DOI:10.13203/j.whugis20140150　　文章编号:1671-8860(2014)06-0641-04

大数据 GIS

李清泉[1,2]　李德仁[2]

1　深圳大学空间信息智能感知与服务深圳市重点实验室,广东 深圳,518060
2　武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079

摘　要:大数据不仅使世界认识到数据的重要性,更引发了许多行业从根本上的变革。大数据时代也对 GIS 提出了诸多挑战,如海量、多源、异构数据的存储与管理以及面对大量噪音的价值挖掘等。作为空间数据管理、分析以及可视化的重要工具,为适应大数据的需求,GIS 必须在大数据时代做出改变和调整。针对大数据的几个"V"特性分析了传统 GIS 所受到的挑战,在前期相关研究的基础上,从 GIS 空间数据管理、空间数据分析以及可视化三方面进一步阐述了大数据 GIS 应具有的特征。

关键词:大数据;GIS;空间数据管理;空间数据分析;可视化
中图法分类号:P208　　**文献标志码**:A

1　大数据时代

大数据在近两年备受关注。由于各类传感器日益普及,通讯技术的飞跃以及网络基础设施的高速发展,越来越多的领域如金融、电商/广告、医疗、生物、物流等开始有意识地收集和积累大量数据,并从中挖掘以前不曾也不可能触及的价值。过去两年所产生的数据量为有史以来所有数据量的 90%[1],其中 2013 年中国产生的数据总量超过 0.8 ZB(相当于 8 亿 TB),是 2012 年所产生的数据量的 2 倍,相当于 2009 年全球的数据总量。大数据中约有 80%的数据与空间位置有关。目前,各种地图和地理位置紧密相关的信息应该每年在数十 PB,一个智慧城市的数据一个季度就是 200 PB 之多。其中高清摄像头一个月产生 1.8 TB 数据,每天北京市的视频采集数据量在 3 PB 左右,一个中等城市每年视频监控产生的数据在 300 PB 左右;国家电网年均产生数据 510 TB(不含视频),累计产生数据 5 PB。其他像搜索、地图、社交、影视娱乐类等互联网公司也拥有 PB 量级的数据储备[2]。

大数据绝不仅仅是数据量大,现在对大数据特征的描述已经从 3V (Volume,Velocity,Variety)[3]、4V(增加 Veracity)(http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg)到 5V(增加 Value)(http://www.ibmbigdatahub.com/infographic/four-vs-big-data),甚至是新的 3V(Veracity,Visualization,Value)[4]。

1) Volume(体量大):大量 TB 级以上已有的数据等待处理;

2) Velocity(速度快):需要响应以 s 甚至 ms 计的流数据不断产生;

3) Variety(模态多样):数据来源和类型繁多,文本、图片、视频等结构化和非结构化数据并存;

4) Veracity(真伪难辨):由于数据的噪音、缺失、不一致性、歧义等引起的数据不确定性;

5) Value(价值):大数据使得人们以前所未有的维度量化和理解世界,蕴含了巨大的价值,大数据的终极目标在于从数据中挖掘价值。

其中,数据的流质特性是因,数据量庞大是果,多样性和真实性是挑战,价值是目标。"海量数据"多年来就一直是 GIS 领域的一个重要问题。对于从海量数据到大数据的跨越,除了数据量庞大,还有快速、异构、分析和挖掘等关键问题。因此,在海量数据研究的基础上,大数据为 GIS 带来了挑战的同时也提供了机遇。

收稿日期:2014-02-24
项目来源:国家自然科学基金资助项目(41371377);深圳市科技研发资金资助项目(ZDSY20121019111146499,JSGG20121026111056204);深圳市战略性新兴产业发展专项资金资助项目(JCYJ20120191111128765)。

第一作者:李清泉,博士,教授。主要研究方向为时空数据挖掘、多传感器集成数据采集、精密工业及工程测量。E-mail:liqq@szu.edu.cn

2 传统GIS面临的挑战

2.1 大数据体量

现有成熟的GIS多依赖于关系型数据库,但是关系型数据库由于在海量数据管理、高并发读写以及扩展性等方面的限制,在大数据时代已经显示出一定的局限性。云端服务模式已经成为解决大数据存储和管理的技术趋势,这对空间大数据的异地多点查询和数据关联与聚合等提出了挑战。在云环境下,数据可能存放在不同磁盘、不同机器甚至不同地点,现有的分布式文件系统、数据索引与查询的方法都具有局限性。所以,针对空间大数据的数据划分,基于内存的索引,历史、当前及未来时空索引的并发控制,以及基于多线程的并发连续查询等仍然是亟需深入研究的问题。

2.2 流质特性

传统的空间数据库以相对静态的数据为主,不能满足流数据的需求。所谓流数据(streaming data)也称数据流,是按照时间序列动态增加的数据观测值向量所组成的数据序列,具有连续性及无限增长性。典型的与空间相关的流数据包括环境、水文、交通等传感器所产生的数据。而现有的空间数据存储是静态的关系型数据记录的结合,具有详细定义的结构、限定的大小及数据持久性,且目前的空间数据查询及分析主要针对可控制的操作,查询为静态的一次查询,所以,现有的空间数据管理方式难以应对高动态的空间流数据。

2.3 异构数据

与空间位置相关的传感器随着应用的不同,类型多样,采集的内容也千差万别,且常具有不同的时间或空间粒度,从数据格式到存储方法都存在着很大差异。虽然多源异构数据也是GIS中的一个经典命题,但是更具挑战性的是越来越多的非结构化数据,如实景图片/街景、3D模型、视频等。传统GIS几乎没有涉及到非结构化数据,常用的关系型数据库也难以管理和使用非结构化数据。对结构化与非结构化数据进行统一的管理、分析和利用是大数据GIS面临的另一个挑战。

2.4 不确定性与价值发现

大数据的真正价值在于数据中所蕴藏的信息和知识。空间现实世界是一个多参数、非线性、时变的不稳定系统,空间数据中的不确定性是无法回避的问题,尤其是其中占重要比重的统计方法。统计的出发点是以样本描述或者推测整体以及识别模式的分布,本质是陈述事实,进行的是确定性分析;而大数据中大量存在的噪音会导致以样本描述整体的偏差可能性大为增加,很可能会增加错误发现的风险或者忽略所需发现的问题,导致某些统计工具失效,以致辅助决策发生偏差甚至错误。正如大数据"巫师"Nate Silver所言,人们常常对数据产生虚有的危险幻想,忽略了数据中存在的缺陷。相比传统分析方法,数据挖掘更强调是有价值的发现而非预期描述。但是,目前空间数据挖掘的理论和方法多给予确定集合理论研究确定问题,对不确定性研究不足[5]。所以,针对大数据噪音多及不确定性高的特性,大数据GIS需要重新思考空间统计模型的选择、参数的训练等问题,以及进一步探索空间数据挖掘算法。

3 大数据GIS特征

一直以来,大家公认的GIS与其他信息系统和电子地图的区别在于同时具有空间数据管理能力、空间分析能力以及基于地图的数据可视化能力[6]。大数据没有改变GIS的基本特征,但是对传统GIS提出了巨大的挑战。为顺应大数据的需求,大数据时代的GIS应具有以下基本特征。

3.1 可扩展的动态数据管理方式

在架构层面,大数据体量大、速度快、模态多等特性带来的挑战终将引起GIS数据存储与管理的质变[7]。相对于静态、有限的数据集,大数据GIS的数据存储管理系统需要具备扩展性,以处理动态无限增长的数据的存储和查询问题。目前,Hadoop系统是大数据处理中常见的一个解决方案,但是其性能越来越难以满足日益增长的数据处理需求,且Hadoop使用的MapReduce模型更适合简单的统计,无法高效地支持更多的算法逻辑。随着大数据处理架构走向多样化,MapReduce框架一统天下的局面将逐渐被打破,因此,随着大数据处理架构的发展,相应的空间数据存储与管理架构也会以开源架构为主逐渐走向多样性。

在数据层面,针对传统关系型数据库难以适应大数据可扩展和非结构化的要求,以及云计算部署环境的问题,近年以键-值数据库(K-V store)为代表的非关系型(NoSQL)数据库迅速发展,如BigTable、HBase、MongoDB以及Cassandra等。此类数据存储方式不用事先修改结构定义,可以自由添加字段,但是会导致数据冗余。虽

然关于非关系型数据库的争议一直存在,但是非关系型数据库已经成为和关系数据库并存的空间数据存储及管理的方式之一。也许未来会出现一个可以同时适应结构化和非结构化的统一的数据模型,或者新型的通用的数据管理系统,或是针对某一应用领域定制的大数据 GIS 管理方案。

在数据处理层面,大数据 GIS 的数据源头以秒、分为间隔采集,且经年累月不间断,数据无限增长,长期积累的数据不可能全部存储在可随机访问的磁盘或内存中。当数据继续不断积累后,必须采用一定的数据粗筛策略,即数据通常在存储前需要进行预处理,保留有价值的信息;原始数据经过处理后,要么被丢弃,要么被存储,但是存储后再次提取代价昂贵。这个预处理过程通常以应用为导向,需要构建适于实时分析的概要结构、时空聚合和多尺度表达等方法,实现高效的数据筛选和聚合机制,以解决数据冗余及噪音问题。这个过程还需要和数据挖掘相结合,如从视频数据中根据知识可以剔除无关人群和车辆,只保留可疑对象,这样可以大大减少数据存储量和分析效率。这个过程自身也是一个知识发现和数据挖掘的过程。

总之,大数据 GIS 的特征之一就是数据存储和管理从传统的面向离线式分析的组织与存储转换为可扩展的、面向实时分析与挖掘的动态处理与管理过程。

3.2　数据驱动的空间分析与挖掘

大数据必然要和数据挖掘相结合,而不仅仅是信息提取,尤其是挖掘隐含的、非显见的模式、规律和知识。传统空间分析的一个重点是空间建模,即计算模拟,考虑的是如何建立一个匹配度或准确度高的模型。已故图灵奖获得者 Jim Gray 提出,科学研究的范式已经从实验科学、理论推演、计算模拟发展到数据密集型科学发现。但是空间数据挖掘不同于传统的数据挖掘,多了一个空间维度以及在不同空间尺度上的挖掘,所以不能完全套用普通事物数据挖掘的策略和方法。早在 1994 年,李德仁院士就提出了“Knowledge discovery from GIS”的理念,建议从纷杂的空间数据中挖掘隐含的模式、规则和知识[8],并针对空间数据挖掘中存在的随机性和模糊性问题系统地提出了云模型、数据场、地学粗空间等挖掘方法[9]。

为了克服大数据的噪音和不确定性,另一种方法是对多源数据进行融合。如公交卡刷卡数据、出租车轨迹数据、自行车租用数据、手机定位数据都是典型的城市大数据,但是单独使用其中任何一种数据都无法全面客观地描述城市交通、人群的移动等信息。智能手机上也具备多种传感器,但是单独使用 GPS 只能进行室外定位,结合 WiFi、陀螺仪、气压计等就可以同时进行室内及高程定位。所以,无论是宏观还是微观层面,如果可能,要尽可能地使用多源数据,并对多源数据进行融合分析和挖掘,以充分发挥大数据的优势。

确定性的地理计算,如 GPS 监测滑坡形变,缓冲区分析仍将继续存在,但是空间分析更多的重点将转移到从积累的定量数据中抽取定性的规则,能够处理不确定性问题和发现蕴含知识及规律的空间数据挖掘算法上,如预测何时、何处、是否会发生滑坡,以及从众多车辆轨迹中识别出道路边界和中心线等。我国正在进行的地理国情监测已经针对我国地表自然、生态以及人类活动的基本情况获得了大量数据,可以通过相关数据对地理国情进行动态测绘、统计,但更应该充分发挥空间数据挖掘的作用,从这些数据中更深层次地发现人与自然和环境的互动关系,为低碳、绿色、可持续发展的社会和生活方式提供重要依据。

数据的极大丰富使人们可以逐渐摆脱对模型和假设的依赖。对于大数据时代,谷歌的研究主管 Peter Norvig 有一句名言:“All models are wrong,and increasingly you can succeed without them”。所以,大数据 GIS 的特征之二就是空间分析方法由模型驱动逐渐转变为数据驱动。大数据 GIS 的空间分析不仅要有建立模型的能力,更要有发现新模式、新知识甚至新规律的能力。

3.3　结合地理计算的可视分析

在经典的“3S”定义中,相对于 GPS 和 RS,GIS 的主要作用在于数据分析和数据可视化。在大数据时代,地球空间信息科学[10]的内涵没有发生改变,但是内容和形式更加丰富。现在可以将 GPS 理解为任何可以标识空间位置的数据,RS 可以理解为多源传感器数据,GIS 则是将这些与空间相关的数据映射到空间基准下统一进行管理、分析和显示。所以,可视化是 GIS 的一个重要功能。

传统 GIS 可视化的重点表现在符号、尺度、三维等问题上。对于大数据,不经过信息提炼和综合直接以点、线、面等符号的形式表现出来,不但达不到传递信息的目的,反而会适得其反,模糊了有效特征;对于相对抽象的数据,如表达真实事

件和虚拟世界交互的社交网络数据，传统GIS则无能为力。所以可视化不仅仅是图形显示功能，还是数据分析和挖掘的一个重要手段，即可视分析[11-12]。

可视分析通过交互可视界面对数据和事件进行分析、推理和决策，对复杂情景进行更深层的理解；同时提供快速、可检验、易理解的评估和更有效的交流手段。在大数据时代，数据量和复杂度的提高带来了对数据探索、分析、理解和呈现的巨大挑战，这决定了数据可视化技术将成为大数据时代的显学[13]。除了上述的统计或者数据挖掘的方法，大数据的海量、动态和不确定整合信息还需要用交互式或动态化展示的表现方式来帮助人们迅速理解和分析数据的模式、特征和内涵，辅助数据的提炼和解释，以及从复杂数据中探索新的发现。可视分析可以放大人类的知觉推理和认知能力，所以发展与地理计算相结合的可视分析是大数据GIS的另一个重要特征和发展方向。

4 结 语

总之，大数据时代，GIS面临着以下挑战和机遇：① 超大规模数据的高效管理，其中包括数据管理体系和架构、流数据的实时处理和分析以及历史数据和模式的高效查询和分析。② 针对大量数据噪音多及数据不确定性大的特性，需要重新思考空间统计模型的选择、参数的训练和使用及计算效率等问题。③ 面对全体数据，需要发展适合的空间数据挖掘算法，发现数据背后所隐藏的模式和价值。④ 高效地显示和分析超大规模的时空数据，发展与地理计算相结合的可视分析理论。大数据GIS需要一整套系统、科学的理论和方法来应对大数据带来的挑战。

致谢：感谢乐阳博士为完成本文所做的工作。

参 考 文 献

[1] Ahalt S C. Why Data Science[J]. *Communications of the CCF*, 2013, 9(12):11-15(Ahalt S C. 为什么需要数据科学[J]. 中国计算机学会通讯, 2013, 9(12):11-15)

[2] History of Bigdata 2013:A Scan of the Industry Data Volume in Year 2013[OL]. http://www.36dsj.com/archives/6285, 2013(大数据史记2013：盘点中国2013行业数据量[OL]. http://www.36dsj.com/archives/6285, 2013)

[3] Zikopoupos P C, Eaton C, de Roos D, et al. Understanding Big Data, Analytics for Enterprise Class Hadoop and Streaming Data[OL]. http://public.dhe.ibm.com/common/ssi/ecm/ en/iml14296usen/IML14296USEN.PDF, 2012

[4] Karel R. See Big Data Through a Different Lens[OL]. https://www.informatica.com/potential-at-work/information-leaders/article/see-big-data.shtml, 2013

[5] Li Deren, Wang Shuliang, Li Deyi. Theories and Application of Spatial Data Mining[M]. 2nd ed. Beijing: Science Press, 2013 (李德仁，王树良，李德毅. 空间数据挖掘理论与应用[M]. 2版. 北京：科学出版社, 2013)

[6] Li Qingquan, Yang Bisheng, Shi Wenzhong, et al. Three Dimensional Spatial Data Real-time Acquisition, Modeling and Visualization[M]. Wuhan: Wuhan University Press, 2003 (李清泉，杨必胜，史文中，等. 三维空间数据的实时获取、建模与可视化[M]. 武汉：武汉大学出版社, 2003)

[7] Li Q Q, Zhang T, Yu Y. Using Cloud Computing to Process Intensive Floating Car Data for Urban Traffic Surveillance[J]. *International Journal of Geographical Information Science*, 2011, 25(8): 1 301-1 322

[8] Li D R, Cheng T. KDG—Knowledge Discovery from GIS[C]. The Canadian Conference on GIS, Ottawa, Canada, 1994

[9] Li Deren, Wang Shuliang, Li Deyi. Theories and Application of Spatial Data Mining[M]. Beijing: Science Press, 2006 (李德仁，王树良，李德毅. 空间数据挖掘理论与应用[M]. 北京：科学出版社, 2006)

[10] Li Deren, Li Qingquan. The Formation of Geospatial Information Science[J]. *Advances in Earth Science*, 1998, 13(4):319-326(李德仁，李清泉. 论地球空间信息科学的形成[J]. 地球科学进展, 1998, 13(4):319-326)

[11] Wong P C, Thomas J. Visual Analytics[J]. *IEEE Computer Graphics and Applications*, 2004, 24(5): 20-21

[12] Kovalerchuk B, Schwing J. Visual and Spatial Analysis: Advances in Data Mining, Reasoning, and Problem Solving[M]. Netherlands: Springer, 2004

[13] The CCF Task Force on Big Data. The Forecast of 2014 Trends of Big Data[J]. *Communications of the CCF*, 2014, 10(1):32-36(CCF大数据专家委员会. 2014年大数据发展趋势预测[J]. 中国计算机学会通讯, 2014, 10(1):32-36)

(下转第666页)

Towards Big Data-Driven Human Mobility Patterns and Models

LIU Yu[1] *KANG Chaogui*[1] *WANG Fahui*[2]

1 Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing 100871, China

2 Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA 70803, USA

Abstract: In the big data era, massive volumes of individual-level movements, extracted from various geospatial data, including mobile phone data, public transportation card records, social media check-in data, taxi trajectories, and bank card records, are available for scholars in different fields to study human mobility patterns. These studies enrich spatio-temporal analysis methods in GIS and provide a new perspective to human-environment interactions. Observed human mobility patterns and models can be applied to many applications such as transportation and public health. This paper presents a generic workflow for big-data-driven human mobility analyses and summaries major movement measures. By comparing a number of models used to interpret and reproduce the observed pattern, this paper emphasizes the geographical impact on human mobility patterns.

Key words: big data; human mobility pattern; movement model; geographical impact

First author: LIU Yu, professor. His research interest is geographical information science. E-mail: liuyu@urban.pku.edu.cn

Foundation support: The National Natural Science Foundation of China, Nos. 41271386, 41271385, 41171296.

(**上接第** 644 **页**)

Big Data GIS

LI Qingquan[1,2] *LI Deren*[2]

1 Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, China

2 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: Big data is changing the world, and also posing challenges for GIS. The volume, velocity, and variety of these data challenge the data management ability of GIS, while the veracity and value issues of big data challenge spatial analysis theory and methods. Thus, as a tool focusing on spatial data management, analysis and visualization, GIS has to make necessary adjustments and changes to meet the big data requirements. This paper discusses the challenges based on the 5V properties of big data, and then, analyzes three characteristics of future GIS in the big data era, which are: ① scalable data management, ②data-driven modeling and data mining, and ③ geo-computational visual analytic.

Key words: big data; GIS; spatial data management; spatial data analysis; visualization

First author: LI Qingquan, PhD, professor, specializes in spatial-temporal data mining, multi-sensor integration, and industry and engineering surveying. E-mail: liqq@szu.edu.cn

Foundation support: The National Natural Science Foundation of China, No. 41371377; Shenzhen Scientific Research and Development Funding Program, Nos. ZDSY20121019111146499, JSGG20121026111056204; Shenzhen Dedicated Funding of Strategic Emerging Industry Development Program, No. JCYJ20121019111128765.