DOI:10.13203/j.whugis20130633　　文章编号:1671-8860(2014)04-0390-07

众源地理数据处理与分析方法探讨

单　杰[1,2]　秦　昆[1]　黄长青[1]　胡翔云[1]　余　洋[1]
胡庆武[1]　林志勇[1]　陈江平[1]　贾　涛[1]

1　武汉大学遥感信息工程学院,湖北 武汉,430079
2　普度大学土木工程学院,美国拉斐特,47907

摘　要:众源地理数据是由大量非专业人员志愿获取并通过互联网向大众或相关机构提供的一种开放地理数据,是有别于传统测绘产品的一种新型地理空间数据。分析和研究了众源地理数据的概念与特点;介绍了众源地理数据的来源和获取方法;讨论了众源地理数据处理与分析的关键技术,包括众源地理数据的质量评价方法,众源地理数据的信息提取与更新方法,众源地理数据的分析与挖掘方法等;指出了众源地理数据处理与分析的研究趋势和发展方向。

关键词:众源地理数据;质量分析与评价;信息提取与更新;空间分析;空间数据挖掘;地理信息服务

中图法分类号:P208　　**文献标志码**:A

众源地理数据(crowd sourcing geographic data,CSGD)是由大量非专业人员志愿获取并通过互联网向大众或相关机构提供的一种开放地理空间数据[1-3]。用户利用智能手机、iPad、GPS 接收机等收集某一时刻的位置信息,然后借助 Web 2.0 的标注和上传功能,使得大众用户成为义务的信息提供者。代表性的众源地理数据有 GPS 路线数据(如 OpenStreeMap,简称 OSM),用户协作标注编辑的地图数据(如 Wikimapia),各类社交网站数据(如 Twitter,Facebook 等),街旁用户签到的兴趣点数据等。这些数据需经过处理才能形成规范的地理信息[4]。与传统地理信息采集和更新方式相比,来自非专业大众的众源地理数据具有现势性高、传播快、信息丰富、成本低、数据量大、质量各异、冗余而不完整、覆盖不均匀、缺少统一规范、隐私和安全难以控制等特点,成为近年来国际地理信息科学领域的研究热点[5-7]。众源地理数据可以用于应急制图、交通分析、早期预警、地图更新、犯罪分析、疾病传播分析等诸多地理空间信息服务领域[7-10]。

众源地理数据的出现与日渐增多将深刻影响现有地理信息科学的发展方向和产业化模式。众源地理数据需要解决如何与已有数据源融合进行低成本、快速、高精度的数据更新问题,其集中了普通大众的地理空间知识并依托互联网为基本平台实现传播与共享。公众对地理信息传播和共享的需求是众源地理数据产生的社会条件,同时,地理信息相关知识与技术逐渐被公众认识和了解也推动了其发展。众源地理数据中蕴含着丰富的人文社会信息和知识,需要利用空间数据分析与挖掘技术提取信息、挖掘知识。所有这些,都迫切需要发展新的众源地理数据处理与分析的理论与方法。

本文主要是针对众源地理数据处理与分析这一新兴研究方向,在分析众源地理数据的概念和特点,及其来源和获取方法的基础上,分析和讨论众源地理数据的质量分析与评价方法,信息提取与更新方法,分析与挖掘方法等,分析这些关键技术的研究现状和存在的问题,提出相应的研究思路和进一步的研究方向。

1　众源地理数据的概念与特点

1.1　众源地理数据的概念

众源地理数据是地理信息科学领域近几年出现的新概念。与此相近的概念包括 Neogeography[11], Volunteered Geographic Information

收稿日期:2013-11-08
项目来源:国家自然科学基金资助项目(61172175)。
第一作者:单杰,教授,博士生导师。近期主要研究方向为地理空间数据处理。E-mail:shanj@whu.edu.cn
通讯作者:秦昆,教授,博士,博士生导师。E-mail:qink@whu.edu.cn

(VGI)[4,12],以及 Crowdsourcing Geospatial Data[3]等。Neogeography 描述的重点在于人们创建并使用属于自己的个性化地图,并与他人分享个人位置信息,从而加强并完善与他人的交流;VGI 是指普通人在参与各种社会活动的过程中主动或无意地创建出许多地理信息;而 Crowdsourcing Geospatial Data 则将"地理信息"概念弱化成"众源"的"地理空间数据",以强调"众源"的描述对象是数据获取这一过程,而不是获取之前的数据建模或之后的数据处理。实际上,众源地理数据涵盖了空间和属性两种类型的数据,同时也涉及一些公共免费的地理数据(非普通民众创建),未能包含在 VGI 和 Neogeography 概念中。

本文讨论的众源地理数据在这些概念上有所延伸,是指在地理空间、属性和拓扑数据的获取和应用方面,从原来单纯依靠专业测绘的方式延伸到使用大众主动或被动提供的相关数据以及免费的公共数据等多种方式,以实现地理数据的快速更新和广泛应用的相关理论和方法。

1.2　众源地理数据的特点

众源地理数据具有以下 10 个特点:

1)现势性高。众源地理数据具有明显的实时更新特点,现势性高。例如,堵在路上的行车者往往会将道路拥堵信息发布于 Twitter、微博、Wikiloc、GPSies 等网站。

2)传播快。众源地理数据大多来自于互联网,借助社交网站和当地新闻等传媒系统的传播能力,进行快速传播和扩散。例如,美国加州 2009 年 5 月的 Jesusita 火灾期间,通过建立地图式火灾监视网站,迅速整合、发布了来自各种 VGI 和当地官方的实时火灾信息[7]。

3)信息丰富。众源地理数据与人类活动及社会发展紧密相关,具有丰富的社会化属性、语义信息和时序信息。其参与创建的广泛性又使得众源地理数据能从更多角度,更多方面对地理要素进行描述。

4)成本低。众源地理数据大多来自网民自发或无意采集的地理数据,其采集和处理的成本很低,极大地降低了地理信息获取和使用的成本,将更有效地促进地理信息技术的推广应用。

5)数据量大。众源地理数据大多来自互联网用户有意或无意提交的地理数据,互联网用户群的迅速发展带来了众源地理数据的激增。无论是像 OSM 这样的共享网站,还是具体的众源地理数据使用者,均需面对海量众源地理数据的高效存储以及网络共享中的快速传输等问题。

6)质量各异。众源地理数据主要由民众提供,其提供过程非常自由,参与人群非常广泛,所采用的数据采集设备精度不一,创建编辑过程中所用比例尺、采样精度不一,使得众源地理数据质量存在较大差异,甚至混杂着错误或恶意扭曲的成分。

7)冗余而又不完整。众源地理数据主要由非专业人员创建,缺乏数据完整性,难以满足一些专业的地理数据要求,同时经过多人多次提交或多次编辑的众源地理数据存在着大量冗余。

8)覆盖不均匀。众源地理数据虽然来源广泛,但是区域覆盖极其不均匀。例如 OSM 数据在英国伦敦的数据覆盖率明显高于中国湖北省的覆盖率。

9)缺少统一规范。众源地理数据来源广泛,数据格式各异,不同数据的内容不同,数据组织和存储方式也千差万别。

10)隐私与安全难以控制。自由创建和分享的众源地理数据有时会对他人及一些组织的隐私和安全问题产生影响。

2　众源地理数据的来源与获取方法

2.1　众源地理数据的来源

众源地理数据的来源主要包括:

1)公共版权数据。这一类数据多由政府部门、企业、公益组织以网站或网络服务的形式发布,例如 Google Map 网站提供的正射影像,OpenStreetMap 网站提供的交通路网等。也有一些部门和企业免费赠送的地理数据,例如 OpenStreetMap 上部分国家的主干交通数据由汽车导航数据公司 AND(Automotive Navigation Data)赠送[13]。

2)GPS 接收机数据。主要包括三类:① 应某些组织和项目请求而特意收集 GPS 数据的志愿者;② 共享自己拥有的有价值的 GPS 数据的普通人或组织;③ 相对被动、无意识上传 GPS 数据的网民,如"街旁网"用户的手机"签到"会上传 GPS 定位数据。

3)网民自发创建的地理数据。OpenStreetMap、Wikimapia 等网站向用户提供了创建地理对象的功能。部分网民出于自我满足、利他主义、兴趣或是描述周围环境等目的[14],主动地在这些网站上创建、编辑、描述各种地理对象。Google Earth 甚至允许用户对感兴趣的地物进行三维

建模。

4) Web 2.0催生的其他地理数据。Web 2.0简化了客户交互过程,出于信息共享和社交目的,部分民众积极地将自己的信息发布到网上,这些信息可能包含了地理数据。例如Flickr提供了上传照片并在地图上关联实际地理位置的功能。类似的数据源使得众源地理数据的种类更多样化、更完整。

2.2 众源地理数据的获取方法

众源地理数据的获取一般包括以下环节:

1) 下载初始化设置。包括设定下载API和登录信息,选定数据范围(包括空间范围和时间范围等)。根据研究目标,指定行政区划或区域边界坐标,或指定用户某时间段所发布的数据等,作为待获取数据的区域或范围。

2) 数据获取。利用开放的众源地理数据网站所提供的API接口,如Google Map API,Google Earth API,街旁Open API,Facebook API等,在网站所提供权限范围内,实现所选区域数据的直接读取。也可以利用网络爬虫技术设计专用的网页分析算法,从互联网上搜索并下载GPS路线数据、矢量地图数据等。

3) 数据规范化分析与转换。众源地理数据具有多源异构性,其存储格式多样、时间版本不一、坐标体系相异。合理有效地利用众源地理数据需要对其数据格式进行分析,利用文本解析、空间数据引擎等技术将众源地理数据转换为在统一存储格式、坐标体系及概念体系下表达的空间数据,并建立相应的众源地理数据表达规范。

4) 数据入库。将众源地理数据按统一规范转换后,将其导入到空间数据库中进行存储和管理。

3 众源地理数据处理与分析的关键技术

3.1 众源地理数据的质量评价方法

众源地理数据一般由缺乏足够地理信息知识和专业训练的非专业人员提供,因此存在数据质量问题,使用时需考虑其冗余性,有效性,完整性和精确性等。如何对众源地理数据的质量进行分析和评价是需要研究的首要问题[4,15]。

众源地理数据的质量是影响众源地理数据广泛应用的重要因素。Oort(2006)总结了空间数据质量需要考虑的11个指标:数据来源、空间精度、时间精度、属性精度、逻辑连贯性、数据完整性、语义准确率、元数据质量、分辨率、数据使用目的和质量变化等。众源地理数据的质量分析在用以上全部或部分指标作为评价标准的同时,还应加入对数据提供者的质量评价,充分考虑人为因素对数据质量的影响,建立更加有效的质量分析和评价模型,从而保证众源地理数据的有效性和可用性。

影响众源地理数据质量的因素主要包括三个方面:① 数据的采集或地图的绘制由非专业人员提供,可能存在一定的人为误差;② 数据可能来自不同的数据源,具有不同等级的精度;③ 不同采集者使用不同精度的GPS,采集的数据精度存在差异。众源地理数据的精度不能依靠常规的地图精度评定方法评估,需要选择合适的质量要素建立质量分析模型,依据质量分析模型与精度更高的数据进行分析对比来评估其数据质量。

目前,国外专家已经对欧洲地区的OSM数据质量问题进行了研究。如对英国地区的OSM数据质量进行分析,从定位精度和数据完整度两个方面建立OSM数据的质量评估模型;在评估希腊首都雅典的OSM数据质量时,将数据质量评估模型扩展到长度完整度、名称完整度、类型精度、名称精度和定位精度等方面。从OSM数据的完整度、专题精度、定位精度三个方面对OSM数据质量进行了分析研究[16-18]。

数据提供者的非专业性是众源地理数据质量不确定性的重要原因,Grira等人指出众源地理数据的使用者和提供者在众源数据上下文中具有认知区别[19],有必要建立针对数据提供者的评价模型。Exel等人提出在众源地理数据的质量控制指标中增加用户指标[20],如用户的数据上传次数、修改次数、反馈意见等,从而建立用户质量测度模型,实现众源地理数据的质量控制。

3.2 众源地理数据的信息提取与更新方法

众源地理数据的信息提取与更新是以众源地理数据作为数据源,以其低成本和高时效的优势实现地理信息的快速提取和及时更新。它是传统地理信息更新方法的重要补充,在特定情况下可以发挥不可替代的作用。如Goodchild指出,在面对印度洋海啸等严重自然灾害时,从传统的遥感影像上获取道路因影像有云或浓烟遮蔽受限制时,利用当地众源用户在Google Earth上及时标识的地物信息来补充数据库就更加高效[4]。

近年来国内外开展了一些利用众源数据进行地理信息提取和更新的方法和应用研究。在灾害快速响应方面,众源地理数据发挥了重大作用。

美国圣巴巴拉市的 4 次大型火灾案例研究表明,尽快建立新的道路数据库可以提供有效的逃生路线[7]。海地大地震后,大众在 OpenStreetMap 上协作完成道路、房屋和其他地物的重新编辑以建立震后地理数据库,利用常规测绘方式需花费上万英镑,同时耗时几年,而利用众源地理数据仅用了三个星期[8]。众源地理数据用于城市道路设计能较大提高人们的满意度,Seeger 指出众源地理数据在良好引导和补充城市规划道路数据库更新和重构中将发挥积极作用[15]。Steffen 在 GeoWiki 项目中利用众源用户对数据库中地物属性信息进行补充和修改[6],能明显提高效率及可靠性。众源地理数据提供的纹理和三维信息,被 Over 等结合 DEM 经综合生成了三维可视化数据库模型[21]。Zhang 等利用众源 GPS 轨迹数据,在行进方向的法线上进行模糊 C-均值聚类,将相邻车道分开,建立了几何精度较好的道路数据库,可用于数据库更新[22]。Mondzech 与 Sester 指出众源地理数据应用于行人导航比现存地形图数据库更有优势,因为它更新快且提供了大量真实的快捷路径,众源地理数据可对传统数据库进行很好的补充[23]。

众源地理数据为建立和更新地理数据库提供了一种不同于传统测绘方式的新途径。它不仅能有效地提取道路等地物和标注属性信息,而且能用于导航数据库的更新。但是由于众源地理数据来源众多且缺少统一规范,存在不足,目前尚未能广泛应用于大区域范围内高精度的数据库更新中。

综合以上分析,众源地理数据信息提取与更新的研究思路为:以建立众源数据的质量模型和多源数据配准和变化检测为核心,研究众源地理数据的信息提取与更新的协作机制和方法。主要研究方向包括:① 研究实现众源地理数据的高覆盖率和高完整性。众源地理数据虽然来源广泛,但是在单一的某个平台上部分区域的覆盖率存在较大的局限性。需要从现有的众源地理数据平台(Wikimapia、BingMap、GoogleMap 等)中获取尽可能多的数据以提高区域覆盖率和属性信息完整性。同时应研究将不同区域的更多用户参与到众源地理数据的协作机制。② 建立规范合理的质量模型。为来源广泛、质量各异的众源地理数据建立有效的综合取舍和聚类机制,以及快速处理算法。需要研究数据量和来源控制的范围,从而最有效地得到能满足数据库更新这一应用的数据覆盖率、完整性和几何精度。③ 提高众源地理数据库的几何精度。可结合高分辨率遥感影像、全景影像、Lidar 数据等对初始建立的数据库的几何位置进行精纠正,对不同数据的配准和修正位置精度的算法进行研究。④ 基于众源地理数据建立地理数据库并进行更新。研究两种不同规范的数据库的配准,对不同时间数据库的几何与属性信息进行比较,发现减少或新增部分,并用判别规则合并两套数据库的信息以实现数据库更新。

3.3　众源地理数据的分析与挖掘方法

众源地理数据作为一种由大众采集并向大众提供的开放地理数据,蕴含着丰富的空间信息和规律性知识。利用空间数据分析与挖掘方法可以从中提取信息、挖掘知识,从而为具体应用提供服务。

1) 众源地理数据拓扑分析。大部分众源地理数据的描述采用的是一种包含拓扑性质的数据结构。如 OSM 数据中的点、线、面等几何要素及其关系是通过顶点、路线和关系等来描述的。通过对某区域内的要素进行拓扑分析,能发现点、线、面的分布规律,挖掘该区域的空间结构和模式。例如,利用瑞典的 OSM 数据进行自然道路网络的提取与拓扑分析,发现自然道路网络存在无标度特性[24];利用香港的街道网络数据和年度平均的每天交通流量数据,通过街道网络的拓扑表示和分析,从而进行交通流量预测[25]。进行拓扑分析时经常用到平均度、平均路径长度和聚类系数等统计指标,结合空间统计方法可以探索地理要素的分布结构和模式。在利用众源地理数据进行网络拓扑分析时,可考虑与其他地理数据源集成和综合分析。

2) 利用众源数据探索地理空间的无标度特性。无标度从数学意义上讲就是某种现象的大小分布服从幂律分布。传统的地理学研究认为地理空间存在高斯分布的特性,而最近基于大量的地理数据的实证研究发现地理空间存在无标度的特性。例如,利用美国的 OSM 数据进行自然城市的提取与统计分析,发现美国的城市大小(无论是人口还是道路节点的个数)满足齐普夫定律[26];利用欧洲三个国家(英国,德国和法国)的 OSM 数据进行街区多边形提取与统计分析,发现所有这些街区大小服从 Lognormal 分布[27];利用 Tele Atlas MultiNet 地理数据库对德国 20 个城市的道路网进行统计分析,发现所有道路上的行车时间服从幂律分布,也就是具有无标度特性[28]。

3) 利用众源地理数据进行导航分析,为人们出行提供帮助。例如,综合应用 OSM 地图与航

空影像研制协作导航系统，结合A* 算法和用户评价在交通网络上进行路线计算和搜索，为行动不便以及有个人偏好的行人提供路线设计[29]；对OSM的一些城市道路网络数据进行分析，计算转弯最少、路径最短的导航路线[30]。

4）众源交通数据挖掘。众源交通数据中隐含着规律性的交通知识。利用空间数据挖掘方法可以挖掘出对交通管理和大众出行具有指导意义的规律性知识。例如，利用伦敦个人摩托车的GPS移动路线研究人们选择路线的偏好，发现路线选择时更多考虑的是角距离而非街区距离，即人们在出行时往往会选择转弯较少的方案[11]；利用OSM的大范围道路数据对人们的出行进行模拟研究，认为人们的出行模式主要受路网结构的影响，由此为出行行为和路径优化研究提供了新的视角[31]；对基于用户轨迹挖掘的智能位置服务进行研究，探讨了基于个人历史轨迹、多人轨迹数据的大众旅游推荐、个性化朋友和地点推荐等[32]。利用志愿者采集的GPS数据，基于历史位置对用户出行的相似性进行分析和挖掘，有助于探测用户之间的关系以及地理位置的相关性[33]；利用志愿者采集的GPS数据研究志愿者的空间移动规律特性，指出该出行距离具有较强的Levy移动特点，并通过基于代理的模型进行了验证，有助于人类动力学的实证与机制研究[34]。

综合以上分析，众源地理数据分析与挖掘的研究思路为：利用网络拓扑分析、空间数据统计建模、地理模拟、时空数据挖掘、统计物理学等方法对众源地理数据进行分析和挖掘，从中提取空间信息，挖掘空间知识。主要研究方向包括：① 众源地理数据的拓扑分析。利用拓扑分析方法研究并构建众源地理数据的网络拓扑关系。② 地理空间的无标度分析。利用空间数据统计建模方法研究地理现象的幂律分布。③ 导航分析：通过众源地理数据的交通流量分析和最优路径分析，为人们的出行提供导航服务。④ 众源地理数据统计建模：对众源地理数据的统计特性进行探索性分析，构建统计模型并分析其统计规律。⑤ 出行行为规律分析：利用空间聚类、频繁模式挖掘等空间数据挖掘方法从众源地理数据中挖掘出规律性知识，分析人们的出行规律，从而为大众旅游推荐、个性化朋友和地点推荐等提供基础。⑥ 人类动力学研究：利用统计物理学方法和地理模拟方法，研究出行距离分布，构建动力学模型，为交通规划设计、传染病控制和无线网络协议设计等提供支持。

4 总结与展望

众源地理数据处理与分析是近年来国际地理信息学科研究和应用的新热点。众源地理数据是一种有别于传统测绘的新型地理数据源，将深刻影响现有地理信息科学的发展方向和产业化模式。众源地理数据可以应用于应急制图、早期预警、地图更新、犯罪分析、疾病传播等地理信息服务领域。

众源地理数据处理与分析的研究才刚刚开始。随着地理信息科学与Web 2.0技术的进一步发展，作为专业测绘领域有效补充的众源地理数据的处理、分析、应用和服务是未来发展的重要趋势。

众源地理数据处理与分析需要完备的技术体系、深厚的理论方法为其提供支撑，需要对众源地理数据源及其特点进行分析和研究，需要研究众源地理数据的质量分析与评价方法、众源地理数据的信息提取与更新方法、众源地理数据的分析与挖掘方法等关键技术，并研究众源地理数据在相关领域的应用，不断扩大其应用领域，从而促进众源地理数据处理、分析与应用服务这一新兴研究方向的发展。

参 考 文 献

[1] Giles J. Wikipedia Rival Calls in the Experts[J]. *Nature*, 2006,443(7 111):493

[2] Howe J. The Rise of Crowdsourcing[J]. *WIRED Magazine*, 2006,14(6):1-4

[3] Heipke C. Crowd Sourcing Geospatial Data[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2010,65(6): 550-557

[4] Goodchild M F. Citizens as Sensors: the World of Volunteered Geography[J]. *GeoJournal*, 2007, 69(4): 211-221

[5] Goodchild M F. Geographic Information Systems and Science: Today and Tomorrow[C]. The International Conference on Mining Science and Technology, Xuzhou, 2009

[6] Fritz S, McCallum I, Schill C, et al. Geo-Wiki. Org: The Use of Crowdsourcing to Improve Global Land Cover[J]. *Remote Sensing*, 2009,1(3):345-354

[7] Goodchild M F, Glennon J A. Crowdsourcing Geographic Information for Disaster Response: a Research Frontier[J]. *International Journal of Dig-*

ital Earth, 2010, 3(3): 231-241

[8] Zook M, Graham M, Shelton T, et al. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake[J]. *World Medical Health Policy*, 2010, 2(2): 7-33

[9] Schade S, Luraschi G, Longueville B D, et al. Citizen as Sensors for Crisis Events: Sensor Web Enablement for Volunteered Geographic Information [C]. WebMGS, Como, Italy, 2010

[10] Ali K, Al-Yaseen D, Ejaz A, et al. Crowd ITS: Crowdsourcing in Intelligent Transportation Systems[C]. IEEE Wireless Communications and Networking Conference, Paris, 2012

[11] Turner A. The Role of Angularity in Route Choice: an Analysis of Motorcycle Courier GPS Traces[C]. COSIT, Aber Wrac'h, 2009

[12] Li Deren, Qian Xinlin. Data Management of Volunteered Geographic Information[J]. *Geomatics and Information Science of Wuhan University*, 2010, 35(4): 379-383 (**李德仁,钱新林. 浅论自发地理信息的数据管理[J]. 武汉大学学报·信息科学版**, 2010, 35(4): 379-383)

[13] Qian Xinlin. Research on the Representation and Management of Geospatial Data from Volunteered Geographic Information[D]. Wuhan: Wuhan University, 2011(**钱新林. 面向自发地理信息的空间数据表达与管理方法研究[D]. 武汉: 武汉大学**, 2011)

[14] Coleman D, Georgiadou Y, Labonte J. Volunteered Geographic Information: The Nature and Motivation of Produsers[J]. *International Journal of Spatial Data Infrastructures Research*, 2009, 4: 332-358

[15] Seeger C J. The Role of Facilitated Volunteered Geographic Information in the Landscape Planning and Site Design Process[J]. *GeoJournal*, 2008, 72(3): 199-213

[16] Ather A. A Quality Analysis of OpenStreetMap Data[D]. London: University College London, 2009

[17] Zulfiqar N. A Study of the Quality of OpenStreetMap. org Maps: A Comparison of OSM Data and Ordnance Survey Data[D]. London: University College London, 2008

[18] Ourania Kounadi. Assessing the Quality of OpenStreetMap Data[D]. London: University College London, 2009

[19] Grira J, Bedard Y, Roche S. Spatial Data Uncertainty in the VGI World: Going from Consumer to Producer[J]. *Geomatica*, 2010, 61(1): 61-71

[20] Van E M, Dias E, Fruijtier S. The Impact of Crowdsourcing on Spatial Data Quality Indicators [C]. The 6th International Conference on Geographic Information Science, Zurich, Switzerland, 2010

[21] Over M, Schilling A, Neubauer S, et al. Generating Web-based 3D City Models from OpenStreetMap: The Current Situation in Germany[J]. *Computers, Environment and Urban Systems*, 2010, 34: 496-507

[22] Zhang L J, Thiemann F, Sester M. Integration of GPS Traces with Road Map[C]. The 2nd International Workshop on Computational Transportation Science, San Jose, CA, 2010

[23] Mondzech J, Sester M. Quality Analysis of OpenStreetMap Data Based on Application Needs[J]. *International Journal for Geographic Information and Geovisualization*, 2011, 46(2): 115-125

[24] Jiang B. Volunteered Geographic Information and Computational Geography: New Perspectives[J]. *Crowdsourcing Geographic Knowledge*, 2013: 125-138

[25] Jiang B. A Topological Pattern of Urban Street Networks: Universality and Peculiarity[J]. *Physica A: Statistical Mechanics and Its Applications*, 2007, 384(2): 647-655

[26] Jiang B, Jia T. Zipf's Law for all the Natural Cities in the United States: a Geospatial Perspective[J]. *International Journal of Geographical Information Science*, 2011, 25(8): 1 269-1 281

[27] Jiang B, Liu X. Scaling of Geographic Space from the Perspective of City and Field Blocks and Using Volunteered Geographic Information[J]. *International Journal of Geographical Information Science*, 2012, 26(2): 215-229

[28] Lämmer S, Gehlsen B, Helbing D. Scaling Laws in the Spatial Structure of Urban Road Networks[J]. *Physica A*, 2006, 363(1): 89-95

[29] Holone H, Misund G, Holmstedt H. Users are Doing it for Themselves: Pedestrian Navigation with User Generated Content[C]. The 2007 International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST 2007), Cardiff, UK, 2007

[30] Jiang B, Liu X T. Computing the Fewest-turn Map Directions Based on the Connectivity of Natural Roads[J]. *International Journal of Geographical Information Science*, 2011, 25(7): 1 069-1 082

[31] Jiang B, Jia T. Agent-based Simulation of Human Movement Shaped by the Underlying Street Structure[J]. *International Journal of Geographical Information Science*, 2011, 25(1): 51-64

[32] Zheng Yu, Xie Xing. Enable Smart Location-based

Services by Mining User Trajectories[J]. *Communications of the CCF*, 2010,6(6):23-30 (**郑宇,谢幸. 基于用户轨迹挖掘的智能位置服务[J]. 中国计算机学会通讯**, 2010,6(6):23-30)

[33] Li Q N, Zheng Y, Xie X, et al. Mining User Similarity Based on Location History[C]. The 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Irvine, CA, USA, 2008

[34] Jia T, Jiang B, Carling K, et al. An Empirical Study on Human Mobility and Its Agent-based Modeling [J]. *Journal of Statistical Mechanics: Theory and Experiment*, 2012,11:1-20

Methods of Crowd Sourcing Geographic Data Processing and Analysis

SHAN Jie[1,2] *QIN Kun*[1] *HUANG Changqing*[1] *HU Xiangyun*[1] *YU Yang*[1]
HU Qingwu[1] *LIN Zhiyong*[1] *CHEN Jiangping*[1] *JIA Tao*[1]

1 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

2 School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA

Abstract: Crowd sourcing geographic data (CSGD) is a new kind of open geospatial data collected and provided to citizens or organizations by non-professional volunteers. Its acquisition differs from conventional methods of surveying and mapping. It has the characteristics of being up-to-date, informative, low cost, and large size. Unique issues of such data are its heterogeneous data quality, redundancy and incompleteness, uneven data coverage, absence of data standards, privacy and safety, etc. It can be employed in a wide range of application fields including emergency cartography, early warning, map and database updating, crime analysis, and epidemiologic studies. Thus, it is important to systematically investigate the key aspects regarding its processing and analysis. The paper reviews the progress and challenges of CSGD processing and analysis. Firstly, the paper summarizes the concept and characteristics of CSGD. Secondly, the paper introduces the acquisition methods and potential data sources of CSGD. Thirdly, the paper discusses some key methods and techniques on the processing and analysis of CSGD, which includes data quality metrics, information extraction and updating, spatial data analysis and mining. Specifically, the paper suggests that firstly data evaluation is a critical prerequisite for CSGD research and application, secondly information extraction and updating via CSGD can complement the conventional methods of geographic database updating, and finally spatial data analysis and mining on CSGD can provide valuable information and knowledge for potential applications using methods like network topological analysis, spatial statistical analysis, and spatial data mining. Lastly, the paper draws some conclusions on the state-of-the-art of CSGD processing and analysis and points out the future research directions.

Key words: crowd sourcing geospatial data (CSGD); quality analysis and evaluation; information extraction and updating; spatial analysis; spatial data mining; geographical information service

First author: SHAN Jie, professor, PhD, PhD supervisor. He is now interested in Geospatial Data Processing. E-mail: shanj@whu.edu.cn

Corresponding author: QIN Kun, professor, PhD, PhD supervisor. E-mail: qink@whu.edu.cn

Foundation support: The National Natural Science Foundation of China, No. 61172175.