DOI:10.14188/j.2095-6045.2016.06.001

文章编号:2095-6045(2016)06-0001-04

时空大数据处理的需求、应用与挑战

边馥苓[1]　杜江毅[1,2]　孟小亮[1]

1 武汉大学国际软件学院,湖北 武汉,430079

2 湖北工业大学计算机学院,湖北 武汉,430068

Requirements, Applications and Challenges of Spatio-Temporal Big Data Processing

BIAN Fuling[1]　*DU Jiangyi*[1,2]　*MENG Xiaoliang*[1]

1 International School of Software, Wuhan University, Wuhan 430079, China

2 Computer School of Hubei University of Technology, Wuhan 430068, China

摘　要:时空信息不仅是大数据的重要组成部分,更可被看成是大数据本身。从大数据时代时空信息处理面临的新需求出发,包括时空信息处理需要大数据平台、大数据库、新的快速计算模型和分析方法,以及时空信息安全和标准化等,阐述了时空信息和大数据相辅相成的紧密联系,并对当今时空大数据的应用和挑战举例分析,描述了现今时空大数据的典型应用及其处理技术存在的问题,并对如何解决和将来的发展方向提出了新的思路。

关键词:时空信息处理;大数据;时空大数据

中图法分类号:P208

文献标志码:A

Abstract: Spatio-temporal information is not only an important part of the Big Data, but also can be seen as the Big Data itself. From the new requirements of spatio-temporal information processing in the era of the Big Data, including the Big Data platform, database, new fast calculation models and analysis methods, the security and standardization of spatio-temporal information. Firstly, this paper expounds the close relationship between the spatio-temporal information and the Big Data, analyzes today's spatio-temporal Big Data applications and challenges, then describes the problems of existing typical application and processing technology. Final-

基金项目:国家自然科学基金资助项目(41501441);湖北省环境保护厅环保科研资助项目(2014HB04)。

Foundation support: The National Natural Science Foundation of China(41501441); The Environmental Protection Project of Hubei Province Environmental Protection Bureau(2014HB04).

ly, the new ideas for how to solve the problems and the future direction are proposed.

Key words: spatio-temporal information processing; big data; spatio-temporal big data

当前,随着通信技术和信息技术的高速发展,互联网上的数据正在激增,人类社会的数据种类和规模正以前所未有的速度增长[1],大数据(big data)时代已经来临。麦肯锡在《大数据:下一个竞争、创新和生产力的前沿领域》的研究报告中认为,医疗保健、零售业、公共领域、制造业和个人位置的数据构成了目前 5 种主要的大数据流,上述无论哪种数据流都具有显著的地理编码与时间标签。从这个角度看,时空信息不仅是大数据的重要组成部分,更可被看成是大数据本身。

目前,很多科学研究都基于大数据进行展开,为更深入地理解大数据时代的时空信息技术,本文从 3 个方面进行了阐述。

1 时空信息和大数据相辅相成

时空信息和大数据关系密切,相辅相成。首先,何为大数据? 科学界目前并没有给出一个完整统一的概念,但归结下来,大数据不仅仅是海量数据,更是一种理念与思维,是在当前信息技术的支撑下,完成从大量、快速、复杂、多源的信息中提取有用价值的技术处理手段,其中,最为核心的就是需要挖掘出数据中蕴含的潜在价值。而时空信息是具有时间要素特征的空间信息,空间信息即是有意义的空间数据[2,3],其 3 要素可以表现为地点、时间和对象。它代表着现实世界地理实体或现象在信息世界中的映射,反映自然界向人类社会传递的信息。海量时空信息离不开大数据,大数据成为时空信息发展的新动力。而以大数据处理为中心的数据科学也需要时空信息的相关理论作为支撑,时空信息处理技术将成为数据科学领域的重要手段和工具,两者相辅相成,互相需要。以个人位置信息研究为例,截至 2009 年,个人位置信息的数据就已经达到了 PB 级,而人类生活中所产生的数据有 80%和空间位置有关[4],大量的隐含时空隐喻的个人位置信息蕴含了巨大的可以利用的价值。面对越来越庞大的时空信息数据,传统的处理方法已经很难跟上时代发展的速度,只有运用大数据的分析与挖掘方法才能够更加快速、实时地提炼出这些信息的价值。

从个人位置信息的实例中可以很清楚地看到,时空信息和大数据之间有着非常密切的联系,主要体现在以下 3 个方面。

1)时空信息的价值空前提升。随着大数据的逐步发展以及数据分析与挖掘的不断进步,时空信息的价值被重新认识。针对海量的地理信息数据研究的相关技术成为热点,地理空间的思维方式成为科学的世界观和方法论,地理信息服务的价值也正逐步升级到认识世界、改造世界的科学理论和科学工具层面。

2)大数据引发了信息采集的全面变革。各种简便、便携式的测量工具层出不穷。而且,随着互联网技术的不断进步,通过网络更快捷地获取各种时空信息,地理信息数据采集开始与其运营服务分离,地理信息生产服务提供者正从专业走向大众。

3)时空信息服务呈现普适化趋势。时空信息服务正全面走进人们的生活,如地图内容实现实时在线服务,基于地理围栏技术(基于地理空间位置围出一个虚拟的地理区域)的精准信息推送服务,以及电子商务、线上线下服务等。

大数据时代的来临促进了时空信息来源的复杂化、多样化、多源化以及巨量化,只有运用大数据的解决手段才能对大量的时空信息进行有效的利用与挖掘,合理利用好时空大数据,将会给人们的生活带来极大的价值和便利。

2 大数据时代的时空信息处理面临新的需求

时空数据是一种结构复杂、多层嵌套的具有空间和时态特性的高维数据,它有效记录了事物的空间位置和时空变化过程,并准确地表达了事物的历史、当前和未来状态,如城市变迁、疾病扩散、环境变化、地质演化、移动对象位置变更等等。传统的时空信息处理平台、方法和技术等已经无法满足大数据时代时空信息处理的需求。这主要体现在时空信息处理需要大数据平台、大数据库、大数据分析方法和快速计算方法等方面。

1)时空大数据平台。在众多领域,包括金融、通讯、生物学、物理学、军事及地球科学等领域,数据量迅速膨胀变大,对于这些海量、多样数据的采集、存储以及应用已经成为这些领域的一个持久的业务趋势,也是地球科学发展到时空大数据阶段的一个必须跨越的门槛。随着全球对地观测系统(earth observing system,EOS)、智慧城市、物联网等行业的深入发展,该趋势和门槛也不断地变化,超出传统平台中基础设施和架构的承载能力,且不能满足于现有计算能力的实时性要求。这就迫切需要一种新的大数据平台,它整合了非结构化和结构化的地理空间数据及时间相关数据,能提供用于大数据处理分析的经济且高效的基础设施,通过虚拟化方法整合计算、存储、网络资源,从而提高资源利用率,增强资

源的弹性伸缩能力,并可通过与大型的存储共享云平台连接,以减少数据存储、处理中心所需的存储空间。

2)时空大数据库。传统用于结构化设计的空间数据处理工具和关系型数据库在管理非结构化时空数据时暴露出了很多的局限性,难以满足时空数据处理的需求。这必然导致对时空大数据库的迫切需求。现有的时空数据主要来源于 GPS、遥感和传感器等设备,以及通过人机交互合作获取的各种类型的数据,每种方式生成的数据格式和数据形式各不相同;数据总量较大甚至巨大,数据类型多样,且非结构化数据所占的份额越来越大。数据产生速度飞快,主要基于手持移动终端、互联网、物联网、车联网等平台产生。因此,整合、清洗、存储和管理不同来源且结构复杂的时空大数据是时空大数据库面临的重要问题。

3)时空大数据分析方法。数据即价值,如何充分利用时空数据中蕴藏的价值,设计出适合海量时空数据的分析方法也是当前时空数据研究的热点问题。时空数据量大,可以通过对计算机系统的扩展加以应对,如采用基于集群的分布式存储与并行计算体系结构和硬件平台,可以在一定程度上缓解数据量大带来的挑战。但是时空数据类型多样,特别是其中非结构化数据的快速处理问题是大数据分析的重大挑战,很难依靠某一种或少数几种方法就能解决所有的大数据分析问题,需要针对不同类型数据的特点研究相应的分析技术。

4)时空大数据的快速计算模型。随着海量时空数据的持续性获取,科学领域获得了多种时空大数据,例如,通过卫星传感设备获取的遥感影像、大规模天气预报数据、天文光谱或测光观测数据等,这些大数据也为科学研究带来了重大机遇[5,6]。数据正以 Moore 定律的方式增长,但显而易见的是,大数据集的出现并没有让我们的发现和知识呈现爆炸式的增长,这就需要人们在研究方法和技术上有所改变乃至突破,才可能在这股数据洪流中冲浪而不是被它所淹没。为了不被科学大数据淹没,对于现有的科学计算模型提出了更高的计算效率要求和挑战,一个典型的需求是研发时空大数据快速计算模型,实现海量数据的快速计算。

5)时空大数据所面临的安全和标准化需求。时空大数据在带给人们巨大收益的同时,也对个人信息的安全产生了威胁。如何更好地利用时空大数据,同时更好地提高数据安全性,已经成为迫切的需求。且时空数据结构复杂,来源多样,与传统数据管理迥然不同的特点使得大数据标准化已成为各国促进大数据产业发展的重要举措。

3　时空大数据处理的应用与挑战

随着信息行业以及互联网的发展,时空大数据迅速发展,这给人们带来了新的机遇和挑战。时空大数据多元异构、高噪,数量巨大。它的产生并非是为了分析而收集,而是信息社会自动化产生,有着开放易得的特点;通过兴起的社交网络,如微博等获得的数据同时也具有很强的交互性;每时每刻都有大量新的网络数据发布,网络信息内容不断变化,导致了信息传播的时序相关性,也具有很高的噪声。结合时空大数据自身的特性,在对其分析方法上,有着如下特点:①重发现不重实证,即在没有理论假设的前提下去预知社会和洞察趋势、规律;②重关系不重因果,问什么而不问为什么;③重预测,相关应用都是预知社会问题,它用逻辑和计算取代了依赖传统和直觉的生产方式;④突发性,短时间内可以引起大量新的网络数据与信息的产生;⑤反映社会,网络数据成了对社会状态的直接反映。

时空大数据自身和其分析的特点使得它能够通过充分搜集数据的时空关系,结合环境等其他因素建模,从而对各行各业中产生的海量数据进行分析,在很多方面都有成功的应用。例如,在智慧交通中解决人们道路出行问题,分析挖掘车辆的运行状况以及人流的移动规律,可以实现对交通状况的跟踪和实时预报。下面再举一些时空大数据处理新兴应用的例子。

1)在经济方面,可以应用大数据分析报告实施产业信贷风险控制,也可以依据客户消费习惯、地理位置、消费时间等要素实现精准营销,并更精确地投放广告,利用大数据分析技术加快内部数据处理速度,利用全局数据了解业务运营薄弱点等。

2)在社会管理方面,以大数据为代表的数据资源来源广泛,数据粒度小,记录单元碎片化,结构多元化,从根本上改变了人文知识的获取、标注、比较、取样、阐释与表现方式。通过对地理、气象、交通运输等自然信息和经济、社会、文化、人口等人文社会信息的挖掘,可以为城市规划提供强大的决策支持,强化城市管理服务的科学性和前瞻性。

3)在医疗健康方面,主要是疾病预测与医疗,通过对相关数据进行分析和建模,除了医疗检测和医师诊断,还充分考虑到大数据分析结果,对海量数据进行分析,并获得深刻的洞见。其典型应用有人群疾病的时空分布规律的分析和了解、疾病成因分析以及疾病的预防和控制等 3 个方面[7]。

4)在生态环境方面,我国的生态环境数据中包

含了大量的时间和空间信息,通过分析可以对将要发生的灾害与污染进行有效的监测、评估与预防。通过加大在指定规划、决策方面的政府支持力度,以对海量时空数据的分析处理为基础,建立生态环境信息管理系统,加强和提升对灾害与污染的防治水平和能力。

5)在竞技体育方面,大数据获取技术和挖掘技术的进步为时空大数据在体育科学研究中的运用提供了技术基础。运用可穿戴设备记录下运动员在赛场上的各种信息,获取充足的比赛中各项信息的数据。这些数据是一种不受无关变量干扰的多维度数据融合,将成为研究和提升球员在球场上表现的重要数据。

综上所述,时空大数据拥有巨大的价值,但也存在挑战。在大数据采集技术方面,集中式网络爬虫采集信息的速度和规模已经难以满足实际应用的需要;在数据预处理方面,对海量数据的清洗和变换会极大地加重应用的负担和开销,而且大数据的预处理可能会造成有效数据的丢失,进而影响数据的完整性和价值;在大数据存储及管理技术方面,到目前为止,世界上并没有一个比较统一的存储接口标准,而且国内并没有出现相关的比较成熟的服务;在大数据分析技术方面,数据日趋庞大,性能易陷入瓶颈,用户对实时性和响应时间的要求越来越高,传统方法无法处理大数据等;在大数据硬件平台方面,传统的方法无法解决对大数据的备份及还原问题,大多数的大数据平台架构基于 Hadoop,没有形成支持大数据平台的完整体系。时空大数据在成功应用于生活的各个方面的同时也暴露出了一些问题。一是对数据的安全和隐私的保护问题。由于时空大数据包含着用户的隐私和其他敏感信息,不恰当的使用会给用户带来严重的威胁。如何实现对时空大数据隐私全面的保护,采用适当的度量方法度量用户敏感信息的泄露程度,并在此基础上兼顾隐私保护的程度和基于位置服务的可用性[8],已成为当前时空大数据隐私保护面对的挑战。二是由于数据来源多样,时空大数据的格式不尽相同,这给时空大数据应用带来了一些阻碍,如何整合、清洗和转换不同来源的时空数据还需要进一步的研究和规范。时空大数据的标准化问题如果不加以解决,将会在很大程度上阻碍时空大数据共享和应用的发展。

4 结束语

随着大数据时代的来临,时空信息的价值将会更加巨大,时空信息和大数据的结合也将更加紧密。面对这一历史机遇,通过研究时空大数据处理的特点,分析其优势与不足,充分挖掘其潜力,必将推动时空大数据在社会生活与科学研究中发挥更加巨大的作用。

参考文献

[1] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展,2013,50(1): 146-169
Meng Xiaofeng, Ci Xiang. Big Data Management: Concepts, Techniques and Challenges [J]. Journal of Computer Research and Development, 2013, 50(1): 146-169

[2] 边馥苓. 空间信息导论[M]. 北京:测绘出版社,2008
Bian Fuling. Introduction to Geo-Spatial Information [M]. Beijing: Surveying and Mapping Press, 2008

[3] 边馥苓. 论我国地理信息产业、人才现状与存在问题[J]. 地理信息世界,2009,7(5): 29-34
Bian Fuling. A Study on China's Geomatics Industry, Its Present Situation of Talents and Problems[J]. Geomatics World, 2009,7(5): 29-34

[4] Xu Guanhua. Pay Much Attention to the Digital Erarth by the Social[J]. Science News Weekly,1999(1):7-8

[5] Jacobs A. The Pathologies of Big Data[J]. Communications of the ACM,2009,52(8):36-44

[6] Crampton J W, Graham M, Poorthuis A, et al. Beyond the Geotag: Situating 'Big Data' and Leveraging the Potential of the Geoweb [J]. Cartography and Geographic Information Science,2013,40(2):130-139

[7] 武继磊,王劲峰,郑晓瑛,等. 空间数据分析技术在公共卫生领域的应用[J]. 地理科学进展,2003,22(3):219-228
Wu Jilei, Wang Jinfeng, Zheng Xiaoying, et al. A Review on Application of Spatial Data Analysis Technology in Public Health[J]. Progress in Geography, 2003, 22(3):219-228

[8] 王璐,孟小峰. 位置大数据隐私保护研究综述[J]. 软件学报,2014,25(4):693-712
Wang Lu, Meng Xiaofeng. Location Privacy Preservation in Big Data Era: A Survey[J]. Journal of Software, 2014, 25(4): 693-712

收稿日期:2016-03-25

第一作者简介:边馥苓,教授,博士生导师,国务院政府津贴享受者,湖北省有特殊贡献的中青年专家,武汉市九届人大代表,湖北省十届人大代表,华中农业大学、桂林工学院兼职教授,主要研究方向是空间信息与数字技术。

First author: BIAN Fuling, professor, PhD supervisor. Her research interest is spatial information and digital technology.

E-mail: flbian@whu.edu.cn

通讯作者:杜江毅,博士生, 主要研究方向是空间数据挖掘。

Corresponding author: DU Jiangyi, PhD candidate, majors in spatial data mining.

E-mail: jydu1980@mail.hbut.edu.cn