HZ BOOKS
华章IT

大 数 据 管 理 丛 书

度量空间数据管理

高云君 陈璐 编著

机械工业出版社
China Machine Press

大/数/据/管/理/丛/书

度量空间数据管理

高云君　陈璐　编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

度量空间数据管理 / 高云君，陈璐编著 . —北京：机械工业出版社，2018.3
（大数据管理丛书）

ISBN 978-7-111-59301-0

I. 度… II. ① 高… ② 陈… III. 度量空间－数据管理 IV. TP274

中国版本图书馆 CIP 数据核字（2018）第 039668 号

度量空间数据管理

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）

责任编辑：余 洁　　责任校对：李秋荣

印　　刷：北京诚信伟业印刷有限公司　　版　　次：2018 年 3 月第 1 版第 1 次印刷

开　　本：170mm×242mm 1/16　　印　　张：14

书　　号：ISBN 978-7-111-59301-0　　定　　价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88379426 88361066　　投稿热线：（010）88379604

购书热线：（010）68326294 88379649 68995259　　读者信箱：hzit@hzbook.com

版权所有 • 侵权必究

封底无防伪标均为盗版

本书法律顾问：北京大成律师事务所 韩光 / 邹晓东

当下大数据技术发展变化日新月异，大数据应用已经遍及工业和社会生活的方方面面，原有的数据管理理论体系与大数据产业应用之间的差距日益加大，而工业界对于大数据人才的需求却急剧增加。大数据专业人才的培养是新一轮科技较量的基础，高等院校承担着大数据人才培养的重任。因此大数据相关课程将逐渐成为国内高校计算机相关专业的重要课程。但纵观大数据人才培养课程体系尚不尽如人意，多是已有课程的“冷拼盘”，顶多是加点“调料”，原材料没有新鲜感。现阶段无论多么新多么好的人才培养计划，都只能在 20 世纪六七十年代编写的计算机知识体系上施教，无法把当下大数据带给我们的新思维、新知识传导给学生。

为此我们意识到，缺少基础性工作和原始积累，就难以培养符合工业界需要的大数据复合型和交叉型人才。因此急需在思维和理念方面进行转变，为现有的课程和知识体系按大数据应用需求进行延展和补充，加入新的可以因材施教的知识模块。我们肩负着大数据时代知识更新的使命，每一位学者都有责任和义务去为此“增砖添瓦”。

在此背景下，我们策划和组织了这套大数据管理丛书，希望能够培

此为试读，需要完整PDF请访问：www.ertongbook.com

养数据思维的理念，对原有数据管理知识体系进行完善和补充，面向新的技术热点，提出新的知识体系/知识点，拉近教材体系与大数据应用的距离，为受教者应对现代技术带来的大数据领域的新问题和挑战，扫除障碍。我们相信，假以时日，这些著作汇溪成河，必将对未来大数据人才培养起到“基石”的作用。

丛书定位：面向新形势下的大数据技术发展对人才培养提出的挑战，旨在为学术研究和人才培养提供可供参考的“基石”。虽然是一些不起眼的“砖头瓦块”，但可以为大数据人才培养积累可用的新模块(新素材)，弥补原有知识体系与应用问题之前的鸿沟，力图为现有的数据管理知识查漏补缺，聚少成多，最终形成适应大数据技术发展和人才培养的知识体系和教材基础。

丛书特点：丛书借鉴 Morgan & Claypool Publishers 出版的 Synthesis Lectures on Data Management，特色在于选题新颖，短小精湛。选题新颖即面向技术热点，弥补现有知识体系的漏洞和不足(或延伸或补充)，内容涵盖大数据管理的理论、方法、技术等诸多方面。短小精湛则不求系统性和完备性，但每本书要自成知识体系，重在阐述基本问题和方法，并辅以例题说明，便于施教。

丛书组织：丛书采用国际学术出版通行的主编负责制，为此特邀中国人民大学孟小峰教授(email：xfmeng@ruc.edu.cn)担任丛书主编，负责丛书的整体规划和选题。责任编辑为机械工业出版社华章分社姚蕾编辑(email：yaolei@hzbook.com)。

当今数据洪流席卷全球，而中国正在努力从数据大国走向数据强国，大数据时代的知识更新和人才培养刻不容缓，虽然我们的力量有限，但聚少成多，积小致巨。因此，我们在设计本套丛书封面的时候，特意选择了清代苏州籍宫廷画家徐扬描绘苏州风物的巨幅长卷画作《姑苏繁华图》(原名《盛世滋生图》)作为底图以表达我们的美好愿景，每

本书选取这幅巨卷的一部分，一步步见证和记录数据管理领域的学者在学术研究和工程应用中的探索和实践，最终形成适应大数据技术发展和人才培养的知识图谱，共同谱写出我们这个大数据时代的盛世华章。

在此期望有志于大数据人才培养并具有丰富理论和实践经验的学者和专业人员能够加入到这套书的编写工作中来，共同为中国大数据研究和人才培养贡献自己的智慧和力量，共筑属于我们自己的“时代记忆”。欢迎读者对我们的出版工作提出宝贵意见和建议。

大数据管理丛书

主编：孟小峰

大数据管理概论

孟小峰　编著

2017 年 5 月

异构信息网络挖掘：原理和方法

[美]孙怡舟(Yizhou Sun)　韩家炜(Jiawei Han)　著

段磊　朱敏　唐常杰　译

2017 年 5 月

大规模元搜索引擎技术

[美]孟卫一(Weiyi Meng)　於德(Clement T. Yu)　著

朱亮　译

2017 年 5 月

大数据集成

[美]董欣(Xin Luna Dong)　戴夫士·斯里瓦斯塔瓦(Divesh Sriva-stava)　著

王秋月　杜治娟　王硕　译

2017 年 5 月

短文本数据理解

王仲远　编著

2017 年 5 月

个人数据管理

李玉坤　孟小峰　编著

2017 年 5 月

位置大数据隐私管理

潘晓　霍峥　孟小峰　编著

2017 年 5 月

移动数据挖掘

连德富　张富峥　王英子　袁晶　谢幸　编著

2017 年 5 月

云数据管理：挑战与机遇

[美]迪卫艾肯特·阿格拉沃尔(Divyakant Agrawal)　苏迪皮托·达斯(Sudipto Das)　阿姆鲁·埃尔·阿巴迪(Amr El Abbadi)　著

马友忠　孟小峰　译

2017 年 5 月

大数据、小数据、无数据：网络世界的数据学术

[美]克莉丝汀 L. 伯格曼(Christine L. Borgman)　著

孟小峰　张祎　赵尔平　译

2017 年 8 月

实体识别技术

申德荣　寇月　聂铁铮　于戈　等编著

2017 年 10 月

度量空间数据管理

高云君　陈璐　编著

2018 年 3 月

前 言

随着计算机、互联网、通信以及定位技术的快速发展，人们获取、存储和传输数据的能力日益增强。“传感器＋互联网”形成的“数据海”融入通信、感知、计算、控制等系统中，不仅量大，而且涉及人类生活和生产的各个方面：文档、通信、新闻、电子邮件和网页上涌现的文本数据；基于位置的服务中产生的时空数据；科研机构、政府机关和企业运行过程中累积的医学、生物、气象、商业、地质数据；各类遥感、测绘、监控设备实时产生的流媒体数据；在线收集的自然环境和社会生活中的传感器数据等。1998 年图灵奖获得者吉姆·格雷曾经断言，现在每 18 个月新增的数据量等于有史以来数据量之和[1]。继人力、资本之后，数据成为一种新的非物质生产要素，是支撑科学研究和各类应用服务不可或缺的战略资源[2]。2008 年 9 月《自然》发表了名为“Big Data：Science in the Petabyte Era”的专刊，“大数据”一词开始被广泛传播。

随着数据规模的不断增长，系统资源消耗量日益增大，系统运行效率显著降低。在大数据时代，当前相关技术很难支撑人们对各类大数据进行复杂而多样的智能处理需求，“数据在，找不到”的问题日益严重[1]。中国计算机学会大数据专家委员会发布的《中国大数据技术与产业发展白皮书》明确指出，大数据相关的索引和查询技术作为大数据的主要入口之一变得尤为重要。现有的(空间)索引和查询处理技术大多

关注欧氏空间，并用欧氏距离来度量对象之间的邻近关系。但在许多实际应用(如地理信息系统、模式识别、生物计算、多媒体检索等)中，各类复杂对象(如图像、文档、基因序列等)不能使用多维向量表示，且对象之间的邻近关系并不能用欧氏距离度量，而需要用其他距离度量方式(如最短路径、编辑距离等)来衡量。因此，如何构建一个通用模型以规范表达各类数据是大数据管理的一大挑战。为此，需要借助度量空间。

本书的内容和组织结构

本书着重介绍度量空间数据管理技术，主要包括度量空间索引、度量空间查询和度量空间查询可用性。本书共分为 6 章，具体如下：

第 1 章阐述数据管理和度量空间的基本概念，并指出度量空间索引和查询存在的问题。

第 2 章介绍高效的度量空间索引，包括确定度量数据索引和不确定度量数据索引。

第 3 章介绍集中式度量空间查询处理技术，包括度量全 k 最近邻查询和度量 k 最近对查询。

第 4 章介绍分布式度量空间查询处理技术，并提出基于 MapReduce 的度量相似连接处理算法。

第 5 章介绍度量空间查询可用性分析技术，引入并解决了度量概率区域查询上的 Why-not 问题。

第 6 章介绍一个集成已有的度量空间索引与查询处理技术的分布式社交图像检索与推荐系统。

致谢

本书是作者在多年研究成果的基础上总结整理而成的。首先感谢国

家自然科学基金委和国家973计划的一贯支持，作者在近六年研究中得到了如下项目的资助：

2011～2013年，国家自然科学青年科学基金项目“障碍环境下的反最近邻查询处理技术研究”，编号：61003049。

2014～2017年，国家自然科学基金面上项目“度量空间下的k最近对查询及变体处理研究”，编号：61379033。

2015～2019年，国家重点基础研究发展计划(973计划)子课题项目“城市多模态数据的语义计算与融合”，编号：2015CB352502。

2016～2018年，国家优秀青年科学基金项目“数据库理论与系统(缺失数据处理理论与方法研究)”，编号：61522208。

2017～2020年，NSFC-浙江两化融合联合基金项目“城市多源异构数据的关联建模与可视分析”，编号：U1609217。

本书凝聚了实验室的集体智慧。特别感谢实验室的博士生柳晴、杨克宇和葛丛丛以及硕士生李信晗、邢郅豪、张远亮和仲启露。

本书可以作为普通高校计算机和信息技术相关专业的数据库与大数据管理研究生课程的教材，也可供从事计算机相关专业的技术人员和学者作为参考书。

感谢机械工业出版社华章公司的编辑，他们在全书的校对和编辑出版过程中付出了巨大的努力。

因作者水平有限，书中错误在所难免，恳请读者批评指正。

此为试读，需要完整PDF请访问：www.ertongbook.com

目 录

第1章

基本概念

本章首先概述数据管理的发展历史，其次介绍度量空间的基本概念，最后给出目前度量空间数据管理面临的挑战。

1.1 数据管理

自20世纪60年代以来，数据从单一型不断演变成复杂多样型。相应地，数据管理系统也逐渐从定制型系统逐渐演变成通用型系统。图1-1描述了数据管理系统的发展历史。

1.1.1 关系型数据管理系统

在20世纪60年代初，随着计算机在企业管理中的普及，部分大公司开始建立自己的信息管理系统，用于记录员工ID、产品价格等数值信息。因而产生了关系型数据库系统，用于管理一维数据，并支持数据选择和连接等基本操作。20世纪70年代，研究人员提出了B^+树索引以提高关系型数据管理系统的查询效率。

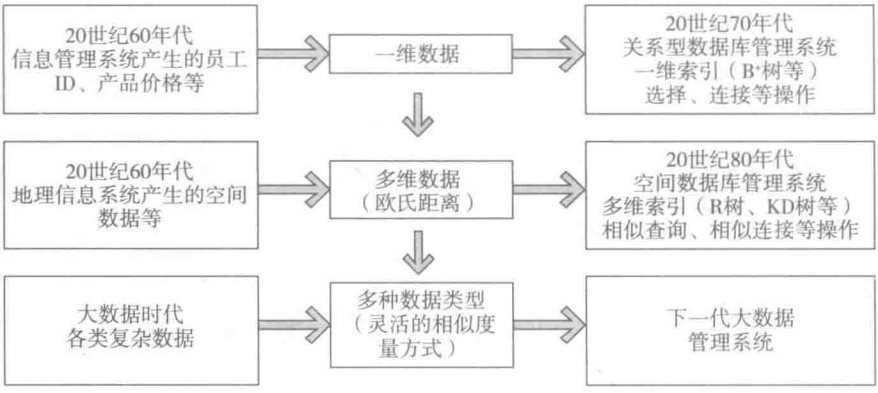

图 1-1　数据管理系统发展历史

1.1.2　空间数据管理系统

随着人造卫星的发明，地理空间信息可以通过卫星获取，地理信息系统应运而生。因此，研究人员开发了空间数据库管理系统以有效管理空间数据，其中，空间数据可以通过多维特征向量来表示，且空间数据间的相似性可以使用欧氏距离来度量。自 20 世纪 80 年代以来，研究人员提出了大量空间索引(如 R 树、KD 树等)，用于提升空间数据库管理系统的查询效率，并探讨了多种空间查询(如相似查询等)以支持多样的地理信息系统应用。

1.1.3　下一代数据管理系统

随着计算机、互联网、通信以及定位技术的快速发展，数据量呈爆炸式增长，导致大数据时代的降临。为此，需要开发通用的下一代大数据管理系统以有效地管理大数据。

大数据通常具有以下四大特性(即 4 个 V)：数据量大(Volume)、数据类型繁多(Variety)、价值密度低(Value)以及更新速度快(Velocity)。针对数据类型繁多的特性，本书采用了一个通用的数据表达模型(即度量空间)，以便对多源异构大数据进行有效建模；针对数据量大、价值密度低和更新速度快的特性，本书介绍了度量空间数据管理技术，以支持高效的多源异构大数据分析。

此为试读，需要完整PDF请访问：www.ertongbook.com