测绘地理信息科技出版资金资助
CEHUI DILI XINXI KEJI CHUBAN ZIJIN ZIZHU

时空大数据的技术与方法

Techniques of Spatio-Temporal Big Data

边馥苓　主　编
孟小亮　崔晓晖　副主编

测绘出版社

测绘出版资助

时空大数据的技术与方法

Techniques of Spatio-Temporal Big Data

边馥苓 主 编
孟小亮 崔晓晖 副主编

测绘出版社
·北京·

内容简介

本书在对时空信息和大数据相关概念认识的基础上,描述大数据应用于时空信息处理方面所需的软硬件平台,分析时空大数据库与传统时空数据库和大数据存储的区别,探讨时空大数据分析对测绘学科的贡献,提出时空大数据快速计算的方法,并探索时空大数据及其处理技术存在问题的解决方法和未来的发展方向。本书沿着时空大数据处理技术主线,从时空信息与大数据两个方面进行结构组织,内容紧跟学术研究前沿,具有一定的前瞻性。

本书不仅可以作为空间信息与数字工程、计算机、测绘和物联网等领域科技工作者的参考书,还可以作为研究生课程的专业教材。

图书在版编目(CIP)数据

时空大数据的技术与方法/边馥苓主编. —北京:测绘出版社,2016.5
ISBN 978-7-5030-3937-9

Ⅰ. ①时…　Ⅱ. ①边…　Ⅲ. ①空间信息技术—研究　Ⅳ. ①P208

中国版本图书馆 CIP 数据核字(2016)第 100000 号

责任编辑	雷秀丽	**封面设计**	李伟	**责任校对**	董玉珍	**责任印制**	陈超

出版发行	测绘出版社	**电　　话**	010—83543956(发行部)
地　　址	北京市西城区三里河路 50 号		010—68531609(门市部)
邮政编码	100045		010—68531363(编辑部)
电子邮箱	smp@sinomaps.com	**网　　址**	www.chinasmp.com
印　　刷	北京京华虎彩印刷有限公司	**经　　销**	新华书店
成品规格	169mm×239mm		
印　　张	10.5	**字　　数**	203 千字
版　　次	2016 年 5 月第 1 版	**印　　次**	2016 年 5 月第 1 次印刷
印　　数	0001—1000	**定　　价**	43.00 元

书　　号　ISBN 978-7-5030-3937-9

本书如有印装质量问题,请与我社门市部联系调换。

主　编：边馥苓

副主编：孟小亮　　崔晓晖

编　委（按姓氏拼音为序）：

韩　波　　李晓剑　　谭喜成

田扬戈　　王黎维　　王志波

前　言

统计研究表明，85%以上的大数据都与时空信息相关。随着集成电路与芯片、传感器网络、移动定位、无线通信、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、人人都是传感器的梦想正在逐步付诸实践。到2020年，全球产生的数据量的总和将达到40 ZB。而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，正是因为有了卫星定位、Wi-Fi、移动通信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，并成为互联网企业的必争之地。泛在、互动、非专业、实时、按需服务的新地理信息时代，无疑也将是时空大数据大放异彩的时代。在新地理信息时代多学科交叉中的大测绘科学快速发展的趋势下，本书作者及其所在的科研团队一直在时空大数据相关教学与科研实践中摸索前进。

哲学、物理、天文、数学、地学分别从各自学科的角度对时空现象进行了研究，计算机科学、人工智能等领域的专家学者更借助计算机技术，根据不同应用需求，对地理实体的空间、时间和属性维度同时进行研究和处理，但由于受计算机软硬件环境和相关技术方面的限制，存在着诸多困难。因此，人们的研究思路主要集中在只考虑空间、时间和属性特征中任两个变量的情况，近年来，随着软硬件环境的改善和大数据技术被广泛研究，关于时空信息处理的研究也逐渐增多起来。本书主要分析时空信息与大数据的关系，讨论时空信息处理面临的新研究需求，以及时空大数据平台架构、时空大数据库、时空大数据分析方法、时空大数据的快速计算模型等关键解决技术，总结时空大数据在社交网络、交通运输、公共卫生与医疗、地质灾害监测与预防、竞技体育等方面的成功应用，具有较高的学术价值和应用价值，对空间信息与数字技术专业及大测绘学科的发展都能起到积极的推动作用。

近年来，国内外大数据领域的书籍层出不穷，大多数是对大数据的产生和发展、基本原理与主流技术，以及应用领域做相关阐述。国内外相关研究多是基于传统测绘学科的研究并发展创新，其中空间信息相关理论和应用书籍也比较多，但目前市面上没有一本涵盖时空信息与大数据结合内容的应用技术专著，本书的目的之一是填补这方面的空白。本书在对时空信息和大数据相关概念的认识基础上，描述大数据应用于时空信息处理方面所需的软硬件平台，分析时空大数据库与传统时空数据库和大数据存储的区别，探讨时空大数据分析对大测绘学科的贡献，提出时空大数据快速计算方法，并探索时空大数据及其处理技术现今存在问题的解

决方法和将来的发展方向。本书沿着时空大数据处理技术主线，从时空信息与大数据两个方向进行结构组织，内容新颖独特，紧跟学术研究前沿，具有一定的前瞻性。

本书由边馥苓主编，负责选题、结构和内容的确定。由下列人员负责各章节的执笔编写：第1章孟小亮，第2章李晓剑，第3章谭喜成，第4章王黎维，第5章田扬戈，第6章韩波，第7章王志波。最后由边馥苓、孟小亮、崔晓晖负责审校和定稿。

由于时间紧促，本书内容在深度和广度上可能存在不足，恳请广大读者多提宝贵意见。

边馥苓

2016年5月

目　录

Contents

第1章　从空间信息到时空信息

1.1　时空信息内涵

1.1.1　空间认知——空间数据、空间信息与空间知识

人们对地理空间的认知主要来源于对地理实体、地理事件、空间对象、空间关系等空间数据、空间信息与空间知识的认知。

空间数据一般是通过地面测绘技术、摄影测量、土地调查、全球定位系统(Global Positioning System,GPS)、遥感(remote sensing,RS)、数字化、传感器及无线传感器网络、空间数据复合等采集手段与技术(边馥苓,2011a),将关于现实世界的模拟信号转换成计算机能够识别、加工处理与分析的具有某种特定数据组织结构的数据。空间数据主要指以地球表面空间位置为参考的自然、社会、人文和经济数据等,用于描述呈现二维、三维,甚至多维分布的实体或现象的空间位置、形态、属性和空间关系。常用的空间数据包括矢量数据、遥感影像、数字高程模型、三维空间目标,可量测的立体影像数据等多源空间数据。

空间数据所表达的信息称为空间信息,反映了空间实体的位置以及与该实体相关联的各种附加属性的性质、关系、变化趋势和传播特性等的总和(边馥苓,2008,2009),其三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。空间性是空间信息最主要的特征,是区别于其他信息的显著标志。空间性表示了空间实体或现象的地理位置、几何特性和拓扑关系,是空间信息处理分析的基础。包含了空间实体的位置和属性信息的空间数据是语义的,但是,通常在应用中人们往往将表示实体位置信息的数据称为空间数据,将表示实体性质、特征等的属性数据独立出来,单独作为属性数据保存,从这个角度讲,空间数据是非语义的,因为单独的坐标(位置)数据可以是任何东西。

实体元数据是对空间信息的一种描述手段,本体元数据为空间信息添加了更多的语义特性,使其可以通过逻辑、规则进行空间知识的推理(Jakus et al.,2013)。空间知识是一个或多个空间信息关联在一起形成的结论性、概括性、描述性、有应用价值的信息结构(邸凯昌,2001)。华裔地理学家 Sui 和美国地理学家 Goodchild 等分析了地理空间知识生产与地方推理和集成化方向的跨学科发展战略,并突出公民对地球科学的影响,认为联系二者的核心是地理信息科学(geographic

information system,GIS),它提供了强大地理信息获取、分享和分析技术(Sui et al.,2013)。麦肯锡报告提出人类目前已步入大数据时代,绝大部分数据被贴上地理标签或者本身就是地理数据(Manyika et al.,2011),附着在数据上的信息与知识也伴随着拥有了地理标签,形成了空间信息与空间知识。

通常,从原始采集的空间数据中可以提炼出其信息价值,一般采用的方法是将数据结构化,总结其要素属性。通过收集信息,在其要素属性的基础上,加入逻辑、规则进行知识的推理、判断和预测。例如,从年复一年日复一日的空气污染物浓度实时监测数据中,可以知道当地的空气质量的信息,通过专家知识或经验规则等还可以推断空气质量对居民健康状况的影响并及时做出预警。如图 1.1 所示,从所占存储空间大小的角度看,空间数据、空间信息与空间知识呈金字塔架构,大量的数据是基石,无处不在,而塔尖的知识并不显而易见。

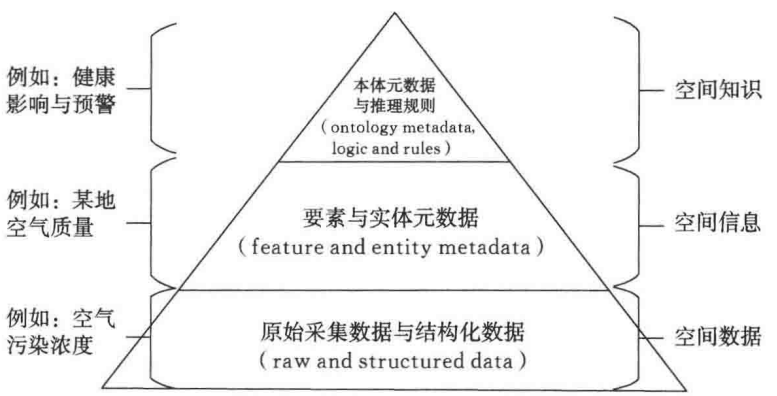

图 1.1 空间数据、空间信息与空间知识的金字塔架构

目前,空间数据、空间信息与空间知识的概念已超越传统的地理数据、地理信息与地理知识的范畴,将空间观从传统地理学的地球表面扩展到任何空间,既包括地图制图、遥感应用、生态环保等领域研究的地球表面空间,也包括医疗影像处理研究的人体内部空间、网络路由研究的计算机网络空间、宇宙空间探索研究的宇宙空间等。尤其是那些全局性、战略性的重大问题,其信息化的大部分内容或者直接与空间数据、空间信息、空间知识相关联,或者间接利用其解决问题。

从数据、信息与知识在人类社会经济生活的认知度来看,空间数据可以由专业数据采集人员或非专业人员采集,空间知识被认为是来自专家或者专家系统,而空间信息是目前信息化浪潮中各行业普遍接受、推广与应用的对象,本书主要讨论的是空间信息,特别是时空信息的应用。

1.1.2 时空认知

本书中的时空信息是指具有时间要素特征的空间信息。

空间和时间也是人类文明中最古老的概念。远古时期原始的耕作、放牧需要丈量大地、顺应天时，产生了简单的空间和时间的概念及其度量方法。

在中国古代，早就有“上下四方曰宇，往古来今曰宙”之说，这里的“宇”和“宙”就是空间和时间的概念，也是原始的三维空间和一维时间的概念，并将宇宙密切联系起来。因此，在汉语中，宇代表了所有的空间，宙代表了所有的时间，所以“宇宙”这个词有“所有的时间和空间”的意思。

而“时空”一词出现于哲学，发展于物理学与数学，是时间与空间的简略集合名词。但对于时间和空间的认识，还存在着一些基本问题有待解决，还在不断地发展（中国大百科全书总编辑委员会，2009）。时、空都是绝对概念，是存在的基本属性。但其测量数值却是相对于参照系而言的。任何事物都处于一定的时空之中。

哲学上，空间和时间的依存关系表达着事物的演化秩序。涉及的发散性概念有周易里的“乾坤”，道家的“道”以及孔孟之道的大成智慧。“时间”是抽象概念，表达事物的生灭排列。其内涵是无尽永前，其外延是一切事件过程长短和发生顺序的度量。“无尽”指时间没有起始和终结，“永前”指时间的增量总是正数。“空间”是抽象概念，表达事物的生灭范围。其内涵是无界永在，其外延是一切物件占位大小和相对位置的度量。“无界”指空间里任一点都居中，“永在”指空间永现于当前时刻。

近代科学的发端，必然涉及空间和时间的概念及其测量方法。近几个世纪以来，物理学和天文学对空间和时间的认识大体上可分为相互交织的两条线索：①从以牛顿力学和麦克斯韦电磁理论为代表的空间-时间概念，经过狭义相对论和广义相对论，发展到现代宇宙论，这是一条线索。②从经典力学经过量子论、量子力学和量子场论，到追求量子引力、超弦和 M 理论，这是另外一条线索。

因为在狭义相对论中，光速是测量时、空的共同尺子，时、空的变化在此共尺上表现依存规律，即遵从洛伦兹变换。所以，时、空的测量数值是相对于具体惯性系的，如同时性在测量上不是绝对的，相对于某一参照系为同时发生的两个事件，相对于另一参照系可能并不同时发生；长度和时段在测量上也不是绝对的，运动的尺相对于静止的尺变短，运动的钟相对于静止的钟变慢。光速在狭义相对论中是绝对量，对于任何惯性参照系光速都是常量 c。

在数学上有各种多维空间，但目前为止，我们认识的物理世界只是四维，即三维空间加一维时间。现代微观物理学提及的高维空间是另一层意思，只有数学意义，我们还无法感知。

地学中的时间地理学是由瑞典地理学家黑格斯特兰德（Hägerstrand）和其带领的隆德学派于 20 世纪 60 年代提出的。20 世纪中后期，西方人文地理学有向社会学渗透的趋势，作为高福利国家的瑞典把社会目标转变为每个社会公民生活质量的提高。黑格斯特兰德把人口统计学中的生命线（life line）概念加上空间轴后，提出生命路径（life path），成为后来时间地理学方法论的基础。时间地理学注重

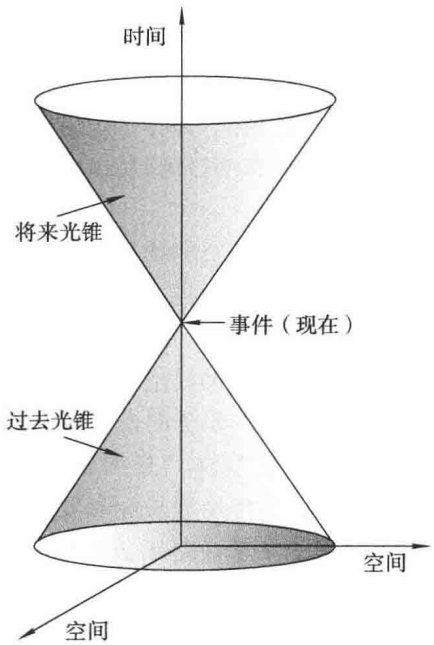

图 1.2 四维时空

分析围绕人类活动的各种具体制约条件，并在时空轴上动态地描述和解释各种人类活动。

时间地理学将时间和空间在微观层面上结合起来，从微观个体的角度去认识人的行动及其过程的先后继承性，去把握不同个体行为活动在不间断的时空间中的同一性。在这里，时间和空间更多的是一种资源的概念，这种资源不仅有限，而且不可转移。时间地理学将传统的空间资源配置和空间秩序动态扩展至时空间资源配置和时空间秩序动态，特别强调了时间秩序的动态。

时间地理学方法不仅是一种动态的方法，而且是一种基于个人行为研究的微观手段，它将微观化个人的研究成果转移到具有某种共同属性的微观人群的研究上，尤其注重个人日常行为的分析。时间地理学研究中经常采用的方法是通过跟踪一个群体中每个人的日常活动路径，研究发生在路径上的活动顺序及时空特征，以得出个人或群体活动行为系统与个人或群体属性之间的匹配关系，从而找到不同类型人群的活动规律，并且利用这种规律进行合理的设施配置。

除哲学、物理、天文、数学、地学领域以外，计算机科学、人工智能等领域的专家学者还借助计算机技术，根据应用部门的不同需求，分别从各自学科的角度对时空现象进行了研究，但由于受计算机软硬件环境和相关技术方面的限制，对地理实体的空间维度、时间维度和属性维度同时进行研究和处理，存在着诸多困难。因此，研究人员的研究思路主要集中在只考虑空间、时间和属性特征中任意两个变量的情况（或者将三个变量压缩转换为两个变量）。近几年来，随着软硬件环境的改善和大数据处理技术的广泛研究，关于时空信息处理的研究也逐渐增多。接下来，主要谈时空信息处理及其目前所面临的研究需求。

1.2 时空信息处理

1.2.1 研究历程

20 世纪 60—70 年代，出现了磁盘、磁鼓等直接存取的存储设备，使得计算机

在实现科学计算的同时，可以完成信息的管理功能。但数据主要以文件形式组织管理，对时间维度的研究更多是在传统学科，尤其是地理学等研究的基础上考虑时间因素。例如，1964 年 Berry 在栅格格式下使用了三维地理矩阵（geographic matrix），以位置、属性和时间分别作为矩阵的行、列和高；1970 年，Hagerstrand 提出了时间地理学的概念；同年，Giederhold 和 Fries 在研制的医疗系统中对时态信息的处理进行了最早的尝试；Thrift 在 1977 年首次提出了历史地理信息系统（historical GIS）的概念；1978 年 Basoglu 和 Morrison 设计了最早的历史地理信息系统。该时期对时间的研究主要是围绕时间的本体和表达、不同领域中时间的作用而展开；时空信息处理的研究更多地表现在以空间为主的地理要素信息处理功能研究和以图形动画为主的表达途径研究。

随着 20 世纪 80 年代图形工作站和计算机性能价格比的迅速提高，数据库技术也日渐成熟，为发展时态技术和数据库技术的融合创造了条件。该时期的研究工作主要集中在时态历史数据库和时态数据库查询语言等方面。理论方面，Ben(1982)、Cliford(1982)和 Ginsburg(1983)三位学者分别在非第一范式时态数据库、关系型历史数据库和对象历史模型方面所进行的时空数据模型的开创性研究具有很强的代表性。实践方面，唐常杰与 Ginsburg 于 1983 年在对象历史模型的基础上，实现了一种基于关系型的历史数据库管理系统——HBase；Snodgrass 于 1985—1986 年开发的时态查询语言（TQuel），是关系数据库管理系统 INGRESS 查询语言的扩展，其目的是不作为属性而根据其语义来处理时间值。总的说来，该阶段对时空信息处理技术的研究处于低潮时期，因为同时涉及时态、空间数据库的文章数量很少，但与此相反，由于数据库技术的不断成熟和受计算机科学领域数据库专家的影响，围绕关系型历史（或时态）数据库方面的研究却蔚然成风。

Langran 于 1992 年撰写了关于时态地理信息系统（time GIS，TGIS）的第一本专著《地理信息系统中的时间》。以此为契机，世界范围内重新点燃了时空信息处理研究的热潮。研究人员意识到仅仅考虑时态数据库的存储、查询等还远远不够，时空信息处理的研究也由原先的注重具体设计细节和算法的研究，到对时空语义、时空拓扑关系、时空查询语言、时空推理过程模拟、地理对象的时空不确定性等理论和应用方面的研究和探讨逐渐增多，时空数据模型、时空信息系统、时空信息服务的构建成为关注的焦点。

1.2.2　时空数据模型

各领域中散布着不同种类的数据、信息和知识，这些数据、信息和知识通过不同媒介在各种对象之间进行传递，即不同对象主要由这些数据和由数据特征抽象出来的模型所构成。

时空数据模型是对现实地理环境的地理实体及其时空关系以及地理事件及其演变规律进行空间认知并进行形式化表达，将数据模型转换为计算机能够识别和操作处理的数据结构。使用数据模型对最终以可计算的数据结构实现的时空概念进行形式化表述为手段。数据模型包括离散的、连续的、动态的和概率性的。数据结构则表达了数据模型在某种特定的计算环境中的具体实现。

以表达和管理时空语义的时空数据模型中代表性的是 Langran(1992)从时变空间数据存储的角度，总结出了文件系统支持下的时空立方体、快照序列、基态修正和时空复合等四种时态数据模型；Hazelton(1991)进行了 4D GIS 的理论研究；Gadia 等(1991)和张师超(1993)引进了时态元素(temporal element)和时态赋值(temporal assignment)的概念，为第一范式(the 1st normal form，INF)关系的属性加上了时间参照，成为时态属性，建立了一种有特色的时态模型；Worboys(1992)提出了面向对象数据库技术的时空数据建模。

可以看出，时空数据模型的建立依赖于时间的表示方法，尽管 20 世纪 90 年代后时空数据模型的研究又经历 20 多年的历史，然而其成果仍然局限于概念模型和原型系统阶段，序列快照模型、基态修正模型、基于事件的时空数据模型、时空复合模型、时空对象模型、面向对象的时空数据模型等都无法全面解决时态理论(王家耀，2004)。目前规范化的时空数据模型正处在探索阶段。

1.2.3 时空信息系统

时空信息系统是建立在时态数据库、地理信息系统、人工智能等基础上的一种综合型应用性技术，其研究对象是时空世界中遵循着诞生、成长、生存，直至死亡等自然规律的事物和现象的时空信息。虽然时空信息系统在理论和实践等环节的研究还不十分成熟，但它是未来时空信息技术发展的一个必然趋势。

时空信息系统采集、存储、管理、分析与显示地理实体随时间变化的信息(时空信息)，它不但包含传统地理信息系统的空间特性，而且涵盖时间特性；它不但反映事物和现象的存在状态，而且表达其发展变化过程及规律。因此时空信息系统的操作对象是时空信息。传统地理信息系统只描述了研究对象的一个快照，没有对时态数据做专门的处理，因而是静态的，它只能反映事物的当前状态，无法反映对象的历史状态，更无法预测未来发展趋势。而客观事物的存在都与时间紧密相联，因此，在系统中增加对时间维的表达、分析能力，提供历史分析与趋势分析的功能，是时空信息系统的独特之处。时间、空间和属性是地理实体和地理现象本身固有的三个基本特征，是反映地理实体的状态和演变过程的重要组成部分。严格地说，空间和属性数据总是在某一特定时间或时间段内采集得到或计算产生的。

1.2.4　时空信息服务

随着人类社会的信息化进程逐步加快，以地球表面位置为参照的自然、社会、人文和经济等数据所代表的时空信息无论是在数量级还是复杂性上都在迅速地增长。人们需要有更有效的手段对存储在网络上的各种大量时空信息进行获取、处理与应用。互联网的迅速发展和普及为时空信息管理提供了新的操作平台，产生了很多相应的网络时空信息服务产品。时空信息需要被广泛共享、交换与使用，这使得时空信息的存储与处理环境从传统的集中式向目前流行的分布式方向发展。分布式就是指数据和程序分散在多个服务器，以网络上分散分布的数据及受其影响的数据处理为研究对象的一种理论计算模型。分布式有利于时空信息处理任务在整个计算机系统上进行分配与优化，克服了传统集中式系统会导致中心主机资源紧张与响应瓶颈的缺陷。很明显，传统的集中式环境很难满足分工明确的现代社会的需求，分布式环境与时空信息服务已成为时空信息处理的主要发展趋势。

1.3　时空信息处理面临新需求

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，云计算、物联网、社交网络等新兴服务促使人类社会的数据种类和规模正以前所未有的速度增长（孟小峰 等，2011），“大数据（big data）”时代已经来临。《Nature》于2008 年推出了大数据的专题讨论（http://www. nature. com/news/specials/bigdata/index. html），涉及环保、生物、互联网等众多领域，已经意识到了数据将在各个领域迅速增长。例如，近年来传感器及传感器网络正在迅猛发展并产生越来越大的数据流，急剧增长的传感器数据已经成为物联网大数据流的重要组成。麦肯锡在《大数据：下一个竞争、创新和生产力的前沿领域》（Manyika et al.，2011）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五种主要的大数据流，上述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。

目前的很多科学研究都随大数据趋势进行展开，使当前科学研究被贴上大数据的标签。时空信息处理技术的发展同样不出意外地迎来了新的挑战，挑战的同时也带来了巨大的机遇，挑战与机遇并存使得大数据成为时空信息处理技术发展的一大新动力。同时，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，时空信息处理与大数据技术互相需要。第 2 章将主要介绍大数据技术及其与时空信息的关系。大数据与时空信息处理的结合，本书称之为时空大数据技术，主要从以下几方面具体讨论时空信息处理所面临的新需求，即时空大数据的新需求。

1.3.1 时空信息处理需要大数据平台

无论是时空信息，还是大数据概念中，非结构化数据都占据极高的比例，是非常重要的资源，因此，对于时空信息处理亟须结构化数据处理，也需要非结构化数据处理技术的支持，有效处理与分析非结构化数据将给时空信息处理应用带来更大竞争优势。一方面，传统用于结构化数据的软件工具和关系型数据库在管理非结构化数据时，暴露出很多的局限性，难以满足目前时空信息处理应用的需求；另一方面，结构化和非结构化数据数量的膨胀导致了时空信息及其处理面临着数据爆炸的压力。因此，必然会催生很多新的时空信息存储、分析以及管理技术，推动面向时空信息处理的大数据平台技术的发展。

在大数据时代，时空信息处理与分析需要一个能自由灵活且高效存储的，区别于传统基础设施的新基础设施。若使用传统型关系型数据库管理系统进行数据管理，需要将非结构化数据转化为结构化数据，这个转变过程将会很耗时、延长数据分析时间并增加管理成本，可行性较差。如何减少服务器使用数量，如何减少数据存储，如何降低处理中心空间，面对这些问题，都需要一种新的平台提供用于时空信息处理的经济且高效的基础设施。

多数情形下，时空大数据平台可以由小型服务器集群组成的，采用分布式存储方式即分布式的本地独立服务器实现数据存储功能，这将必然取代原先的共享服务器集中存储的方式。分布式存储方法有很多优点，例如，可以利用其经济、高效灵活的特性，在不进行服务器和存储设备升级的前提下，快速扩展以包含数以千计的低成本服务器，这种无共享分布式模式因为不需要有限数量的共享存储设备与传输设备，所以可极大消除处理海量时空大数据时所遇到的性能瓶颈。

第 3 章将详细的从时空大数据处理硬件平台的发展、基础硬件平台与软件平台等方面阐述时空大数据平台的相关技术。

1.3.2 时空信息处理需要大数据库

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。Yannis 等(1996)定义时空数据库是一个包含时态数据和空间数据，并能同时处理数据对象的时间和空间属性的数据库。

现有的时空数据主要来源于卫星导航定位、遥感和传感器等，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。主要包含地图(矢量和栅格)、影像数据、移动对象轨迹、社交网络关系、个性化地理信息和传感器数据等。由于时空数据涉及领域广泛，多渠道的数据积累形成海量

数据。数据总量较大甚至巨大，通常以 GB、TB 或 PB 为基础单位。数据类型繁多，数以千计，包含结构化、半结构化甚至非结构化数据，且非结构化数据所占份额越来越大。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。另外，基于“价值密度与数据总量成反比”的规律，如何从价值密度较低的时空大数据中挖掘出有价值的资源对于数据挖掘研究也至关重要。

第 4 章将从时空数据的清洗，时空数据的存储，时空数据的索引和查询几个方面来探讨时空大数据管理中的相关技术。

1.3.3　时空信息处理需要大数据分析

通过提取文档、图片中的位置信息和地理标签，结合空间分析技术，可以研究各类地理事物的空间分布和时空变化。目前，以提取文本网页地理语义为目的研究集中在信息检索领域。受益于搜索引擎应用日益广泛，与地理信息相关的搜索也日益增长。据微软和雅虎公司的统计报告，用户查询的信息中 12%～16%包含地名。旺盛的需求，刺激了地理信息检索相关技术的快速发展，网络地理内容分析日趋流行。

与此同时，地理社交网络分析伴随着社交网络同步发展着，社交媒体网站的兴起是近年来一个引人注目的现象。社交媒体上的地理信息不仅具有空间特征，而且还带有用户的社交网络特征，可以结合社会网络分析技术，从中挖掘更为丰富的信息。例如，新浪微博默认在用户发表文字或者图片微博时嵌入当前的位置；大众点评、百度身边以及街旁等签到类应用使人们可以在餐馆、酒店等各种商家进行“签到”，并对商家的产品和服务进行点评；照片分享应用如 Instagram、Flickr 等使得用户在拍摄照片的时候除了可以添加文本描述信息之外，还可以附加当前的位置信息等。

随着手机、卫星导航定位设备、智能公交卡等泛在传感器被大量使用，携带用户个人信息与时空轨迹信息的数据越来越多，通过分析这类轨迹数据特征，有助于深入认识人类活动规律，并以此为基础研究来改进城市交通、优化公共设施布局。轨迹分析能够帮助分析锚点、出行范围、轨迹的形状、起止点(origin destination，OD)流、时间等方面。

集成各类时空大数据，通过对多种异构数据的整合、分析和挖掘，提取知识，还能帮助解决城市本身所面临的挑战。第 5 章将分类介绍各类时空大数据的分析方法及应用，结合新兴的城市计算技术，探讨时空大数据的集成与应用。