路网做了初步探索.未来更多类型的时空数据(如激光雷达等三维数据)的融合、更多类型的地理实体信息推测有待研究,如融合轨迹数据和卫星图像,联合识别危险小化工厂等等.此外,随着无人车、无人机等无人设备的普及,以人机协同的方式实现数据的收集也是可行途径.

(2)激励机制设计.同样为了缓和数据倾斜的问题,除了参与式感知模式下平台分配任务外,设计激励机制让人群去收集数据缺失的区域和时间段的数据也是解决数据倾斜的问题的途径之一.如Wu等^[25]设计了一红包机制鼓励用户收集当前时段某区域的数据.红包的金额随着该区域信息增益以及距上一个观测值的时间跨度而变化.但是奖励金额的设置本质是一个较为复杂的问题,受用户心理、任务难度、天气情况、交通状况等因素的影响,如何设置一个合理的激励金额有待进一步研究.

(3)隐私保护.虽然众包时空数据驱动的地理信息推测是基于移动群体的聚合信息实现,不依赖单个移动对象的信息,但是现阶段个体的原始时空活动序列仍然需要作为各种地理信息推测方法的输入,存在隐私泄露的风险.隐私保护在基于位置的服务、众包时空数据的任务分配中已经存在不少研究^[130,131],然而面向地理信息推测的众包时空数据隐私保护研究较少.Chen等^[132]探索了在轨迹数据驱动路网生成中保护轨迹数据上传者的隐私.但为了隐私保护需要,路网生成算法的输入仅包含离散的时间无序的空间点集,丢失了对于路网生成来说非常关键的轨迹序列信息.如何在确保隐私的情况下,推测更高质量的地理信息值得继续探索.

4.2 地理实体建模

从地理实体建模角度,更复杂的地理实体位置表示、三维建模的需求在逐渐增多.

(1)更复杂的地理实体位置表示.现有研究中大多数对POI的表示还是以二维空间点为主,虽然能够满足部分应用需求,但是真实世界中POI在空间上往往会占据一定的面积,而且因为不同的建筑设计风格,有些POI的形状可能是不规则的.因此,以单个二维点表示是不够精确的,这会致使下游电子地理围栏等一些应用包含关系判定不准确的问题.由于采样频率较低,利用时空活动序列前后点连线勾勒地理实体边界信息较为困难.虽然DBSCAN等基于密度聚类的方法能够

带有定位噪声的离散时空活动记录中得到不规则的区域^[106],但是其效果受参数影响较大,且难以为所有地点设置统一的参数.利用监督学习建模驻留点空间分布,同时结合卫星遥感数据识别POI的边界信息,是潜在的优化方向.

(2)三维地理实体建模.三维地图能够带来更为丰富的信息,能够为自动驾驶、蜂窝基站的选址规划、楼房日照计算等需要依赖三维信息的应用提供支持.Rodrigues等^[116]基于GNSS设备信号的干扰情况就三维楼宇估计做了初步探索,未来随着高度信息的感知更加精准,各种众包时空数据驱动的三维地理实体生成将成为可能.

4.3 地理信息整合

从地理信息整合角度,更精准、更自动化的地理信息更新是未来的趋势.Stanojevic等^[38]和Balsebre等^[39]分别就路网的融合和POI的融合做了初步探索,但目前尚且缺乏统一的地理实体融合框架.此外,融合完成的结果还需要依大量的人工验证和编辑才可供使用,成本巨大且时效性不强.因此,地理信息的自动化可靠更新还尚待解决.

5 结束语

本文从推测地理信息角度阐述了众包时空数据挖掘相关的研究进展.首先,本文给出了众包时空数据的概念,比较了基于众包时空数据的地理信息推测与传统地理信息收集方法的优劣,提出了众包时空数据驱动的地理信息推测框架和对应的研究挑战.其次,本文以轨迹、LBSN和街景三种常见的众包时空数据为例,对地图匹配、名称提取、位置发现、统计属性推测四个具体问题的研究脉络和现状进行了介绍和讨论.最后,基于研究现状,本文从数据可用性、地理信息建模、地理信息自动化整合三个角度展望了未来研究方向,供未来从事众包时空数据驱动地理信息推测的研究人员参考.

参考文献

- [1] BAO J, HE T F, RUAN S J, et al. Planning bike lanes based on sharing-bikes' trajectories[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 1377-1386.
- [2] HE T F, BAO J, RUAN S J, et al. Interactive bike lane planning using sharing bikes' trajectories[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(8): 1529-1542.
- [3] PAN Z Y, ZHANG W T, LIANG Y X, et al. Spatio-tempo-

- ral meta learning for urban traffic prediction[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 34(3): 1462-1476.
- [4] ZHANG J B, ZHENG Y, QI D K. Deep spatio-temporal residual networks for citywide crowd flows prediction[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, 31(1): 1655-1661.
- [5] LI Y H, BAO J, LI Y H, et al. Mining the most influential k -location set from massive trajectories[C]//*Proceedings of the 24th ACM SIGSpatial International Conference on Advances in Geographic Information Systems*. New York: ACM, 2016: 1-4.
- [6] ZHANG P, BAO Z F, LI Y C, et al. Trajectory-driven influential billboard placement[C]//*Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York: ACM, 2018: 2748-2757.
- [7] RUAN S J, BAO J E, LIANG Y X, et al. Dynamic public resource allocation based on human mobility prediction[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2020, 4(1): 1-22.
- [8] HU Y, RUAN S J, NI Y T, et al. SALON: A universal stay point-based location analysis platform[C]//*Proceedings of the 29th International Conference on Advances in Geographic Information Systems*. New York: ACM, 2021: 407-410.
- [9] 王家耀, 魏海平, 成毅, 等. 时空 GIS 的研究与进展[J]. *海洋测绘*, 2004, 24(5): 1-4.
WANG J Y, WEI H P, CHENG Y, et al. The research and development of spatio-temporal GIS[J]. *Hydrographic Surveying and Charting*, 2004, 24(5): 1-4. (in Chinese)
- [10] 李德仁, 陈小明, 郭丙轩, 等. 车载 GPS 道路信息采集和更新系统研究[J]. *武汉测绘科技大学学报*, 2000, 25(2): 95-99.
LI D R, CHEN X M, GUO B X, et al. Study on car-borne GPS road information collecting and update system[J]. *Journal of Wuhan Technical University of Surveying and Mapping*, 2000, 25(2): 95-99. (in Chinese)
- [11] 王龙波. 基于 GPS 与实景影像的 POI 快速采集技术[J]. *中国科技信息*, 2007(12): 121-122.
WANG L B. Fast POI collecting technologies based on GPS and real-scene imagery[J]. *China Science and Technology Information*, 2007(12): 121-122. (in Chinese)
- [12] 皮晓亮, 王正, 韩皓, 等. 基于环形线圈检测器采集信息的交通状态分类方法应用研究[J]. *公路交通科技*, 2006, 23(4): 115-119.
PI X L, WANG Z, HAN H, et al. Application research of traffic state classification method based on collected information from loop detector[J]. *Journal of Highway and Transportation Research and Development*, 2006, 23(4): 115-119. (in Chinese)
- [13] 童咏昕, 袁野, 成雨蓉, 等. 时空众包数据管理技术研究综述[J]. *软件学报*, 2017, 28(1): 35-58.
TONG Y X, YUAN Y, CHENG Y R, et al. Survey on spatiotemporal crowdsourced data management techniques[J]. *Journal of Software*, 2017, 28(1): 35-58. (in Chinese)
- [14] CAMPBELL A T, EISENMAN S B, LANE N D, et al. People-centric urban sensing[C]//*Proceedings of the 2nd Annual International Workshop on Wireless Internet*. New York: ACM, 2006: 18-es.
- [15] 陈荟慧, 郭斌, 於志文. 移动群智感知应用[J]. *中兴通讯技术*, 2014, 20(1): 35-37.
CHEN H H, GUO B, YU Z W. Mobile crowd-sensing application[J]. *ZTE Technology Journal*, 2014, 20(1): 35-37. (in Chinese)
- [16] CAPPONI A, FIANDRINO C, KANTARCI B, et al. A survey on mobile crowdsensing systems: Challenges, solutions, and opportunities[J]. *IEEE Communications Surveys & Tutorials*, 2019, 21(3): 2419-2465.
- [17] GOODCHILD M F. Citizens as sensors: The world of volunteered geography[J]. *GeoJournal*, 2007, 69(4): 211-221.
- [18] ZHENG Y, CAPRA L, WOLFSON O, et al. Urban computing: Concepts, methodologies, and applications[J]. *ACM Transactions on Intelligent Systems and Technology*, 2014, 5(3): 38.
- [19] ZHAO L S, MAO J L, PU M, et al. Automatic calibration of road intersection topology using trajectories[C]//*2020 IEEE 36th International Conference on Data Engineering (ICDE)*. Piscataway: IEEE, 2020: 1633-1644.
- [20] KRAVI E, KANZA Y, KIMELFELD B, et al. Geosocial location classification: Associating type to places based on geotagged social-media posts[C]//*Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. New York: ACM, 2020: 167-170.
- [21] SONG Y T, LI J W, CHEN L Y, et al. A semantic segmentation based POI coordinates generating framework for on-demand food delivery service[C]//*Proceedings of the 29th International Conference on Advances in Geographic Information Systems*. New York: ACM, 2021: 379-388.

- [22] XU Q X, RUAN S J, LONG C, et al. Traffic speed imputation with spatio-temporal attentions and cycle-perceptual training[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. New York: ACM, 2022: 2280-2289.
- [23] ZHAO Y, LI Y, WANG Y, et al. Destination-aware task assignment in spatial crowdsourcing[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York: ACM, 2017: 297-306.
- [24] JI S G, ZHENG Y, LI T R. Urban sensing based on human mobility[C]//Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. New York: ACM, 2016: 1040-1051.
- [25] WU D, XIAO T, LIAO X W, et al. When sharing economy meets IoT[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020, 4(2): 1-26.
- [26] CHAO P F, HUA W, MAO R, et al. A survey and quantitative study on map inference algorithms from GPS trajectories[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 15-28.
- [27] SILVA T H, VIANA A C, BENEVENUTO F, et al. Urban computing leveraging location-based social network data[J]. ACM Computing Surveys, 2020, 52(1): 1-39.
- [28] 唐炉亮, 赵紫龙, 杨雪, 等. 大数据环境下道路场景高时空分辨率众包感知方法[J]. 测绘学报, 2022, 51(6): 1070-1090.
- TANG L L, ZHAO Z L, YANG X, et al. Road crowdsensing with high spatio-temporal resolution in big data era[J]. Acta Geodaetica et Cartographica Sinica, 2022, 51(6): 1070-1090. (in Chinese)
- [29] Data West Research Agency. A Practitioner's Guide to GIS Terminology: A Glossary of Geographic Information System Terms[M]. University Place: Data West Research Agency, 2000.
- [30] 刘海, 郝克刚. 软件缺陷数据的分析方法及其实现[J]. 计算机科学, 2008, 35(8): 262-264.
- LIU H, HAO K G. Method of software defect data analysis and its implementation[J]. Computer Science, 2008, 35(8): 262-264. (in Chinese)
- [31] YI X W, DUAN Z W, LI T, et al. CityTraffic: Modeling citywide traffic via neural memorization and generalization approach[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM, 2019: 2665-2671.
- [32] HE T F, BAO J, LI R Y, et al. Detecting vehicle illegal parking events using sharing bikes' trajectories[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 340-349.
- [33] KWON J, VARAIYA P, SKABARDONIS A. Estimation of truck traffic volume from single loop detectors with lane-to-lane speed correlation[J]. Transportation Research Record: Journal of the Transportation Research Board, 2003, 1856(1): 106-117.
- [34] DEMIR I, KOPERSKI K, LINDENBAUM D, et al. DeepGlobe 2018: A challenge to parse the earth through satellite images[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2018: 172-17209.
- [35] PAOLA J D, SCHOWENGERDT R A. A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 1995, 33(4): 981-996.
- [36] HE S T, BASTANI F, JAGWANI S, et al. RoadTagger: Robust road attribute inference with graph neural networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 10965-10972.
- [37] KHALIL M, LI J P, SHARIF A, et al. Traffic congestion detection by use of satellites view[C]//2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). Piscataway: IEEE, 2018: 278-280.
- [38] STANOJEVIC R, ABBAR S, THIRUMURUGANATHAN S, et al. Road network fusion for incremental map updates[C]//LBS 2018: 14th International Conference on Location Based Services. Cham: Springer, 2018: 91-109.
- [39] BALSEBRE P, YAO D Z, CONG G, et al. Geospatial entity resolution[C]//Proceedings of the ACM Web Conference 2022. New York: ACM, 2022: 3061-3070.
- [40] YUAN J, ZHENG Y, ZHANG C Y, et al. An interactive-voting based map matching algorithm[C]//2010 Eleventh International Conference on Mobile Data Management. Piscataway: IEEE, 2010: 43-52.
- [41] RUAN S J, LONG C, MA Z P, et al. Service time prediction for delivery tasks via spatial meta-learning[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2022: 3829-3837.
- [42] RUAN S J, LI R Y, BAO J, et al. CloudTP: A cloud-based flexible trajectory preprocessing framework[C]//

- 2018 IEEE 34th International Conference on Data Engineering (ICDE). Piscataway: IEEE, 2018: 1601-1604.
- [43] ZHENG Y. Trajectory data mining: An overview[J]. *ACM Transactions on Intelligent Systems and Technology*, 6(3): 29.
- [44] ZHENG Y, XIE X, MA W Y. GeoLife: A collaborative social networking service among user, location and trajectory[J]. *IEEE Data Engineering Bulletin*, 2010, 33: 32-39.
- [45] LI Q N, ZHENG Y, XIE X, et al. Mining user similarity based on location history[C]//*Proceedings of the 16th ACM SIGSpatial International Conference on Advances in Geographic Information Systems*. New York: ACM, 2008: 1-10.
- [46] YUAN J, ZHENG Y, ZHANG L H, et al. Where to find my next passenger[C]//*Proceedings of the 13th International Conference on Ubiquitous Computing*. New York: ACM, 2011: 109-118.
- [47] LEUNG D, NEWSAM S. Exploring Geotagged images for land-use classification[C]//*Proceedings of the ACM Multimedia 2012 Workshop on Geotagging and Its Applications in Multimedia*. New York: ACM, 2012: 3-8.
- [48] ALVARES L O, BOGORNY V, KUIJPERS B, et al. A model for enriching trajectories with semantic geographical information[C]//*Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*. New York: ACM, 2007: 1-8.
- [49] CHOI D W, PEI J A, HEINIS T. Efficient mining of regional movement patterns in semantic trajectories[J]. *Proceedings of the VLDB Endowment*, 2017, 10(13): 2073-2084.
- [50] BHATTACHARYA T, KULIK L, BAILEY J. Automatically recognizing places of interest from unreliable GPS data using spatio-temporal density estimation and line intersections[J]. *Pervasive and Mobile Computing*, 2015, 19: 86-107.
- [51] KELES I, SCHUBERT M, KRÖGER P, et al. Extracting visited points of interest from vehicle trajectories[C]//*Proceedings of the 4th International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data*. New York: ACM, 2017: 1-6.
- [52] ZHANG F Z, WILKIE D, ZHENG Y, et al. Sensing the pulse of urban refueling behavior[C]//*Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. New York: ACM, 2013: 13-22.
- [53] ZHANG F Z, YUAN N J, WILKIE D, et al. Sensing the pulse of urban refueling behavior: A perspective from taxi mobility[J]. *ACM Transactions on Intelligent Systems and Technology*, 2015, 6(3): 1-23.
- [54] WU H, SUN W W, ZHENG B H. Is only one GPS position sufficient to locate You to the road network accurately? [C]//*Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. New York: ACM, 2016: 740-751.
- [55] NEWSON P, KRUMM J. Hidden Markov map matching through noise and sparseness[C]//*Proceedings of the 17th ACM SIGSpatial International Conference on Advances in Geographic Information Systems*. New York: ACM, 2009: 336-343.
- [56] LOU Y, ZHANG C Y, ZHENG Y, et al. Map-matching for low-sampling-rate GPS trajectories[C]//*Proceedings of the 17th ACM SIGSpatial International Conference on Advances in Geographic Information Systems*. New York: ACM, 2009: 352-361.
- [57] ZHAO K, FENG J, XU Z, et al. DeepMM: Deep learning based map matching with data augmentation[C]//*Proceedings of the 27th ACM SIGSpatial International Conference on Advances in Geographic Information Systems*. New York: ACM, 2019: 452-455.
- [58] REN H M, RUAN S J, LI Y H, et al. MTrajRec: Map-constrained trajectory recovery via Seq2Seq multi-task learning[C]//*Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. New York: ACM, 2021: 1410-1419.
- [59] YAN Z X, CHAKRABORTY D, PARENT C, et al. SeMiTri: A framework for semantic annotation of heterogeneous trajectories[C]//*Proceedings of the 14th International Conference on Extending Database Technology*. New York: ACM, 2011: 259-270.
- [60] NISHIDA K, TODA H, KURASHIMA T, et al. Probabilistic identification of visited point-of-interest for personalized automatic check-in[C]//*Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. New York: ACM, 2014: 631-642.
- [61] SUZUKI J, SUHARA Y, TODA H, et al. Personalized visited-POI assignment to individual raw GPS trajectories [J]. *ACM Transactions on Spatial Algorithms and Systems*, 5(3): 16.
- [62] RUAN S J, FU X, LONG C, et al. Filling delivery time automatically based on couriers' trajectories[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(2): 1528-1540.

- [63] JONES C B, PURVES R S, CLOUGH P D, et al. Modeling vague places with knowledge from the Web[J]. *International Journal of Geographical Information Science*, 2008, 22(10): 1045-1065.
- [64] LI L N, GOODCHILD M F. Constructing places from spatial footprints[C]//*Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. New York: ACM, 2012: 15-21.
- [65] HOLLENSTEIN L, PURVES R. Exploring place through user-generated content: Using Flickr tags to describe city cores[J]. *Journal of Spatial Information Science*, 2010(1): 21-48.
- [66] RATTENBURY T, NAAMAN M. Methods for extracting place semantics from Flickr tags[J]. *ACM Transactions on the Web*, 2009, 3(1): 1-30.
- [67] BAEZA-YATES R, RIBEIRO-NETO B. *Modern Information Retrieval*[M]. Austin: Addison-Wesley Longman, 1999.
- [68] LLOYD S. Least squares quantization in PCM[J]. *IEEE Transactions on Information Theory*, 1982, 28(2): 129-137.
- [69] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-08-09)[2023-02-15]. <https://arxiv.org/abs/1508.01991>.
- [70] ABID N, UL HASAN A, SHAFAIT F. DeepParse: A trainable postal address parser[C]//*2018 Digital Image Computing: Techniques and Applications (DICTA)*. Piscataway: IEEE, 2019: 1-8.
- [71] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2023-02-15]. <https://arxiv.org/abs/1810.04805>.
- [72] HE T F, CHEN G C, MENG C S, et al. POI alias discovery in delivery addresses using user locations[C]//*Proceedings of the 29th International Conference on Advances in Geographic Information Systems*. New York: ACM, 2021: 225-228.
- [73] GAO S, LI L N, LI W W, et al. Constructing gazetteers from volunteered big geo-data based on Hadoop[J]. *Computers, Environment and Urban Systems*, 2017, 61: 172-186.
- [74] INTAGORN S, LERMAN K. Learning boundaries of vague places from noisy annotations[C]//*Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York: ACM, 2011: 425-428.
- [75] KEßLER C, MAUÉ P, HEUER J T, et al. Bottom-up gazetteers: Learning from the implicit semantics of geotags [C]//*International Conference on GeoSpatial Semantics*. Berlin: Springer, 2009: 83-102.
- [76] RUAN S J, XIONG Z, LONG C, et al. Doing in one go: Delivery time inference based on couriers' trajectories [C]//*Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York: ACM, 2020: 2813-2821.
- [77] SRIVASTAVA V, TEJASWIN P, DHAKAD L, et al. A geocoding framework powered by delivery data[C]//*Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. New York: ACM, 2020: 568-577.
- [78] FORMAN G. Getting your package to the right place: Supervised machine learning for geolocation[C]//*Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Cham: Springer, 2021: 403-419.
- [79] RUAN S J, LONG C, YANG X D, et al. Discovering actual delivery locations from mis-annotated couriers' trajectories[C]//*2022 IEEE 38th International Conference on Data Engineering (ICDE)*. Piscataway: IEEE, 2022: 3241-3253.
- [80] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//*Proceedings of the Advances in Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 2017: 5998-6008.
- [81] DAHIYA M, SAMATIA D, RUSTOGI K. Learning locality maps from noisy geospatial labels[C]//*Proceedings of the 35th Annual ACM Symposium on Applied Computing*. New York: ACM, 2020: 601-608.
- [82] FANG Z H, WANG G A, XIE X Y, et al. Urban map inference by pervasive vehicular sensing systems with complementary mobility[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021, 5(1): 1-24.
- [83] DING X, FAN H C, GONG J Y. Towards generating network of bikeways from Mapillary data[J]. *Computers, Environment and Urban Systems*, 2021, 88: 101632.
- [84] BIAGIONI J, ERIKSSON J. Inferring Road maps from global positioning system traces[J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2012, 2291(1): 61-71.
- [85] EDELKAMP S, SCHRÖDL S. Route planning and map inference with global positioning traces[M]//*Lecture*

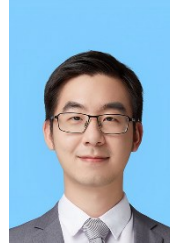
- Notes in Computer Science. Berlin: Springer, 2003: 128-151.
- [86] CHEN C, LU C W, HUANG Q X, et al. City-scale map creation and updating using GPS collections[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1465-1474.
- [87] STANOJEVIC R, ABBAR S, THIRUMURUGANATHAN S, et al. Robust road map inference through network alignment of trajectories[C]//Proceedings of the 2018 SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics, 2018: 135-143.
- [88] CAO L L, KRUMM J. From GPS traces to a routable road map[C]//Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2009: 3-12.
- [89] NIEHOEFER B, BURDA R, WIETFELD C, et al. GPS community map generation for enhanced routing methods based on trace-collection by mobile phones[C]//2009 First International Conference on Advances in Satellite and Space Communications. Piscataway: IEEE, 2009: 156-161.
- [90] BIAGIONI J, ERIKSSON J. Map inference in the face of noise and disparity[C]//Proceedings of the 20th International Conference on Advances in Geographic Information Systems. New York: ACM, 2012: 79-88.
- [91] WANG S Y, WANG Y S, LI Y J. Efficient map reconstruction and augmentation via topological methods[C]//Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2015: 1-10.
- [92] YANG X E, TANG L L, REN C, et al. Pedestrian network generation based on crowdsourced tracking data[J]. International Journal of Geographical Information Science, 2020, 34(5): 1051-1074.
- [93] HE S T, BASTANI F, ABBAR S, et al. RoadRunner: Improving the precision of road network inference from GPS trajectories[C]//Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2018: 3-12.
- [94] WU H, ZHANG H Y, ZHANG X Y, et al. Deep Dual Mapper: A gated fusion network for automatic map extraction using aerial images and trajectories[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 1037-1045.
- [95] RUAN S J, LONG C, BAO J E, et al. Learning to generate maps from trajectories[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 890-897.
- [96] YANG J Z, YE X Q, WU B, et al. DuARE: Automatic road extraction with aerial images and trajectory data at Baidu maps[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2022: 4321-4331.
- [97] WANG Y, LIU X M, WEI H, et al. CrowdAtlas: Self-updating maps for cloud and personal use[C]//MobiSys' 13: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services. New York: ACM, 2013: 469-470.
- [98] WU H, TU C C, SUN W W, et al. GLUE: A parameter-tuning-free map updating system[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2015: 683-692.
- [99] SHAN Z Q, WU H, SUN W W, et al. COBWEB: A robust map update system using GPS trajectories[C]//Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. New York: ACM, 2015: 927-937.
- [100] CAO C, LIU Z D, LI M, et al. Walkway discovery from large scale crowdsensing[C]//2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). Piscataway: IEEE, 2018: 13-24.
- [101] SIBSON R. SLINK: An optimally efficient algorithm for the single-link cluster method[J]. The Computer Journal, 1973, 16(1): 30-34.
- [102] KEGL B, KRZYZAK A, LINDER T, et al. Learning and design of principal curves[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(3): 281-297.
- [103] WANG T, MAO J L, JIN C Q. HyMU: A hybrid map updating framework[C]//International Conference on Database Systems for Advanced Applications. Cham: Springer, 2017: 19-33.
- [104] LEE J G, HAN J W, WHANG K Y. Trajectory clustering: A partition-and-group framework[C]//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2007: 593-604.

- [105] CHAO P F, HUA W, ZHOU X F. Trajectories know where map is wrong: An iterative framework for map-trajectory co-optimisation[J]. *World Wide Web*, 2020, 23(1): 47-73.
- [106] ZHU Z, REN H M, RUAN S J, et al. ICFinder: A ubiquitous approach to detecting illegal hazardous chemical facilities with truck trajectories[C]//*Proceedings of the 29th International Conference on Advances in Geographic Information Systems*. New York: ACM, 2021: 37-40.
- [107] ASHBROOK D, STARNER T. Learning significant locations and predicting user movement with GPS[C]//*Proceedings of Sixth International Symposium on Wearable Computers*. Piscataway: IEEE, 2003: 101-108.
- [108] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//*Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. New York: AAAI Press, 1996: 226-231.
- [109] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. Optics: Ordering points to identify the clustering structure[C]//*Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. New York: ACM, 1999: 49-60.
- [110] ASSAM R, SEIDL T. Context-based location clustering and prediction using conditional random fields[C]//*Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia*. New York: ACM, 2014: 1-10.
- [111] ZHENG Y, ZHANG L Z, XIE X, et al. Mining interesting locations and travel sequences from GPS trajectories [C]//*Proceedings of the 18th International Conference on World Wide Web*. New York: ACM, 2009: 791-800.
- [112] HU Y J, GAO S, JANOWICZ K, et al. Extracting and understanding urban areas of interest using geotagged photos[J]. *Computers, Environment and Urban Systems*, 2015, 54: 240-254.
- [113] DOYTSHER Y, GALON B, KANZA Y. Emotion maps based on geotagged posts in the social media[C]//*Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. New York: ACM, 2017: 39-46.
- [114] ZHENG V W, ZHENG Y, XIE X, et al. Collaborative location and activity recommendations with GPS history data[C]//*Proceedings of the 19th International Conference on World Wide Web*. New York: ACM, 2010: 1029-1038.
- [115] WARD J H Jr. Hierarchical grouping to optimize an objective function[J]. *Journal of the American Statistical Association*, 1963, 58(301): 236-244.
- [116] RODRIGUES J G P, AGUIAR A. Extracting 3D maps from crowdsourced GNSS skyview data[C]//*MobiCom' 19: The 25th Annual International Conference on Mobile Computing and Networking*. New York: ACM, 2019: 1-15.
- [117] VAN WINDEN K. Automatically Deriving and Updating Attribute Road Data from Movement Trajectories [D]. *Mekelweg: Technische Universiteit Delft*, 2014.
- [118] ZHANG L J, THIEMANN F, SESTER M. Integration of GPS traces with road map[C]//*Proceedings of the Third International Workshop on Computational Transportation Science*. New York: ACM, 2010: 17-22.
- [119] CHEN Y H, KRUMM J. Probabilistic modeling of traffic lanes from GPS traces[C]//*Proceedings of the 18th SIGSpatial International Conference on Advances in Geographic Information Systems*. New York: ACM, 2010: 81-88.
- [120] DING Y, ZHENG J C, TAN H Y, et al. Inferring Road type in crowdsourced map services[C]//*International Conference on Database Systems for Advanced Applications*. Cham: Springer, 2014: 392-406.
- [121] QIN H L, ZHAN X Y, LI Y X, et al. Network-wide traffic states imputation using self-interested coalitional learning[C]//*Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. New York: ACM, 2021: 1370-1378.
- [122] SHANG J B, ZHENG Y, TONG W Z, et al. Inferring gas consumption and pollution emission of vehicles throughout a city[C]//*Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2014: 1027-1036.
- [123] HE S T, SADEGHI M A, CHAWLA S, et al. Inferring high-resolution traffic accident risk maps based on satellite imagery and GPS trajectories[C]//*2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway: IEEE, 2022: 11957-11965.
- [124] WANG J Y, CHEN C, WU J J, et al. No longer sleeping with a bomb: A duet system for protecting urban safety from dangerous goods[C]//*Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining. New York: ACM, 2017: 1673-1681.

- [125] PAN G, QI G D, WU Z H, et al. Land-use classification using taxi GPS traces[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(1): 113-123.
- [126] ZHU Y, NEWSAM S. Spatio-temporal sentiment hot-spot detection using geotagged photos[C]//Proceedings of the 24th ACM SIGSPatial International Conference on Advances in Geographic Information Systems. New York: ACM, 2016: 1-4.
- [127] PEDERSEN K F, TORP K. Geolocating traffic signs using crowd-sourced imagery[C]//Proceedings of the 28th International Conference on Advances in Geographic Information Systems. New York: ACM, 2020: 199-202.
- [128] 张国龙. 高精定位与高精地图众包更新现状、挑战[EB/OL]. (2021-10-01) [2022-08-08]. <https://www.163.com/dy/article/GL7IU5F805278R4J.html>.
ZHANG G L. High precision positioning and crowd-sourcing high precision maps updates: Current status and challenges[EB/OL]. (2021-10-01) [2022-08-08]. <https://www.163.com/dy/article/GL7IU5F805278R4J.html>. (in Chinese)
- [129] 李勤. 中海庭CTO罗跃军谈高精地图: 众包更新是必须,未来2-3年会落地[EB/OL]. (2021-04-25) [2022-08-08]. <https://auto-time.36kr.com/p/1196970721626626>.
LI Q. CTO of zhonghaiting, talks about high precision map: Crowd-sourcing is a must, which would be in real use in the next 2-3 years[EB/OL]. (2021-04-25) [2022-08-08]. <https://auto-time.36kr.com/p/1196970721626626>. (in Chinese)
- [130] JIANG H B, LI J, ZHAO P, et al. Location privacy-preserving mechanisms in location-based services: A comprehensive survey[J]. ACM Computing Surveys, 2021, 54(1): 1-36.
- [131] WANG H Z, WANG E, YANG Y J, et al. Privacy-preserving online task assignment in spatial crowdsourcing: A graph-based approach[C]//IEEE INFOCOM 2022 - IEEE Conference on Computer Communications. Piscataway: IEEE, 2022: 570-579.
- [132] CHEN X, WU X P, LI X Y, et al. Privacy-preserving high-quality map generation with participatory sensing [C]//IEEE INFOCOM 2014 - IEEE Conference on Computer Communications. Piscataway: IEEE, 2014: 2310-2318.

作者简介



阮思捷 男,1994年12月出生于浙江省杭州市。现为北京理工大学计算机学院特别副研究员。主要研究方向为城市计算与时空数据挖掘。曾获ACM SIGSPATIAL中国分会优博奖。在国内外顶尖期刊和会议上发表学术论文30余篇。

E-mail: sjruan@bit.edu.cn



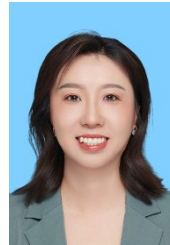
熊可钦 女,2002年4月出生于江西省抚州市。现为北京理工大学本科生。专业为计算机科学与技术。

E-mail: keqin0127x@163.com



王树良(通讯作者) 男,1974年7月生,山东梁山人。现为北京理工大学教授、博士生导师,电子政务研究院执行院长,中国制造企业双创发展联盟副理事长,国家重点研发计划项目首席科学家。主要研究方向为空间数据挖掘与社会智能治理。获国家科技进步一等奖等。

E-mail: slwang2011@bit.edu.cn



耿晶 女,1986年2月出生于山东省威海市。现为北京理工大学计算机学院特别副研究员。主要研究方向为地理信息系统、时空数据分析、大数据智能挖掘、电子政务相关理论。

E-mail: janegeng@bit.edu.cn



鲍捷 男,1985年5月生,浙江金华人。美国明尼苏达大学计算机系博士。现为京东科技研发总监。主要研究方向为时空大数据管理与挖掘。曾在贵阳、上海以及北京落地多个智慧城市项目,并在世界顶尖期刊和会议上发表学术论文60余篇。

E-mail: baojie@jd.com



郑宇 男,1975年10月生,湖南衡阳人。现为京东集团副总裁、京东科技首席数据科学家,上海交通大学讲座教授,南京大学、香港科技大学等多所知名高校的客座教授和博士生导师,国家重点研发计划项目首席科学家。他是城市计算领域的先驱和奠基人,也是大数据和人工智能领域的领军人物和实践者,被评为IEEE Fellow。

E-mail: msyuzheng@outlook.com