DOI：10.13203/j.whugis20140148 文章编号：1671-8860(2014)06-0645-10

大数据与广义 GIS

陆 锋[1] 张恒才[1]

1 中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室，北京，100101

摘 要：普适计算基础设施和数据处理技术的发展催生了大数据概念，而大数据时空粒度的不断细化加速了地理空间信息的泛化过程。阐述了大数据时代地理空间信息泛化的显著特征，进而提出 GIS 概念广义化的迫切需求，从数据采集与整理、数据管理与集成、数据分析与计算三个方面分析了广义 GIS 所面临的技术挑战，重点探讨了互联网蕴含地理空间数据采集、移动对象数据库和异构动态数据管理、移动对象轨迹数据挖掘、复杂网络分析等方面的研究进展与存在的问题，并展望了广义 GIS 时代地理计算与城市计算、社会计算的融合趋势。

关键词：广义地理信息系统；互联网文本搜索；移动对象数据库；轨迹数据挖掘；复杂网络

中图法分类号：P208 **文献标志码**：A

随着集成电路与芯片、传感器网络、移动定位、无线通讯、移动互联网、高性能计算与存储技术的快速发展与普及，数据采集与计算单元的外延不断扩展，地球电子皮肤[1]、人人都是传感器[2]的梦想正在逐步付诸实践。普适计算基础设施和技术的日益完善为各种信息技术应用的推陈出新提供了无限可能，同时也大幅降低了很多专业数据采集过程的技术壁垒，减弱了政府、行业和大众在信息获取方面的不对称性。在此背景下，全球数据呈现爆发式增长态势。IDC(international data corporation)最新的研究结果显示，全球每 18 个月新增的数据量是人类有史以来全部数据量的总和。到 2020 年，全球每年产生的数据将达到 40 ZB，其中传感器网络产生的数据将超过 40%[3]。而且这些数据中 95%是不精确的、非结构化的数据[4]。业界把这些超出正常处理规模，难以采用传统方法在合理时间内管理、处理并整理成为辅助决策信息的非结构化和半结构化数据称为大数据(big data)，以区别于大众思维中的“海量数据”，并归纳了大数据的 5V 特征(volume, variety, velocity, veracity, value)。

除了传感器网络支持下相对封闭的专业数据采集与处理渠道外，大数据最大的贡献来自于大众。人们无时不刻不在积累着人生数据，生理指标与健康档案、出行轨迹、通讯记录、网络浏览记录、发表言论、消费记录、社交网络关系等全方位、真实地反映着自身、自然环境与社会动态，同时也为传统制造业、互联网与电子商务、金融保险业、零售业、医疗卫生业、交通运输业等开展产品设计与优化、生产流程与调度优化、商品推荐与广告投递、店铺规划与商业分析等实时提供着分析数据源。大数据将成为人们下一个观察人类自身社会行为的“显微镜”和监测大自然的“仪表盘”[5]。

尽管业界一直存在对大数据维护成本、处理原则、技术壁垒、回报周期、隐私侵犯，甚至包括人性伦理影响的担忧，及其对大数据概念有意无意的过度消费，但大数据无疑将重构很多行业的商业思维和商业模式，其价值不言而喻。与此同时，随着定位技术的进步，大数据的位置标签越发精确，空间隐喻越发显著。例如，正是有了卫星定位、WiFi、移动通讯蜂窝定位技术及其微陀螺、加速度计等各种微小传感器的支持，才使得之前虚拟的社交网络系统(social network system, SNS)发展到客观世界与虚拟世界相融合的基于位置的社交网络(location based social network, LBSN)，并成为互联网企业的必争之地。人类生活中所产生的数据有 80%和空间位置有关[6]，具有泛在、互动、非专业、实时、按需服务、SOA 特征的新地理信息时代[7]无疑也将是地理空间大数据(big geo-data)大放异彩的时代。

收稿日期：2014-03-04

项目来源：国家 863 计划资助项目(2013AA120305，2012AA12A211)；国家自然科学基金资助项目(41271408)。

第一作者：陆锋，博士，研究员，博士生导师。研究方向为导航与位置服务、空间数据库技术、交通地理信息系统等。E-mail：luf@lreis.ac.cn

1 地理信息泛化与广义GIS

历经半个世纪,GIS已发展成为由地理信息科学、地理信息技术和地理信息工程组成的综合科学技术体系,功能日臻复杂完善,应用领域不断扩展。地理信息已经成为生活必需品,处理地理信息的各种软件平台也日益增多,这推动着相关应用系统功能和技术水平的不断进步。

然而需要注意的是,一直以来,GIS是以处理分析精确的、位置相对固定的测绘地理空间信息为目标(当然这些信息绝大多数也是非结构化的,如4D产品),并不擅长处理模糊、实时、海量异构的地理空间隐喻信息,即泛化地理信息。大数据时代,地理信息的泛化具有以下两个显著特征。

1) 数据类型丰富多样,异构特征显著

传统的地理信息采集是专业任务,强调几何精确性,测绘遥感是主要的采集手段。而泛在地理信息采集手段更为丰富和自由,传感器网络、个体出行过程、网络行为、消费记录等均可能成为泛在地理信息采集手段,强调非专业性、实时性和全面性,地理空间信息形态很多是隐式的。数据内容涉及自然环境、政务信息、民意调查、商业信息、社会动态、人口流动等,半结构化和非结构化信息大量出现,同时,地理信息个性化特征凸显。

2) 时空粒度不断细化,位置动态关联

传统上的地理信息强调对地表要素的静态描述,涉及几何的时态变化时,可采取类似版本管理的方法,属性数据的动态变化亦不频繁。而泛在地理信息更强调动态性,对于地表要素或区域,强调与其关联的实时属性变化(很多是社会属性,如人口密度、交通状态、温湿度、空气质量、噪声、光照等),对于移动对象,强调其几何位置的连续表达和其他社会属性的实时变化表达,同时,数据时空粒度不断细化,从而使得地理空间概念越发重要。在覆盖范围上,强调从室外到室内、从二维到三维,追求精细化、高动态,强调以位置为核心的时空大数据动态关联。

在泛地理信息时代,实时自然环境、社会环境监测与扫描日趋多样,数据以惊人的速度在增长。然而,如何利用泛在地理空间数据挖掘、信息提取、空间智能技术为决策提供依据,如何实时处理各种UGC(user generated contents)方式产生的地理空间隐喻大数据,使之成为生活助理,这些针对政府智能管理、企业商业决策、大众现代生活的需求很多已经超越了传统GIS的能力范畴。因此,泛地理信息时代GIS的概念也需要泛化,广义GIS应运而生。

广义GIS有两个层面的含义,一是处理广义地理信息的系统,这里的广义地理信息与狭义的测绘地理信息对应,包括各种具有地理空间分布特征的异构信息,如移动传感器网络方式采集的个体出行轨迹,具有位置标签的照片、移动定位指纹库,具有地理空间语义的网络文本,基于位置的复杂社交网络关系与内容等,地理信息的社会属性更加显著;二是广义的地理信息系统,即处理地理位置相关信息的技术系统。事实上,GIS之所以能够长期地持续发展,正是在与IT技术的不断融合过程中寻求广义化,以各种应用形态呈现在用户面前,成为大众生活的组成部分。因此,从早期的管理信息系统、字处理系统到ERP、CIMS,甚至包括竞技体育数据分析与网络游戏平台,地理位置信息的份量在不断提升。我们认为,追求GIS的广义化即广义GIS,是GIS的终极目标。本质上,这也是追求地理控制(GeoControl)的过程。它与传统GIS技术和应用模式不同,将融合到相关技术之中共同发挥作用,渗透到日常生活的方方面面,直至意识不到它的存在[8]。

2 广义GIS技术需求

地理空间数据采集与整理、数据集成与管理、数据分析与计算是GIS的基本内容。在测绘遥感、地理学、计算机科学等学科的共同努力下,经过几十年的发展,传统GIS技术体系已经较为成熟。然而,泛在地理信息的出现对GIS的广义化提出了新的技术需求。

2.1 数据采集与整理

广义GIS所面临的时空数据类型众多,对于数据采集技术的新需求包括室内地图采集技术、真三维可量测地理信息采集技术、移动终端多源定位技术、互联网蕴含的地理信息采集技术等。数据采集技术强调空间无缝、自动化、实时性、非专业、协同交互,发挥群体智慧。同时,需要对数据进行实时清洗和甄别,尽可能去伪存真。其中,互联网蕴含地理空间数据采集技术是目前广义GIS数据采集的薄弱环节。

互联网已成为人类历史上最为庞大的图书馆与知识库,是公众获取与分享信息的重要渠道,同时也是全社会、多领域、广纵深、近实时的动态映像。大量的互联网文本直接或间接表达了地理信息,使得互联网文本成为获取地理信息或地理空

间知识的重要来源。互联网蕴含地理空间数据采集的目标是从网络文本,如网页、论坛、百科、微博客与社交网络消息描述中获取地理对象或事件的空间位置、范围、语义和时空演化特征,以支持与地理对象或用户群体的属性、状态、规模等的关联分析,及其地理知识图和地理语义网络的构建。基本技术流程如图 1 所示。目前,相关研究主要着眼于网络文本蕴含地名分词与消歧[9-11]、地理空间位置推断[12-15]、交通信息提取[16]、传染病空间分布与扩散[17-18]、突发事件识别与实时监测[19-20]等。

虽然自然语言文本分析技术已经比较成熟,然而,在当前以社会大众为主体的 Web 2.0 时代,用户产生的大量文本不可避免地包含文法错误、情感符号、网络用语、个人表达习惯、语言混杂现象及噪声,标准化的文本分析方法难以真正发挥其价值。此外,微博客等社交媒体对文本篇幅的限制又使得文本缺乏上下文线索,也影响了文本包含内容的正确解析。互联网蕴含地理空间数据采集技术需要寻求新的突破。

首先,需要利用地理语境知识,在类型定义、值域提取、结果筛选环节控制地理信息抽取质量,并需设计高效智能方法将自然语言描述的模糊空间关系映射到空间数据模型中。第二,现有的自然语言处理工具难以实现大规模分布式的地理信息实时抽取,需要研究如何降低时空维度的计算复杂度,设计合理的数据存储、索引和更新方式,并加强并行处理技术的应用。第三,对于互联网文本涌现的新地理信息的感知和理解,手动维护知识库的方式已完全不能适应,需要寻求自动化的方法实现地名词典的动态更新与维护,设计不依赖于训练语料的方法。第四,地理空间信息内在的复杂性增加了信息抽取的不确定性,可能导致误差的累积与传递,需要研究如何提高复杂句法结构和隐式语义的理解能力。

对于互联网蕴含地理空间数据采集与处理而言,从专业的角度看,最终的目标是构建类似于 Google 的 Knowledge Graph 及 Facebook 的 GraphSearch 的知识图,当然 GIS 领域更关注的是地理知识图,即如何自动化地探测地理实体间的空间关系与语义关系,实现地理信息的自动聚合过程。此外,当获取了地理实体很多的资源信息后,需要有地理实体的领域分类结构图的支持,才能构建起最后的地理知识图结构,建立地理语义网(geographic semantic Web)。图 2 展示了笔者所研发的地理知识图构建所涉及的网络搜索地理信息分层。

2.2　数据管理与集成

与海量数据最大的区别在于,大数据更强调数据的多源异构性和动态性,而不仅仅是数据规模。即使是中小企业和社会公众,在普适计算时代,依然有着迫切的大数据需求。广义 GIS 所涉及的时空大数据包括多源地理空间信息、全景实景影像、视频、移动对象轨迹、社交网络关系、空间隐喻文本、生活服务信息、个性化地理信息等,对数据管理与集成提出了更高的要求。

1) 在数据表达方面,由于大数据天然的异构性,来源混杂,传统 GIS 中对于地理实体和过程的精确几何描述难以实现,数据融合压力极大,这对地理空间认知方法提出了新的挑战。而且由于传感器的复杂性、多样性、多尺度性和不确定性,利用现有的 GIS 数据模型难以真正实现物联网环境下传感器采集信息的描述与管理[21]。此外,对于多维复杂地理对象的统一表达,需要重新设

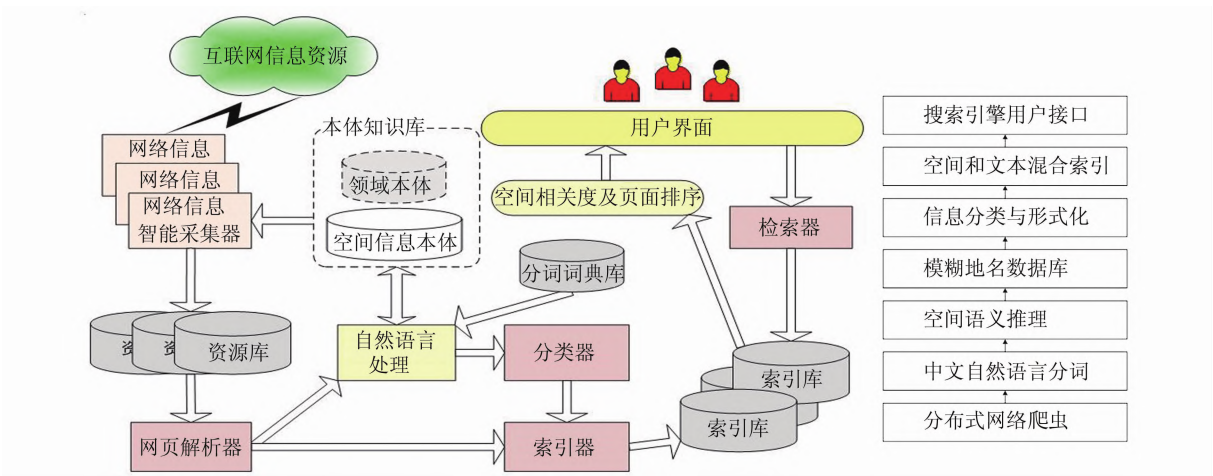

图 1　互联网蕴含地理空间数据采集基本技术流程

Fig. 1　Flowchart of Internet Text Mining for Geo-spatial Data Collection

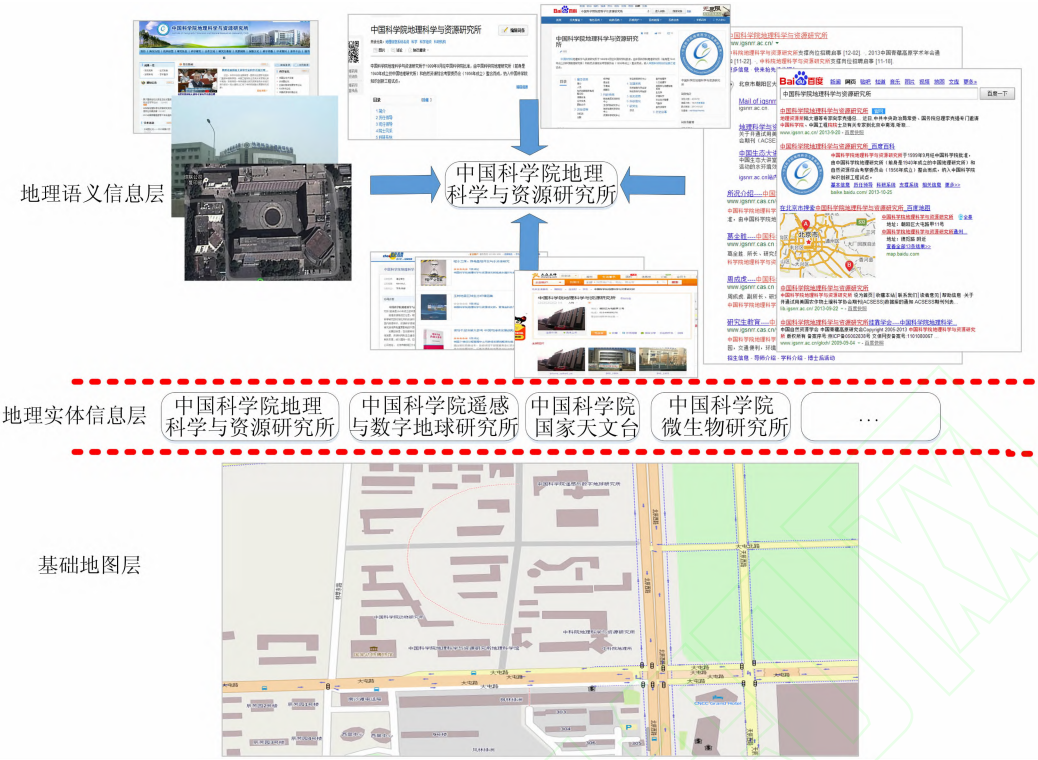

图 2 地理知识图构建中网络搜索地理信息分层示例

Fig. 2 Geo-information Layer for Geo-knowledge Graph Building with Internet Text Mining

计支持复杂地理计算的多维空间数据模型[22]，针对某些应用场景，甚至需要考虑从平面几何过渡到微分几何模型。

2）长期以来，由于数据采集手段和应用需求不足，移动对象数据库(moving object database，MOD)在 GIS 领域没有引起足够的重视。随着目前各种移动监测系统应用的深入，海量移动对象将成为广义 GIS 非常重要的数据源，而移动对象管理与分析技术在 GIS 领域技术的积累比较薄弱。例如，由于无线通讯业务剧增而数据处理能力未能同步增长，美国国家安全局搜集国民通话记录的比例已经由 2006 年的 100%降至目前的 30%以下，但每年的记录仍然在数十亿以上[23]。笔者所实现的用于实时交通信息处理的特大城市移动通讯信令采集系统，每 min 的定位记录数高达数百万条，仅仅只包括通话号码、通话时间、通话位置等简单的文本信息，也有上百兆数据量。如何有效管理这些数据，并支持多种应用所急需的移动对象连续轨迹查询、密度查询、近邻查询、Geo-social 联合查询，以辅助群体与个体决策，对 GIS 和数据库领域都是一个严峻的挑战，亟需加强移动对象数据模型、位置更新策略、索引、实时查询、位置预测、不确定性、隐私保护等 7 个方面的研究与实践[24]。

3）对于与传统地理空间数据完全不同且实时变化的移动对象轨迹、文本、照片、视频、社交关系等数据的管理与实时处理，目前成熟的扩展 RDBMS 技术方案无能为力，以 HBase、MongoDB、Neo4J 等列数据库、文档数据库和图数据库为代表的 NoSQL 数据库技术能够实现海量异构数据的高效存储与访问，且具有高扩展性与高并发特点，有可能脱颖而出。图 3 展示了笔者研发的基于 MongoDB 的移动对象数据库系统架构。图 4 为所部署的基于云计算的移动对象轨迹数据处理平台 Trajectory Cloud 的系统架构。该平台以 Hadoop 为基础，增加了对应的 SQL 解析功能，数据缓存层采用 Redis 解决方案，采用 MongoDB 存储轨迹数据。需要注意的是，虽然现有的 NoSQL 比经典的关系数据库系统的存储效率有所提升，但是在数据模型、索引及查询方面都还存在很多不足，需要深入研究。

此外，大数据的真正价值在于各种异构数据之间的关联性。而数据之间众多的关系会带来大量的连接(join)操作，传统的 RDBMS 和现有的大数据系统如 Hadoop 均难以胜任。而 NoSQL 阵营中的重要分支——图数据库系统对数据关联处理具有先天优势。现有的很多应用处理的数据都具有图结构特征，而且这种趋势随着应用及数据的日趋复杂变得愈来愈明显。目前对图数据的研究仍然处于起步阶段。研究重点包括图算法、图系统结构、图数据基本操作、图数据访问模式等[25]。

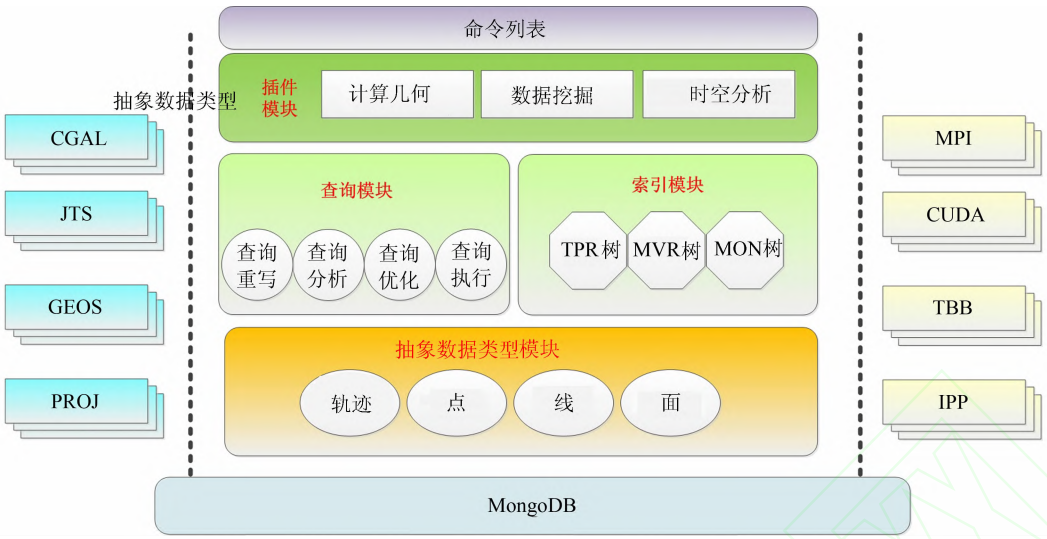

图 3　基于 MongoDB 的移动对象数据库系统架构

Fig. 3　System Structure of Moving Object Database Based on MongoDB

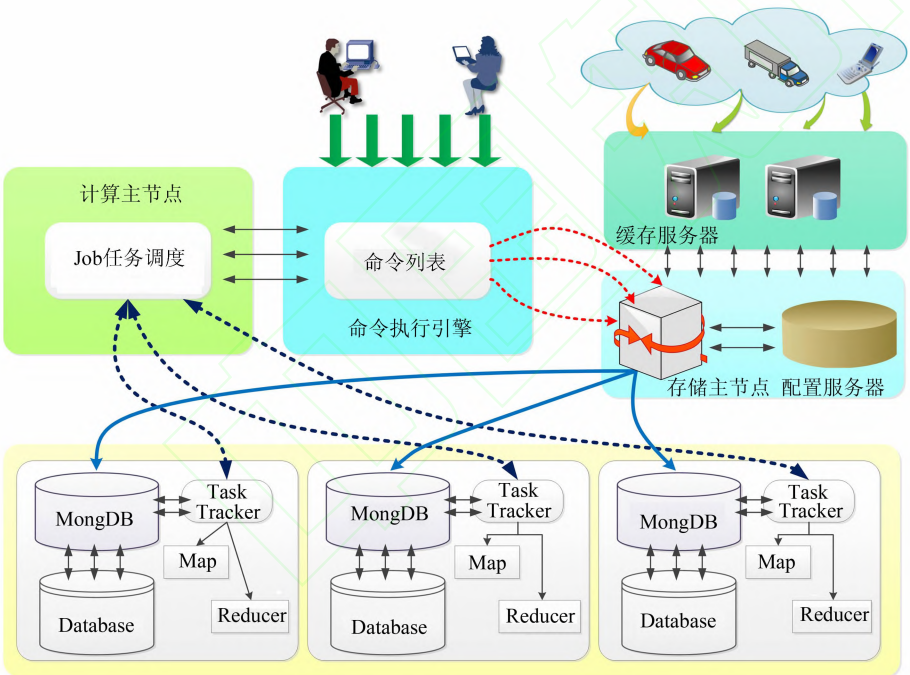

图 4　移动对象轨迹数据处理平台 Trajectory Cloud 系统架构

Fig. 4　System Structure of Moving Object Trajectory Processing Platform

2.3　数据分析与计算

除了传统的时空数据分析需求外,广义 GIS 在数据分析与计算方面需要从基于位置的多要素时空关联、海量移动对象轨迹分析、复杂网络的地理空间特征分析等方面入手,加强人工智能、机器学习等领域的理论研究和技术研发,丰富广义 GIS 的分析能力。这也是大数据产业的核心和生命力所在。此外,由于大数据处理更强调全数据模式,以尽可能发现数据中隐藏的细节或异常,并强调数据处理效率比数据的精确性更为重要。这将极大地撼动传统的随机抽样和数据分析的理念[4]。

个性化、智能化服务需要多维度、无缝、精细时空粒度的地理信息的支持,亟需架构灵活的信息聚合与服务平台。基于位置的多要素时空关联旨在异构数据集成与管理技术的支持下,研究以位置为锚点的多要素(显式和隐式地理空间信息、移动对象信息、社交网络信息、社会环境信息、商业信息等)的实时聚合与关联分析方法,在全息位置地图平台[26]上开展时空大数据分析与预测。

移动对象不仅限于通过各种方式定位的交通运输工具或通讯终端,还包括个体出行或迁移记

录、信用卡消费记录、流通纸币等。移动对象的运动轨迹在一定程度上反映了个体或群体的意图、喜好和空间行为模式。从大量的移动对象轨迹数据中提取出蕴含的关联规则和时空序列模式,对社会动态研究和个性化推荐至关重要。应用包括路网交通状态分析[27]、城市交通出行模式分析[28]、空间相互作用模型标定[29]、人类行为时空特性[30]、群体事件监测等[31]。

移动对象轨迹数据挖掘主要包括轨迹聚类、轨迹频繁模式挖掘、轨迹异常检测三个研究方向。轨迹聚类的目的是对轨迹集合进行分组,使得组内轨迹之间具有较高的相似度,而组间的差别较大。业界提出了很多轨迹聚类算法[32-34]。地理约束在轨迹聚类中也得到了广泛关注[35-36]。轨迹频繁模式挖掘旨在发现在同一时间切面中移动对象所呈现的聚合模式及该模式的时间演化规律[37]。微软亚洲研究院近年来在轨迹数据频繁集挖掘方面开展了深入研究,涉及临近搜索[38]、聚集模式[39]、旅伴搜索[40]、出行模型分类[41]等。在轨迹异常检测方面,目前业界已发展出一批成熟算法,如iBAT[42]、RTOD[43]、iBOAT[44]等。文献[45]系统评述了轨迹异常检测方法的研究进展,文献[46]对移动对象轨迹数据挖掘的常用方法进行了分类和评述。近年来,在大数据背景下,出现了一批新兴的移动对象轨迹数据挖掘方法,包括非参数建模[47]、流数据挖掘[48-49]、图数据挖掘[50]等。另外,基于大数据的可视化分析[51-52]与其他学科的交叉研究(如DNA测序[53])也为移动对象轨迹分析提供了新的思路。

复杂网络科学是研究城市和社会动态卓有成效的理论方法,研究内容涉及网络几何性质、形成机制、结构稳定性、演化规律与动力学机制等,多采用图论和统计物理学研究方法。其中,研究网络中顶点与边的度值及权值等微观性质与网络的几何性质、效率与稳定性等宏观性质之间的关系是复杂网络研究的核心[54]。复杂网络研究在物理学、社会学、交通工程学科非常活跃,如美国东北大学的Barabási教授领导的团队从复杂系统的角度出发,通过挖掘即时的移动通讯数据,审视社会关系的健壮性及其社会群体对大规模紧急事件的反应,甚至预测人类行为等[55]。北京交通大学高自友教授领导的团队系统研究了出行者博弈、网络结构与城市交通系统的复杂性问题,以期建立城市交通网络演化模型,探明不同网络拓扑上交通流的典型动力学特性等[56]。GIS和计算机学科更注重采用复杂网络科学的一些算子,基于转换网络的度分布、介中心性、社区聚类等研究交通网络的健壮性[57],或挖掘网络个体或群体的喜好与习惯,据此开展个性化推荐[58]。现实中的地理网络的动力学过程可能会呈现出与静态网络和非空间网络极为不同的规律,需要探索这种随时空演化的网络上的动力学特性,以及节点、连边的活跃特性与动力学的关联规律[59]。此外,从GIS学科角度,需要更关注网络的几何形态和空间分布,将地理空间分布引入到经典复杂网络算子和模型中,研究城市与区域问题。

需要注意的是,在大数据时代,需要特别重视机器学习的研究。机器学习被视为大数据的核心[4]。和数据挖掘相比,机器学习的算法不是固定的,而是带有自调适参数,能够随着计算技术的增多,让计算机通过学习自我完善,使挖掘和预测的结果更准确[5]。机器学习是使用计算机模拟人的学习过程,识别现有知识、获取新知识、不断改善性能和实现自身完善的方法。随着统计方法成为研究主流,机器学习技术也被引入到自然语言处理、轨迹数据分析等时空大数据分析中,隐马尔可夫模型、条件随机场、最大熵、支持向量机、贝叶斯等机器学习模型和算法得到广泛应用[60]。例如,采用基于集成学习理论的层叠泛化方法来集成时序数据分析经典模型(LLSR、ARMA、HM、ANN、RBF-NN、SVM),结果表明,层叠泛化学习模型的RMSE与MAE均小于单一模型,同时也小于其他统计混合模型(EW、OW、ME、MV),并且具有极好的开放性,模型成功用于城市路网交通状态时序分析[61]。

3 从城市计算到社会计算

地理空间大数据的出现使得地理计算(geo-computation)受到追捧。而云计算和普适计算在提升地理计算规模、精度和效率的同时,也大幅促进了地理计算和城市计算(urban computing)乃至社会计算(social computing)的交叉和融合。

地理计算旨在利用各种不同类型的地理空间数据,基于高性能计算发展相关的工具和模型,以解决地理空间问题[62]。城市计算是将遍布城市的传感器、建筑、路网、车辆和居民均作为计算单元,协同完成城市级别的计算过程。通过城市感知、数据挖掘、智能提取和服务提供四个环节感知城市脉搏,为城市居民提供更美好的城市生活,同时也让城市变得更加智能[63]。关于城市计算的最新进展可参见文献[64,65]。同时,由于个体在

真实社会中的活动得到了前所未有的记录,记录的时空粒度在不断细化,为社会科学的定量分析提供了极为丰富的数据,帮助加深对社会生活、机构组织和社会的理解[66]。社会计算旨在研究如何利用计算技术研究社会系统的运行规律与发展趋势,并帮助人们进行沟通与协作。研究内容包括社交网络、社会媒体内容计算、群体智慧、人工社会等。

从广义 GIS 的角度看,很多城市问题和社会问题本质上就是地理空间问题。城市计算和社会计算都是涉及复杂系统、数据挖掘、网络科学、社会学、管理科学、自然语言处理、信息检索等多个学科的交叉研究领域,可以理解为大数据技术的城市与社会应用过程。其中很多研究场景和内容与大数据的空间隐喻息息相关。从面向地理空间的计算到面向城市场景的计算,乃至面向全社会场景(自然环境+社会环境)的计算,可以预见,地理计算与城市计算、社会计算必将融为一体,服务于现代社会的管理与大众生活,这也是 GIS 社会化[22]发展的必然趋势。

4　结　语

经过多年的发展,地理空间数据表达、管理与分析方法已经比较成熟。然而,随着大数据时代的来临,对 GIS 社会化的应用又提出了更高的要求。基于位置的社交网络、城市计算、社会计算等新兴领域的不断涌现,云计算、物联网、普适计算基础设施不断成熟,移动对象轨迹、网络行为、社交关系等作为典型大数据的重要性会越加凸显,同时也将极大地推进广义 GIS 的核心研究内容,如网络文本蕴含地理信息搜索、海量移动对象管理与轨迹数据挖掘、复杂时空网络分析、图像内容时空语义分析等方面的研究进程。上述研究已经成为当前业界及学界共同关注的问题,大量新出现的理论和技术问题亟待解决。从 GIS 到广义 GIS,挑战才刚刚开始。

参　考　文　献

[1] Gross N. The Earth will Don an Electronic Skin [EB/OL]. http://www.businessweek.com/1999/99_35/b3644024.htm,1999

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

[3] Adshead A. Data to Grow More Quickly[EB/OL]. http://www.computerweekly.com/news/2240174381/Data-to-grow-more-quickly-says-IDCs-Digital-Universe-study,2012

[4] Mayer-Schönberger V, Cukier K. Big Data[M]. Sheng Yangyan, Zhou Tao. Hangzhou: Zhejiang People's Press,2013 (维克托·迈尔-舍恩伯格,肯尼思·库克耶. 大数据时代[M]. 盛杨燕,周涛. 杭州:浙江人民出版社,2013)

[5] Tu Zipei. The Big Data Revolution[M]. Guilin: Guangxi Normal University Press,2013(涂子沛. 大数据:正在到来的数据革命[M]. 桂林:广西师范大学出版社,2013)

[6] Xu Guanhua. Pay Much Attention to the Digital Earth[J]. *Science News Weekly*, 1999(1):7-8(徐冠华. 全社会要高度关注"数字地球"[J].《科学新闻》周刊,1999(1): 7-8

[7] Li Deren, Shao Zhenfeng. The New Era of Geographic Information[J]. *Science in China (Series F: Information Sciences)*, 2009, 39(6):579-587(李德仁, 邵振峰. 论新地理信息时代[J]. 中国科学 F 辑(信息科学),2009, 39(6): 579-587)

[8] Zhong Ershun. GeoControl and Live Geography: Some Thoughts on the Direction of GIS[J]. *Journal of Geo-Information Science*, 2013,15(6):783-792(钟耳顺. 地理控制与实况地理学:关于 GIS 发展的思考[J]. 地球信息科学学报,2013,15(6):783-792)

[9] Tang Xuri, Chen Xiaohe, Zhang Xueying. Research on Toponym Resolution in Chinese Text[J]. *Geomatics and Information Science of Wuhan University*, 2010, 35(8):930-935(唐旭日,陈小荷,张雪英. 中文文本的地名解析方法研究[J]. 武汉大学学报·信息科学版, 2010, 35(8):930-935)

[10] Zhang Xueying, Zhu Shaonan, Zhang Chunju. Annotation of Geographical Named Entities in Chinese Text[J]. *Acta Geodaetica et Cartographica Sinaca*, 2012,41(1): 115-120(张雪英,朱少楠,张春菊. 中文文本的地理命名实体标注[J]. 测绘学报, 2012,41(1): 115-120)

[11] Mónica M, Julián U, Sonia S C, et al. Named Entity Recognition: Fallacies, Challenges and Opportunities[J]. *Computer Standards & Interfaces*, 2013, 35(5): 482-489

[12] Cheng Z, Caverlee J, Lee K. You are Where you Tweet: a Content-based Approach to Geo-locating Twitter Users[C]. The 19th ACM International Conference on Information and Knowledge Management, Toronto, Canada, 2010

[13] Ikawa Y, Enoki M, Tatsubori M. Location Inference Using Microblog Messages[C]. The 21st International Conference Companion on World Wide Web, Lyon, France,2012

[14] Zhang Xueying, Zhang Chunju, Zhu Shaonan. Annotation of Geo-spatial Relations in Chinese Text[J]. *Acta Geodaetica et Cartographica Sinica*, 2012,41(3):468-474(张雪英,张春菊,朱少楠. 中文文本的地理空间关系标注[J]. 测绘学报,2012,41(3):468-474)

[15] Vasardani M, Winter S, Richter K F. Locating Place Names from Place Descriptions[J]. *International Journal of Geographical Information Science*, 2013, 27(12): 1-24

[16] Zhang Hengcai, Lu Feng, Chen Jie. Extracting Traffic Information from Massive Micro-blog Messages[J]. *Journal of Image and Graphics*, 2013, 18(1):123-129(张恒才,陆锋,陈洁. 微博客蕴含交通信息的提取方法研究[J]. 中国图像图形学报,2013,18(1):123-129)

[17] Gomide J, Veloso A, Meira Jr W, et al. Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter[C]. ACM Conference on Web Science, Koblenz, Germany,2011

[18] Sadilek A, Kautz H A, Silenzio V. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data[C]. The 26th AAAI Conference on Artificial Intelligence,Toronto,Canada,2012

[19] Sakaki T, Okazaki M, Matsuo Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors[C]. The 19th International Conference on World Wide Web, Raleigh, USA, 2010

[20] Watanabe K, Ochi M, Okabe M, et al. Jasmine: A Real-time Local-event Detection System Based on Geolocation Information Propagated to Microblogs [C]. The 20th ACM International Conference on Information and knowledge Management, Glasgow, Scotland, UK, 2011

[21] Gong Jianya, Wang Guoliang. From Digital City to Smart City: New Challenges to Geographic Information Technology[J]. *Journal of Geomatics*, 2013,38(2):1-6(龚健雅,王国良. 从数字城市到智慧城市:地理信息技术面临的新挑战[J]. 测绘地理信息,2013,38(2):1-6)

[22] Lv Guonian, Yuan Linwang, Yu Zhaoyuan. Challenges to Development and Socialization of GIS Technology[J]. *Journal of Geo-Information Science*, 2013,15(4):483-490(闾国年,袁林旺,俞肇元. GIS技术发展与社会化的困境与挑战[J]. 地球信息科学学报,2013,15(4):483-490)

[23] Nakashima E. NSA is Collecting Less than 30 Percent of U. S. Call Data[DB/OL]. The Washington Post,2014-02-08

[24] Zhang Hengcai, Lu Feng, Chen Jie. Advance in Moving Object Data Modeling under Geographic Network Environment[J]. *Journal of Geo-Information Science*, 2013, 15(3): 328-337(张恒才,陆锋,陈洁. 路网空间移动对象模型的应用与发展[J]. 地球信息科学学报, 2013, 15(3):328-337)

[25] Wang Haixun, Management and Mining of Graph Data[J]. *Communications of CCF*, 2012,8(11): 10-11(王海勋. 图数据的管理与挖掘[J]. 中国计算机学会通讯,2012,8(11):10-11)

[26] Zhou Chenghu, Zhu Xinyan, Wang Meng, et al. Panoramic Location Based Map[J]. *Progress in Geography*, 2011,30(11):1 331-1 335(周成虎,朱欣焰,王蒙,等. 全息位置地图研究[J]. 地理科学进展,2011. 30(11):1 331-1 335)

[27] Li Qingquan, Yin Jianzhong, He Fenqin. A Coverage Rate Model of GPS Floating Car for Road Networks[J]. *Geomatics and Information Science of Wuhan University*,2009, 34(6):715-718(李清泉,尹建忠,贺奋琴. 面向道路网的GPS浮动车覆盖率模型研究[J]. 武汉大学学报·信息科学版,2009,34(6):715-718)

[28] Liu Y, Kang C, Gao S, et al. Understanding Intra-urban Trip Patterns from Taxi Trajectory Data[J]. *Journal of Geographical Systems*, 2012, 14(4): 463-483

[29] Yue Y, Wang H D, Hu B, et al. Exploratory Calibration of a Spatial Interaction Model Using Taxi GPS Trajectories[J]. *Computers, Environment and Urban Systems*, 2012, 36(2): 140-153

[30] Ganti R, Mudhakar S, Ranganathan A, et al. Inferring Human Mobility Patterns from Taxicab Location Traces[C]. The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland, 2013

[31] Bagrow J P, Wang D, Barabasi A L. Collective Response of Human Populations to Large-scale Emergencies[J]. *PLoS ONE*, 2011, 6(3):1-8

[32] Somayeh D, Patrick L, Robert W. Movement Similarity Assessment Using Symbolic Representation of Trajectories[J]. *International Journal of Geographical Information Science*, 2012, 26(9): 1 563-1 588

[33] Zeinalipour-Yazti D, Laoudias C, Costa C, et al. Crowdsourced Trace Similarity with Smartphones [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 25(6): 1 240-1 253

[34] Pei T, Gong X, Shaw S L, et al. Clustering of Temporal Event Processes[J]. *International Journal of Geographical Information Science*, 2013, 27(3): 484-510

[35] Onnela J P, Arbesman S, González M C, et al. Geographic Constraints on Social Network Groups[J].

PLoS ONE, 2011, 6(4): e16939, doi:10.1371/journal.pone.0016939

[36] Hung C C, Peng W C, Lee W C. Clustering and Aggregating Clues of Trajectories for Mining Trajectory Patterns and Routes[J]. *The VLDB Journal*, 2011,20(5):1-24

[37] Liu H Y, Lin Y, Han J W. Methods for Mining Frequent Items in Data Streams: An Overview[J]. *Knowledge and Information Systems*, 2011, 26(1): 1-30

[38] Tang L A, Zheng Y, Yuan J, et al. Retrieving *k*-nearset Neighboring Trajectories by a set of Point Locations[C]. The 12th International Symposium on Spatial and Temporal Databases, Minneapolis, USA, 2011

[39] Zheng K, Zheng Y, Yuan J, et al. On Discovery of Gathering Patterns from Trajectories[C]. The 29th International Conference on Data Engineering, Brisbane, Australia,2013

[40] Tang L A, Zheng Y, Yuan J, et al. On Discovery of Traveling Companions from Streaming Trajectories[C]. The 28th International Conference on Data Engineering, Washington D C, USA, 2012

[41] Zheng Y, Chen Y, Li Q, et al. Understanding Transportation Modes Based on GPS Data for Web Applications[J]. *ACM Transactions on the Web*, 2010, 4(1):1- 36

[42] Zhang D, Li N, Zhou Z, et al. iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces[C]. The 2011 ACM International Conference on Ubiquitous Computing, Beijing, China, 2011

[43] Liu L X, Qiao S J, Zhang Y P, et al. An Efficient Outlying Trajectories Mining Approach Based on Relative Distance[J]. *International Journal of Geographical Information Science*, 2012, 26(10): 1 789-1 810

[44] Chen C, Zhang D, Castro P S, et al. iBOAT: Isolation-Based Online Anomalous Trajectory Detection [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2013,14(2):806-818

[45] Chandola V, Banerjee A, Kumar V. Anomaly Detection for Discrete Sequences: A Survey[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2012, 24(5): 823-839

[46] Long J A, Nelson T A. A Review of Quantitative Methods for Movement Data [J]. *International Journal of Geographical Information Science*, 2013, 27(2):292-318

[47] Liu X L, Lu F, Zhang H C, et al. Intersection Delay Estimation from Floating Car Data via Principal Curves: A Case Study on Beijing's Road Network [J]. *Frontiers of Earth Science*, 2013, 7(2): 206-216

[48] Gama J, Sebastião R, Rodrigues P P. On Evaluating Stream Learning Algorithms [J]. *Machine Learning*, 2013, 90(3):317-346

[49] Sarah M, Andrew D B, David D B, et al. Developing Systems for Real-Time Streaming Analysis[J]. *Journal of Computational and Graphical Statistics*, 2012, 21(3):561-580

[50] Baruch B, Barabási A L. Network Link Prediction by Global Silencing of Indirect Correlations [J]. *Nature Biotechnology*, 2013, 31:720-725

[51] Lee J Y, Kwan M P. Visualization of Socio-spatial Isolation Based on Human Activity Patterns and Social Networks in Space-time [J]. *Tijdschriftvoor Economische en Sociale Geografie*, 2012, 102(4): 468-485

[52] Wang Z C, Lu M, Yuan X R, et al. Visual Traffic Jam Analysis Based on Trajectory Data[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(12):2 159-2 168

[53] Jawad A, Kersting K, Andrienko N. Where Traffic Meets DNA: Mobility Mining Using Biological Sequence Analysis Revisited[C]. The 19th International Conference on Advances in Geographic Information Systems, Chicago, USA, 2011

[54] Wu Jinshan, Di Zengru. Complex Networks in Statistical Physics[J]. *Progress in Physics*, 2004, 24(1):18-46(吴金闪,狄增如.从统计物理学看复杂网络研究[J]. 物理学进展,2004,24(1):18-46)

[55] Barabási A L. Bursts: the Hidden Pattern Behind Everything We Do[M]. Ma Hui. Beijing: China Renmin University Press(艾伯特-拉斯洛·巴拉巴西.爆发:大数据时代预见未来的新思维[M]. 马慧. 北京:中国人民大学出版社,2012)

[56] Gao Ziyou, Wu Jianjun. Travelers Game, Network Structure and Urban Traffic System Complexity[J]. *Complex Systems and Complexity Science*, 2010,7(4):55-64(高自友,吴建军.出行者博弈、网络结构与城市交通系统复杂性[J]. 复杂系统与复杂性科学,2010,7(4):55-64)

[57] Duan Y Y, Lu F. Structural Robustness of City Road Networks Based on Community[J]. *Computers, Environment and Urban Systems*, 2013, 41:75-87

[58] Bao J, Zheng Y, Mokbel M F. Location-based and Preference-Aware Recommendation Using Sparse Geo-Social Networking Data[C]. The 20th International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 2012

[59] Zhou Tao, Han X P, Yan Xiaoyong, et al. Statisti-

cal Mechanics on Temporal and Spatial Activities of Human[J]. *Journal of University of Electronic Science and Technology of China*, 2013, 42(4): 481-540 (周涛,韩筱璞,闫小勇,等. 人类行为时空特性的统计力学[J]. 电子科技大学学报,2013,42(4): 481-540)

[60] Wu Jun. Beauty of Mathematics[M]. Beijing: Posts & Telecommunications Press,2012(吴军. 数学之美[M]. 北京:人民邮电出版社,2012

[61] Liu X L, Lu F, Zhang H C. Estimating Beijing's Intersection Delays from Floating Car Data: A Boosting Approach [C]. The 12th International Conference on Geocomputation, Wuhan, China, 2013

[62] Openshaw S, Abrahart R J. GeoComputation[M]. New York: Taylor & Francis Ltd, 2000

[63] Zheng Yu. Features and Scopes for Urban Computing[J]. *CEO & CIO*, 2012(1):76-77 (郑宇. 城市计算的内涵和边界[J]. IT经理世界,2012(1):76-77)

[64] Zheng Yu. Urban Computing and Big Data[J]. *Communications of CCF*, 2013, 9(8):6-16(郑宇. 城市计算与大数据[J]. 中国计算机学会通讯,2013,9(8):6-16)

[65] Zhang Daqing, Chen Chao, Yang Dingqi, et al. From Digital Footprints to Urban Computing[J]. *Communications of CCF*, 2013, 9(8):17-24 (张大庆,陈超,杨丁奇,等. 从数字脚印到城市计算[J]. 中国计算机学会通讯,2013,9(8):17-24)

[66] Lazer D, Pentland A, Adamic L, et al. Life in the Network: the Coming Age of Computational Social Science[J]. *Science*, 2009, 323(5 915): 721-723

Big Data and Generalized GIS

LU Feng[1] *ZHANG Hengcai*[1]

1 State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

Abstract: The development of an ubiquitous computing infrastructure and data processing technologies are giving birth to big data, and the continuous refining of the spatial-temporal granularities of big data speeds up the generalization of geo-spatial information. In this paper, the distinctive characteristics of generalized geo-spatial information are investigated, and the urgent need for a more generalized concept of GIS are clarified. Then the technical challenges for general GIS are set forward in terms of data collection and cleaning, data management and integration, and data analysis and computing. The progress is summarized and the research issues are discussed in geo-spatial data collection with Internet text mining, moving object database, dynamic and heterogeneous data management, moving trajectory data mining, and complex network analysis. The fusion of geocomputation, urban computing and social computing in the near future is considered at the end of the paper.

Key words: generalized GIS; Internet text mining; moving object database; trajectory mining; complex network

First author: LU Feng, PhD, researcher, PhD supervisor. His research interests cover location based services, spatial DBMS, and GIS for transportation. E-mail: luf@lreis.ac.cn

Foundation support: The National High Technology Research and Development Program of China(863 Program), Nos. 2013AA120305, 2012AA12A211; the National Natural Science Foundation of China, No. 41271408.