

众源地理数据分析与应用

单杰 贾涛 黄长青 胡庆武 秦昆 黄润 编著



科学出版社

众源地理数据分析与应用

单 杰 贾 涛 黄长青 编著
胡庆武 秦 昆 黄 润



科学出版社

北 京

版权所有,侵权必究

举报电话:010-64030229;010-64034315;13501151303

内 容 简 介

本书较为全面系统地论述众源地理数据的基本概念、分析方法、前沿技术以及最新的应用成果。全书针对众源地理数据的特点进行分析利用与挖掘,共分8章。第1章介绍众源地理数据的来源、特点与典型应用方向;第2章结合多个实际案例论述众源地理数据的采集与获取;第3章论述众源地理数据质量分析的理论框架、关键问题与相关分析案例;第4章论述众源地理数据在道路网演化及志愿者绘图行为方面的特点和规律;第5~6章论述众源地理数据在道路提取与更新中的应用与相关方法;第7章论述众源轨迹数据在城市信息提取与人类行为分析中的应用;第8章探讨众源地理数据的共享平台建设与可视化技术。

本书可供测绘遥感、地理信息、城市地理、交通工程等专业领域的科研人员和高校师生阅读参考。

图书在版编目(CIP)数据

众源地理数据分析与应用/单杰等编著. —北京:科学出版社,2017.3

ISBN 978-7-03-050866-9

I. ①众… II. ①单… III. ①地理信息系统-数据-分析 ②地理信息系统-数据-应用 IV. ①P208

中国版本图书馆 CIP 数据核字(2016)第 279428 号

责任编辑:杨光华 / 责任校对:肖 婷

责任印制:彭 超 / 封面设计:苏 波

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

武汉中远印务有限公司印刷

科学出版社发行 各地新华书店经销

*

开本:787×1092 1/16

2017年3月第 一 版 印张:17 1/2

2017年3月第一次印刷 字数:412 000

定价:120.00 元

(如有印装质量问题,我社负责调换)

前言

众源地理数据主要是指那些来源广泛,通常由非专业个人或单位生产的地理数据。它随着导航定位、互联网等现代技术的发展和普及而产生,并在近几年引起相当一些学科和领域的关注。近十年来,一方面导航定位技术不断普及和大众化,另一方面包括 Web 2.0 在内的计算机技术极大地方便了信息的生产和共享。在这样的背景下,普通民众(团体)在日常生活(尤其是网络活动)中直接或间接地生产出大量的地理数据,典型的例如 OpenStreetMap 道路数据、新浪微博签到数据、浮动车数据。考虑到来源多样性和生产自发性是这类数据区别于传统地理数据的本质属性,本书将这些数据称为“众源地理数据”。相较于传统的专业测绘地理数据,众源地理数据有着数据量大、现势性强、覆盖面广、成本低廉等优点,成为近几年测绘地理信息提取与更新、城市热点提取与分析、交通分析及其可视化、时空行为分析等领域的热点数据。

尽管近年来基于众源地理数据的研究不少,但是国内尚缺少能系统、集中、深入地论述这种新型数据及其分析和应用的书籍。相对传统地理数据而言,众源地理数据有更强的异构性、多样性和不确定性,想要清楚地认识这种数据的全貌、要点及潜在价值并不容易,因此编撰系统论述这类数据的相关书籍十分必要。本书依托国家自然科学基金项目“基于众源 GPS 路线数据的城市道路网自动更新和重构”(项目编号:61172175),从数据获取、质量分析、特点分析、多领域应用等多个角度切入,尝试系统地论述众源地理数据(尤其是道路数据)及其分析和应用,以期一定程度上填补这个空缺。

本书采用以下流程化体系展开论述:概述、获取、质量分析、特点及规律分析、不同领域的应用、共享和可视化。第 1 章介绍众源地理数据的来源、特点及典型应用方向,以引入数据全貌、总领全书;第 2 章论述众源地理数据的采集和获取过程,深入认识这种数据的来源;第 3 章论述众源地理数据的质量及其分析方法,这是开展分析和应用的基础;第 4 章论述众源道路数据演变的特点及规律;第 5~7 章则结合城市热点分析、道路提取与更新、交通信息及交通设施的提取与分析、市民行为规律分析等多个方向,论述众源地理数据的不同应用;第 8 章论述众源地理数据的共享平台构建及可视化。

本书以我们近期在众源地理数据方面所做研究为基础,对此进行整理、归纳和总结,并最终形成众源地理数据分析与应用的系统化框架。各个章节的主要撰写人员分别为:第 1 章,单杰、王英;第 2 章,单杰、黄润、黄长青;第 3 章,胡庆武、王明、单杰、陈潇健;第 4 章,贾涛、赵鹏祥;第 5 章,黄玉春、崔卫红、陈昱霖、单杰;第 6 章,单杰、胡翔云、刘长勇、王振华;第 7 章,秦昆、黄长青、王英、单杰;第 8 章,黄长青、黄润、单杰、熊恢、桂志鹏。本

书由单杰和贾涛共同指导黄润完成统稿,最后由单杰对全书进行了定稿。

周檬、李依娇、刘成堃、熊恋、陈江平、曹劲舟、武红宇、王健、周天、徐静、杜胜兰、何韬、刘子政、李默颖、李枫、栗法、孟凡泽、刘亚奇等对本书的内容有直接贡献,在此表示感谢。在本书的准备和撰写过程中,感谢余洋、付建红、王玉龙、康朝贵、黄培旗、姜卓君、周溢波等提供的指导或帮助。感谢参与、组织、协助“2015 地理日记伴我行”的人员和组织,包括:约 659 名武汉大学遥感信息工程学院的学生志愿者、武汉大学遥感信息工程学院研究生会及学生代表项皓东等、潘婷和刘亚奇等 25 名活动工作人员、为活动组织提供资料支持的何凌和谢云等,以及指导老师卢宾宾等。感谢组织 2014 年武汉市社会状况综合调查的林曾等武汉大学社会学系师生。

感谢以下团体在数据或资金上对本书提供的帮助:国家自然科学基金项目“基于众源 GPS 路线数据的城市道路网自动更新和重构”、北京四维图新科技股份有限公司、广州建通测绘地理信息技术股份有限公司、中国公路工程咨询集团有限公司、高分综合交通遥感应用示范系统(一期)项目、测绘遥感信息工程国家重点实验室 3S 集成与通信研究室、北京吉威时代软件股份有限公司、武汉安捷联合在线信息科技有限公司等。

本书自成体系,具有专著与教材相结合的写作风格,希望能帮助读者全面、系统、深入地认识众源地理数据及其应用。对于科研工作者,本书可作为工具书或参考书;对于高等院校的本科生或研究生,本书可作为教材或参考书。

由于作者水平有限,书中难免存在纰漏,敬请各位读者不吝指正。

作者

2016 年 8 月

目 录

第 1 章 绪论	1
1.1 众源地理数据的概念	2
1.2 众源地理数据的来源与分类	3
1.2.1 众源地理数据的来源	3
1.2.2 众源地理数据的分类	4
1.3 众源地理数据的特点	4
1.4 众源地理数据的主要应用方向	6
1.4.1 众源地理数据拓扑及无标度分析	6
1.4.2 城市交通基础地理信息提取	7
1.4.3 人类时空活动分析	7
1.4.4 其他分析与应用	8
1.5 本书内容和组织结构	9
参考文献	10
第 2 章 众源地理数据的采集与获取	13
2.1 概述	14
2.1.1 众源地理数据的采集	14
2.1.2 众源地理数据的获取	20
2.2 OpenStreetMap 项目及其数据的采集与获取	23
2.2.1 OpenStreetMap 项目概述	23
2.2.2 OpenStreetMap 数据的采集	24
2.2.3 OpenStreetMap 数据的获取	26
2.3 基于手机的志愿者位置数据采集与 GeoDiary 系统	26
2.3.1 背景和一般方法	26
2.3.2 典型案例	28
2.3.3 GeoDiary 系统与数据采集实验	30
2.3.4 基于手机的位置数据采集面临的挑战和应对策略	32
2.4 网络众源地理数据的获取与方案设计	38
2.4.1 网络地理数据的组织结构	38
2.4.2 获取基础和限制	40
2.4.3 获取方案设计与实验	42
2.5 本章小结	44
参考文献	44

第3章 众源地理数据质量分析与评价	47
3.1 众源地理数据的质量概述	48
3.1.1 众源地理数据质量问题及其研究进展	48
3.1.2 众源地理数据的误差源	48
3.2 众源地理数据的质量分析框架	49
3.2.1 概述	49
3.2.2 众源地理数据质量分析的关键问题	50
3.3 OpenStreetMap 道路数据的质量分析	51
3.3.1 概述	51
3.3.2 OpenStreetMap 道路数据质量分析方法	52
3.3.3 质量要素及其计算模型	53
3.3.4 实例应用与分析	55
3.4 位置签到数据质量分析	64
3.4.1 位置签到数据及其特点	64
3.4.2 位置签到数据质量分析方法	64
3.4.3 实例应用与分析	67
3.5 众源 GPS 轨迹的预处理	71
3.5.1 GPS 轨迹粗差	71
3.5.2 GPS 轨迹粗差剔除的常用方法	72
3.5.3 一种基于趋势的粗差剔除法	74
3.5.4 随机误差及轨迹的平滑	79
3.6 本章小结	81
参考文献	82
第4章 OpenStreetMap 路网演变分析	85
4.1 OpenStreetMap 道路数据结构	86
4.2 网络演变理论与方法	87
4.2.1 网络演变理论	87
4.2.2 网络模型构建	88
4.2.3 网络分析方法	88
4.3 开放道路网结构演变分析	90
4.3.1 OpenStreetMap 道路网演变概况	90
4.3.2 OpenStreetMap 道路网络建模	93
4.3.3 属性分析	93
4.3.4 几何分析	94
4.3.5 拓扑分析	99
4.3.6 中心性分析	101
4.4 志愿者绘图行为演变分析	103
4.4.1 志愿者数量	104

4.4.2 志愿者类型	104
4.4.3 志愿者上传的数据类型	105
4.5 本章小结	106
参考文献	106
第5章 OpenStreetMap 辅助的影像道路提取	109
5.1 高分辨率影像道路提取概述	110
5.1.1 道路采集的意义与传统方法	110
5.1.2 基于高分辨率影像的道路提取	111
5.1.3 OpenStreetMap 辅助道路提取的潜力	112
5.2 基于种子点追踪的道路提取与更新	113
5.2.1 概述	113
5.2.2 矢量引导下的路域缓冲带生成	114
5.2.3 超高分辨率遥感影像的道路提取与更新	116
5.2.4 高分辨率遥感影像的道路提取与更新	121
5.3 基于机器学习的道路提取	127
5.3.1 概述	127
5.3.2 道路分级	129
5.3.3 道路样本库建立	132
5.3.4 HOG+SVM 训练提取道路特征	135
5.3.5 实验及分析	138
5.4 本章小结	144
参考文献	144
第6章 基于众源数据的路网提取与更新	147
6.1 引言	148
6.2 基于众源 GPS 数据的道路提取	148
6.2.1 主要方法	148
6.2.2 一种基于栅格图的道路中心线提取方法	155
6.3 基于 OpenStreetMap 的城市道路数据库更新	159
6.3.1 研究背景	159
6.3.2 路网匹配的自适应缓冲区增长法	161
6.3.3 自适应概率松弛法路网特征点匹配	179
6.4 本章小结	184
参考文献	184
第7章 基于众源轨迹的交通出行信息提取与分析	187
7.1 概述	188
7.2 利用出租车轨迹提取城市热点区域	192
7.2.1 空间聚类的主要方法	192
7.2.2 基于决策图和数据场的轨迹聚类方法	193

7.2.3 利用轨迹聚类方法提取城市热点区域	196
7.3 基于出租车轨迹的城市拥堵区域提取及其分布模式分析	200
7.3.1 基于浮动车轨迹的拥堵事件提取	201
7.3.2 拥堵事件的统计分析	201
7.3.3 通过聚类提取拥堵易发区域	203
7.3.4 基于 K 函数的拥堵时空分布模式探测	207
7.3.5 利用轨迹数据场模拟拥堵强度分布	208
7.4 基于 OpenStreetMap 轨迹的交通附属设施提取	210
7.4.1 OpenStreetMap 轨迹数据及预处理	210
7.4.2 基于低速极值点的收费站提取	211
7.4.3 基于轨迹始末点的停车场提取	214
7.4.4 分时段道路流速信息建模	217
7.5 基于志愿者 GPS 轨迹数据的大学生时空行为分析	219
7.5.1 概述	219
7.5.2 志愿者 GPS 轨迹数据获取与质量分析	220
7.5.3 数据预处理	223
7.5.4 停留行为提取	226
7.5.5 基于停留行为的大学生时空行为分析	229
7.5.6 实验验证与总结	234
7.6 本章小结	235
参考文献	236
第 8 章 众源地理数据共享平台设计与可视化	241
8.1 概述	242
8.1.1 数据共享及共享平台建设	242
8.1.2 众源地理数据可视化的研究进展	243
8.2 共享平台设计和实现	246
8.2.1 总体设计	246
8.2.2 数据体系和标准体系	247
8.2.3 软硬件技术采纳	250
8.3 可视化应用实例	252
8.3.1 出租车轨迹的高维可视化系统	252
8.3.2 GeoDiary 地理日记系统	257
8.3.3 大众在线制图系统设计	262
8.4 本章小结	269
参考文献	269



第 1 章

绪 论

近年来,众源(crowdsourcing)地理数据的出现与快速增长日益影响着地理信息科学与技术的发展方向,同时也为地理信息产业的发展提供了新的研究数据和应用方向,成为测绘领域新兴的基础数据和研究热点。本章首先概括性地描述众源地理数据产生的背景和相关概念;随后对众源地理数据的来源、分类和特点做综合介绍,指出当前众源地理数据的主要应用方向;最后,介绍本书的内容和组织结构。本章有助于读者快速宏观地认识众源地理数据,为其他章节的阅读奠定基础。

1.1 众源地理数据的概念

几个世纪以来,地图制图一直由国家组织实施。早期,地图制图主要由军事机构主体控制,将其用于军事方面。20 世纪开始,越来越多的国家致力于将制图用于民用服务,地理制图开始由一些政府测绘和地籍机构或政府授权的公司主导负责,由受过训练的人员根据定义良好的制图规范及质量保证程序进行地图生产、地理要素发布等活动,用于土地管理、基础设施建设和环境监测(Heipke,2010)。

近年来,定位导航和计算机技术快速发展。一方面,借助便携、低成本的 GPS 接收器,如具有 GPS 定位功能的手机、相机等,全球定位系统(global positioning system, GPS)的应用快速普及,普通民众能够方便地记录个人位置信息;另一方面,Web 2.0 技术的出现在很大程度上改变了以往信息交流的方式,普通民众开始参与到绘制其生活环境的地图活动中来(O'Reilly,2007)。这些技术发展带来的数据极大地丰富了以往由政府或公司提供的地理信息,对地理数据、地理信息,甚至地理知识的产生和传播产生了深刻影响(Sui et al.,2013)。

面对由公众产生的日益增长的地理数据,来自不同领域的学者逐渐意识到这些数据对地理空间信息的意义,从不同角度对这些数据进行了定义。如 Turner(2006)将其定义为“NewGeography”(新地理),意指普通民众通过现有工具,按照自己的方式制作和使用地图,向朋友、游客分享地理位置,理解地理知识;Goodchild(2007)提出“volunteered geographic information”(VGI,志愿者地理信息)概念,指普通民众在参与各种社会活动的过程中主动或无意地创建出的地理信息;Heipke(2010)将公众产生的地理数据定义为“crowdsourcing geospatial data”(CGD,众源地理空间数据),将其描述为由大量非专业人员志愿获取并通过互联网向大众或各种机构提供的一种开放地理空间数据;单杰等(2014)将其定义为“crowdsourcing geographic data”(众源地理数据),强调“众源”描述的对象同时包含数据获取过程、数据建模或数据处理。此外,还有“user-generated content”等不同概念,尽管这些概念表述不一,但其核心的关键词一致,均是指普通民众通过多种计算设备,使用互联网生成、分享、分析地理信息(Sui et al.,2013)。

众源最早被维基百科(Wikipedia)^①定义为通过一大群人的贡献获得所需服务、想法和内容的过程。该概念由美国 Wired 杂志记者 Howe(2006)提出,用来描述从“外包”基础上发展出来的新的商业模式。随着计算机技术、移动终端设备、GPS 定位和导航的发展,众源逐渐发展成为一种新的信息交互模式,它改变了信息传播的方式,使每个人都可能兼具信息的生产者、传播者和消费者三重身份。相比由政府部门、大型测绘遥感公司生产的传统地理数据,众源方式使大众在地理数据的生产和传播中起到了越来越重要的作用,地理数据由原先自上而下的生产方式转变为自下而上的生产方式

^① 引自:Wikipedia-Crowdsourcing. <https://en.wikipedia.org/wiki/Crowdsourcing>

(Sui et al., 2013)。

广义上,本书所要讨论的众源地理数据的内容在基本概念之外有所延伸,除了上述公众通过互联网、移动终端设备产生的带有地理位置信息的数据外,还包括在地理空间、属性和拓扑数据的获取和应用方面,从传统有明确目的、单纯依靠专业测绘的方式延伸到使用大众主动或被动产生的带有地理信息的数据及开源地图等多种方式,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 众源地理数据的来源与分类

1.2.1 众源地理数据的来源

相较于传统由政府部门、大型测绘遥感公司生产的专业测绘地理数据,众源地理数据往往并不具有明确的地理测绘目的,其来源广泛。典型的来源包括如下几种。

(1) 由特定部门或公司发布的公共版权数据。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,如美国地质调查局(United States Geological Survey, USGS)官网上提供的最新、最全面的全球卫星影像可以免费下载。对于一些特定的众源项目,也有一些部门和企业愿意免费赠送其持有的地理数据,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(钱新林, 2011)。

(2) 开源地图要素数据。OpenStreetMap^①、WikiMapia 等网站向用户提供了创建地理对象的功能。一部分网民出于自我满足、利他主义或是描述周围环境等目的(Coleman et al., 2009),参照正射影像、GPS 轨迹,主动在这些网站上创建、编辑、描述各种地理对象。谷歌地球甚至允许用户对自己感兴趣的地方进行三维建模。这些开源地图的出现和兴起,见证了由地图制作爱好者生产、更新地图的成功。以 OpenStreetMap 为例,相较于传统测绘部门生产的地图,OpenStreetMap 地图的制作参与者往往具有更优质的本地知识,贡献者使用航空影像、GPS 设备和传统的地区地图来确保 OpenStreetMap 的精确性和实效性,使其具有更高现势性。

(3) 来自城市公共交通管理部门的行驶数据。为便于公共交通团队管理和提升公共交通服务水平,当前不少公共交通部门都利用 GPS 记录仪器记录运输工具的轨迹,如航空公司飞机航线数据、浮动车(一般是指安装了车载 GPS 定位装置并行驶在城市主干道上的公交汽车和出租车)GPS 轨迹数据。这些带有时空地理信息的交通数据对于动态了解城市交通流变化、城市居民移动规律和城市热点具有重要研究价值。

(4) 由公众日常生活中有意或无意产生的空间数据。公众在日常生活中,无意间产生了大量地理时空数据,如信用卡刷卡数据、手机通信记录、地铁及公交等公共交通刷卡数据,这些数据产生于居民日常生活,能被服务商记录下来,具有丰富的语义信息,对解读

① 引自:OpenStreetMap, <http://www.openstreetmap.org>

居民出行习惯、出行范围、交通方式选择等人类行为分析具有研究价值。

(5) 公众在社交网站上共享的带有地理信息的数据。Web 2.0 的变革,改变了公众在互联网中的作用,公众不再仅仅是信息的“阅读”者,更是信息的“创造”者和“传播”者,同时,Web 2.0 也简化了客户交互过程。出于信息共享的目的,许多民众以即兴和松散的方式记录发生在某些地点和时间的事件,并将这些信息通过文字、图片或者录像片段等格式标注在相应的网站上,或将包含了位置数据的个人信息发布到网上(如签到(check-in)数据)。在此背景下产生的一系列社交网站成为用户分享个人生活状态、发表观点和传播观点的媒介,如国外的 Wikiloc^① 和国内的六只脚网站^②等轨迹共享网站,Facebook^③、Flickr^④、新浪微博、QQ 空间等社交网站。这些网站上的信息具有丰富的地理信息和语义信息,对研究网络群体的地理空间分布、聚落规模、区位、空间结构及功能区分布具有重要的研究价值(Feick et al.,2013)。

1.2.2 众源地理数据的分类

众源地理数据虽来源广泛,但按内容划分,基本可以分为两种类型:一类是空间数据,一类是描述型数据。

空间数据包括点、线、面三种类型,按数据类型划分,又包括矢量数据和栅格数据。点数据如用户的签到位置数据、刷卡等位置信息,线数据和面数据如用户上传至网络的旅程线路 GPS 轨迹、网民自发编辑的地理对象(道路、湖泊等)、第三方无偿提供的矢量数据等。

描述型数据,又称属性数据,这些数据多来源于社交网站,形式多种多样,包括各种带有地理位置的照片、文本、视频、音频,表现出零散、无规则的特点。例如,Wikimapia 允许用户针对某一块地域进行文字描述,用户上传到 Flickr 上带有时空标签的照片不仅反映了用户活动轨迹,还能构成一定区域的影像数据。此外,用户的签到数据、刷卡数据等,往往具有附属信息,一定程度上含有描述该位置属性或活动行为的内容。需要说明的是,有时描述型数据既包含了地理对象的属性信息,又隐含了其拓扑信息(李德仁等,2010)。

1.3 众源地理数据的特点

综合比较不同来源的众源地理数据与传统测绘地理数据,其特点、优势和不足如下。

(1) 数据量大。很大一部分众源地理数据来自互联网用户有意或无意提交至网络的数据,互联网用户群的迅速发展使众源地理数据激增,任何人都可以参与数据的生产和传

① 引自:Wikiloc-GPS trails and way points of the world. <http://www.wikiloc.com/wikiloc/home.do>

② 引自:六只脚-GPS 轨迹记录_户外自助游_自助游路线. <http://www.fooooooot.com>

③ 引自:欢迎使用 Facebook-登录、注册或详细了解. <http://www.facebook.com>

④ 引自:Flickr. <http://www.flickr.com>

播,也能便捷地通过互联网获取开源地理数据。这也意味着,无论是像 OpenStreetMap 这样的共享网站,还是具体的众源地理数据使用者,经常需要面对数据量大带来的一系列技术难题,如高效存储、网络共享中的快速传输等。

(2) 现势性强。不少众源地理数据产生于实时在互联网上发布状态和共享位置数据的网民,具有明显的实时更新特点。大众实时发布的与位置相关的数据使众源地理数据在灾难制图、人道主义援助和救灾中具有显著的应用价值(Feick et al., 2013)。例如,2010 年海地地震后,当地民众使用手机等移动设备将自己获知的消息发送到 Ushahidi 平台^①,平台将收到的受灾人口分布状况、各地物资储备和救灾物资缺乏等情况进行分类,供救援人员、医护人员使用。此外,这种现势性强的地理数据能极大地缩短信息获取和更新时间,对交通状况实时分析和道路更新也具有重要意义。例如,当人们遇上某条道路因施工而无法通行的情况时,有时会愿意将这个道路信息发布于推特、微博;对于 Wikiloc、GPSies 等一些通过上传 GPS 轨迹来分享行程路线的网站,其数据更新频率也较快、实时性强。

(3) 传播速度快。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,美国加利福尼亚州 2009 年 5 月的杰苏斯塔(Jesusita)火灾期间,通过建立地图式火灾监视网站,迅速整合、发布了各种志愿者地理信息和当地的实时火灾信息(Goodchild et al., 2010)。

(4) 信息覆盖面广。传统的地理空间数据缺乏社会化属性信息,众源地理空间数据来源于大众,具有很强的社会性。而大众用户活动的场所多集中在城市,并以交通为主要空间表现,其本身包含丰富的位置信息、语义信息和行为信息,与人类活动及社会发展紧密相关。另外,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,信息覆盖面广。

(5) 成本低廉。众源地理数据大多来自网民自发或无意采集的地理数据,相对专业的地理数据生产,其采集和处理的成本很低。这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。

(6) 缺乏统一规范。众源地理数据来源广泛,数据格式各异,不同数据的内容、精度、格式不同,数据组织和存储方式也千差万别,缺乏统一的标准规范,有时难以满足一些专业的地理数据要求。此外,数据缺乏统一的规范还表现在元数据的标准不一,部分众源地理数据常常缺乏元数据或元数据描述不清晰,难以检索和查询。

(7) 质量不确定。与传统规范的地理数据相比,众源地理数据来源广泛,包括政府、公司发布的公共服务数据,以及普通民众由移动终端、互联网分享的带有地理坐标信息的数据。一方面,由于缺乏市场监管力量和专业生产标准,众源地理数据的生产方式和过程不同,所采用的数据采集设备精度、方法不一,质量差异大(Goodchild et al., 2010),具体表现在数据精度和完整性上;另一方面,志愿者创建、分享地理数据的动机多样、主观性高低和技术水平也增加了数据的不确定性(Coleman et al., 2009; Budhathoki et al., 2008;

^① 引自:Ushahidi. <https://www.ushahidi.com>

Flanagin et al., 2008)。数据质量不可预测,可能存在偏差、重复、错误(李德仁等, 2010), 甚至恶意扭曲的成分, 政府部门和决策者往往对众源地理数据抱有怀疑态度, 拒绝将其用于决策过程(Johnson et al., 2013)。

(8) 覆盖不均匀。首先, 众源地理数据在空间上分布不均匀, 尽管当前众源地理数据增长速度快, 数据量巨大, 一些地区被海量数据淹没, 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, 而另一些地区却严重匮乏。例如, OpenStreetMap 数据在伦敦的覆盖率明显高于中国湖北省的覆盖率。造成这种不均匀的原因有很多, 包括世界各地网络覆盖率、使用率不均匀(Miniwatts Marketing Group, 2015), 还包括隐含的某些决定网络上传内容内容的社会政治因素(Engler et al., 2007)、网络审查(Warf, 2011)、权力法(Shirky, 2003)等。其次, 众源地理数据的来源不均匀, 以社交网站上公民自发产生的地理数据为例, 生产者的年龄大多为使用智能手机的年轻人, 年长者较少。此外, 这些生产者的性别比例也不均衡, 社会分工也不均匀。

(9) 开放性。众源模式大多采用开放的开源平台、开放的数据标准和协同工作规范, 其数据成果多采用开放、共享的免费应用模式。

(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生产产生影响。

1.4 众源地理数据的主要应用方向

众源地理数据作为一种开放地理数据, 蕴含着丰富的空间信息和规律性知识。利用空间数据分析和挖掘方法可以从中提取信息、挖掘知识, 为具体应用提供服务。本书结合近年来众源地理数据的应用研究, 将众源地理数据的主要应用方向总结如下。

1.4.1 众源地理数据拓扑及无标度分析

众源地理数据的拓扑和无标度分析, 指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, 利用空间数据统计建模方法研究地理现象的幂律分布。

大部分众源地理数据被描述为一种包含拓扑关系的数据结构, 如 OpenStreetMap 数据中的点、线、面等几何要素, 它们的关系是通过顶点、路线的关系来描述的。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, Jiang(2007)利用香港的街道网络数据和年度平均每天交通数据流量, 借助街道网络的拓扑表示和分析, 进行交通流量预测; Jiang(2013)利用瑞典 OpenStreetMap 数据进行自然道路网络的提取和拓扑分析, 发现道路网络存在无标度特性。拓扑分析经常用到平均度、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搜索地理要素的分布结构和模式。

无标度特性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, Jiang 等(2011)利用美国的 OpenStreetMap 数据进行自然城市的提

取和统计分析,发现美国城市的大小(无论是人口还是道路结点的个数)满足齐普夫定律;Lämmer 等(2006)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。

1.4.2 城市交通基础地理信息提取

城市交通是人类科学技术进步和工业发展过程中形成的社会产物,是由人、车、道路网络、私人交通、公共交通、专业运输组成的一个动态、复杂的系统,是城市中极为重要的组成部分。城市交通基础地理信息是包括城市道路网、路网附属信息和城市拥堵状况等内容的综合地理信息,在交通管理、车载导航、城市规划和网络地图服务等领域都具有重要作用(Wilson et al.,1998),也是智慧城市的重要内容之一。很大一部分众源地理数据源于城市中生活的民众、行驶的公共交通工具和私人交通工具,其蕴含丰富的城市交通信息,且具有现势性强、容易获取、成本低廉等优点,相比于测绘部门或测绘型、导航服务型公司所生产的路网地图,往往还含有丰富的小道路、人行道等信息。这些数据能够用来提取、更新路网,挖掘城市交通附属物,如停车场、高速路收费站、加油站等道路交通基础设施,作为城市基础地理信息提取的基础数据源。

Schroedl 等(2004)采用配备了差分 GPS 接收器的车辆获得高精度的 GPS 数据,将轨迹点划分为道路段轨迹和十字路口两种类型,从而得到道路中心线,提取高精度道路地图;Li 等(2012)综合利用了 GPS 数据的空间及语义信息从浮动车数据中提取道路;陈舒燕(2010)利用 OpenStreetMap 数据实现了基于在线位置信息服务方式的出行可达性分析。众源数据还反映了城市中居民的出行信息,对其进行时空聚类能实时提取交通拥堵情况,获得城市动态拥堵状况,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并用于道路规划、商场选址等决策。例如, Lee 等(2008)利用 K 均值方法对出租车的上下车点进行分析,从而为空出租车进行位置推荐;Carisi 等(2011)利用 GPS 轨迹信息的时间、速度等信息对城市交通信号位置信息进行估测;陈漪(2011)则借助 GPS 轨迹信息的时间、速度、高程等识别城市立交桥。

1.4.3 人类时空活动分析

计算机技术和移动定位技术的飞速发展,使得普通民众开始成为地理信息的生产者和传播者。众源地理数据的生产主体大多为民众,因而这些时空数据一定程度上能够反映个体、群体的活动规律,可用于人类时空活动分析。一方面,众源数据能用于人类动力学研究,研究个体的行为活动规律和偏好。Turner(2009)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,发现道路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路线;Jia 等(2012)利用 OpenStreetMap 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。另一方面,众源数据还能用于城市居民群体活动规律研究,挖掘城市热点信息。例如,Yue 等(2009)利用 Single-Linkage 聚类算法对不同时段内出租车轨迹的上下车点进行分析,从而挖掘出依赖于时间的兴趣区域和移动模式。

1.4.4 其他分析与应用

众源地理数据来源广泛,包含丰富的时空地理信息、语义信息、属性信息,在众多领域都能弥补传统时空数据更新慢、成本高、语义缺失等不足。

在灾害应急方面,众源地理数据,特别是志愿者地理信息,在近几年的重大自然灾害中发挥了重要作用(Goodchild et al.,2010)。近年来全球范围内重大灾害频发,造成了巨大的人员伤亡和经济损失,紧急情况下政府机构和管理部門不可避免地捉襟见肘,特别是那些对生命和财产安全有巨大威胁的社会危机和自然灾害。政府机构和管理部門往往人员有限,缺乏当地地形分布的有效知识,对重要地理信息的有效反映能力差,而普通民众对所在地具有更好的本地知识,更了解灾害发生或蔓延情况,也具备通过互联网生产、发布救灾消息的能力(Budhathoki et al.,2008)。快速、准确地获取灾害相关消息,如灾害现场信息(灾害级别、范围等)、受灾人员信息(位置、身体状况、环境等)、基础设施信息(道路、供水、供电等)、救援人员信息(位置、环境等)、救援物资信息(数量、类别、需求等),是启动灾害应急响应、制定救援方案和实施救援的重要依据之一(高原等,2013)。良好的灾害应急系统和平台设计,有利于政府和救援部門在灾害发生时迅速获得受灾信息和民众需求,并将它们综合到容易理解的地图和状态报告中,以在灾害中迅速做出合理、正确的反应,挽救人民财产和生命安全。

在公共卫生和流行病蔓延方面,GIS 因其强大的空间分析和可视化能力已经受到国内外专家的重视(武继磊等,2003)。众源地理数据的出现,特别是志愿者地理信息,依靠智能手机和其他位置感知设备等简单易用的数据收集工具,能提供高现势性、来源广泛的数据池,为公共卫生健康研究提供新的机遇(Goranson et al.,2013)。2009年2月19日,*Nature* 上刊登了一篇关于谷歌预测流感的文章(Ginsberg et al.,2009),谷歌工程师根据汇总的搜索数据,近乎实时地对全球当前的流感疫情进行了估测,推出了“谷歌流感趋势”(Google flu trend,GFT),并于2009年H1N1暴发前,成功预测了H1N1在全美的传播,比疾病中心的数据更及时。尽管2013年2月,*Nature* 上有文章表示GFT预测的全国范围的流感样疾病病例(占全国人口的比例)近乎是实际值的2倍,与真实值偏差较大(Butler,2013),但毫无疑问,GFT至少为基于众源地理数据研究流行疾病传播、公共卫生健康提供了一定的引导。

此外,众源地理数据还能用来做导航分析,为人们出行提供帮助。例如,Holone等(2007)综合应用OpenStreetMap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,为行动不便及有个人偏好的行人提供路线设计。也有学者指出,当前大多数众源地理数据的研究集中于数据准备、数据融合、异常值处理和数据平滑等数据处理工作,较少关注社会经济层面的问题(Granell et al.,2016),但是,可以肯定的是,众源地理数据的价值已然越发明显地体现在城市交通、社会经济学、人类学等学科研究中。