粒计算研究丛书

大数据挖掘的原理与方法

——基于粒计算与粗糙集的视角

李天瑞　罗 川　陈红梅　张钧波　著

科学出版社

粒计算研究丛书

大数据挖掘的原理与方法——基于粒计算与粗糙集的视角

李天瑞　罗　川　陈红梅　张钧波　著

科学出版社

北京

内 容 简 介

现代信息社会已经迈入大数据时代，但大数据给人们带来了前所未有的挑战，如何有效地从动态变化，结构化、半结构化和非结构化等多模态数据共存的大数据中进行高效实时的数据挖掘并发现有价值知识已成为当前信息科学领域亟待解决的问题。本书针对大数据呈现的体量巨大、多源异构、动态性和不确定性等特点，以粒计算理论为基础，以典型粗糙集模型为对象，以增量学习技术为手段，以云计算并行框架为支撑平台，构建大数据分析与挖掘的原理和方法及其算法，并融入了相关领域学者在动态知识发现、数据融合和大数据并行处理等成果，力图展现基于粒计算和粗糙集视角处理大数据的最新进展。

本书可供计算机科学与技术、智能科学与技术、软件工程、自动化、控制科学与工程、管理科学与工程和应用数学等专业的教师、研究生、高年级本科生和科研技术人员参考。

图书在版编目(CIP)数据

大数据挖掘的原理与方法：基于粒计算与粗糙集的视角/李天瑞等著. —北京：科学出版社，2016.6

(粒计算研究丛书)

ISBN 978-7-03-048368-3

Ⅰ. ①大… Ⅱ. ①李… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字 (2016) 第 114899 号

责任编辑：任 静／责任校对：桂伟利

责任印制：张 倩／封面设计：华路天然

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.scieneep.com

文林印务有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 6 月第 一 版 开本：720 × 1000 1/16

2016 年 6 月第一次印刷 印张：12

字数：211 000

定价：68.00 元

(如有印装质量问题，我社负责调换)

《粒计算研究丛书》编委会

丛 书 序

粒计算是一个新兴的、多学科交叉的研究领域。它既融入了经典的智慧，也包括了信息时代的创新。通过十多年的研究，粒计算逐渐形成了自己的哲学、理论、方法和工具，并产生了粒思维、粒逻辑、粒推理、粒分析、粒处理、粒问题求解等诸多研究课题。值得骄傲的是，中国科学工作者为粒计算研究发挥了奠基性的作用，并引导了粒计算研究的发展趋势。在过去几年里，科学出版社出版了一系列具有广泛影响的粒计算著作，包括《粒计算:过去、现在与展望》、《商空间与粒计算——结构化问题求解理论与方法》、《不确定性与粒计算》等。为了更系统、全面地介绍粒计算的最新研究成果，推动粒计算研究的发展，科学出版社推出了《粒计算研究丛书》。本丛书的基本编辑方式为：以粒计算为中心，每年选择该领域的一个突出热点为主题，邀请国内外粒计算和该主题方面的知名专家、学者就此主题撰文，来介绍近期相关研究成果及对未来的展望。此外，其他相关研究者对该主题撰写的稿件，经丛书编委会评审通过后，也可以列入该系列丛书。本丛书与每年的粒计算研讨会建立长期合作关系，丛书的作者将捐献稿费购书，赠给研讨会的参会者。中国有句老话，“星星之火，可以燎原”，还有句谚语，“众人拾柴火焰高”。《粒计算研究丛书》就是基于这样的理念和信念出版发行的。粒计算还处于婴儿时期，是星星之火，在我们每个人的爱心呵护下，一定能够燃烧成燎原大火。粒计算的成长，要靠大家不断地提供营养，靠大家的集体智慧，靠每一个人的独特贡献。这套丛书为大家提供了一个平台，让我们可以相互探讨和交流，共同创新和建树，推广粒计算的研究与发展。本丛书受益于粒计算研究每一位同仁的热心参与，也必将服务于从事粒计算研究的每一位科学工作者、老师和同学。《粒计算研究丛书》的出版得到了众多学者的支持和鼓励，同时也得到了科学出版社的大力帮助。没有这些支持，也就没有本丛书。我们衷心地感谢所有给予我们支持和帮助的朋友们！

《粒计算研究丛书》编委会

2015 年 7 月

序

近年来，随着信息技术、网络技术在国家治理、科学研究、生产实践、日常管理等各个领域的蓬勃发展，数据的产生、收集和处理的期望和驱动更加迫切，“大数据”应运而生。在过去几十年，众多学者已在机器学习、数据挖掘、人工智能、统计分析、知识工程等领域的研究中取得了丰硕的成果。但大数据带来了前所未有的挑战和机遇：数据的不确定性攀升、数据的计算规模激增、数据的实时性凸显、数据结构的复杂性和数据稀疏性等参错重出。大数据蕴涵更加丰富的知识，如何以简御繁，有效地挖掘提炼知识以满足不同应用场景的需求是一个亟须解决的问题。

粒计算理论是近年来新兴的一个研究领域，是信息处理的一种新的计算范式，主要用于描述和处理不确定的、模糊的、不完整的和海量的信息及提供一种基于粒和粒间关系的问题求解方法。粒计算的模型与现实世界的结构、人们的思维模式及行为方式是一致的。它提供了对所解决问题多视角、多层次的理解、概括和操作。用粒计算指导的思维模式和行为方式可将复杂问题分解成若干小问题，再进行分而治之。粒计算可对问题进行不同层次的抽象和处理，寻求不同粒度上的近似解。在大数据环境下，充分利用粒计算的特性进行问题求解和智能信息处理是一个行之有效的方式。

粒计算的主要理论分支包括粗糙集、词计算、商空间理论和三支决策等。其中，粗糙集理论近年来越来越受到人们的重视。它聚焦于不确定信息的近似逼近，用上下近似集对不确定信息进行近似描述，无需先验知识，进而拓展到特征选择、逻辑推理、关联规则、粒化模型构造等相关的研究。基于统计学的智能信息处理需要数据分布等先验知识，而在大数据环境下，数据的统计信息的获取和假设有诸多限制，不同的假设可能影响最终的结果。粗糙集理论从数据的粒度结构出发，可以刻画出不同的粒度结构和近似描述，能有效地适应大数据环境去繁就简和数据融合的需要。近年来，粗糙集理论得到了蓬勃发展，其有效性已在许多科学与工程领域的成功应用得到证实，尤其是将概率论、贝叶斯决策理论和粗糙集理论相结合的决策粗糙集理论，已在风险决策和不确定信息处理方面得到了很好的应用。

基于粒计算和粗糙集理论与方法的大数据处理技术已得到了研究者的广泛关注。本书针对大数据呈现的体量巨大、多源异构、动态性和不确定性等特点，以粒计算理论为基础，以典型粗糙集模型为对象，以增量学习技术为手段，以云计算并行框架为支撑平台，构建大数据分析与挖掘的原理和方法及其算法，并融入了相关领域学者在动态知识发现、数据融合和大数据并行处理等方面的成果，反映了基于

粒计算和粗糙集视角处理大数据的最新进展。本书汇聚了作者多年来在这个领域的创新研究成果，各章节内容选题恰到好处，不但系统梳理了大数据挖掘的产生背景、基于粒计算和粗糙集理论的动态知识发现国内外研究进展，而且分别针对大数据挖掘中几个主要挑战性问题给出作者的独特解决方案。其主要贡献包括三方面：①全面刻画了大数据中动态特性的三种粒度变化类型，提出了面向大数据基于粒计算和粗糙集理论的动态知识发现系列方法及相应的算法，设计了数据并行、模型并行以及数据-模型并行方法，并在流行的云平台 Hadoop 和 Spark 上进行多机和多核并行的全方位性能评测；②充分利用增量学习技术，即能够有效利用已有知识，通过对新数据的增量分析处理，从而实现知识的渐进性修改和更新，揭示了不同粒层次之间转换的数学关系，刻画了不同粒度之间的变化规律，提高了数据快速增长时的知识获取效率；③借助矩阵这一通用的表达和运算工具，精巧地刻画了数据矩阵、关系矩阵、诱导对角矩阵、等价类特征矩阵和特征值矩阵、决策特征矩阵和特征值矩阵以及属性重要度矩阵等概念，创新了近似集的布尔矩阵表示，阐明了等价类的泛化决策与近似集之间的关系，展示了决策特征矩阵和分配辨识矩阵的更新来实现知识的动态更新过程，所设计的基于矩阵增量学习方法不但简明、直观性较好，而且更新知识效率高，为大数据中的动态知识发现提供了一个新的技巧与工具。

另外，书中在刻画概念、引理、定理和算法等内容时配有相当数量的图解和实例，以便于读者充分理解书中的知识点，突出基于粒计算和粗糙集理论应用于大数据分析与挖掘的直观性。该书充分展示了利用粒计算和粗糙集理论解决大数据复杂问题的优势，为大数据分析、处理与挖掘提供了新的理论支撑和技术支持，同时对于推动大数据产业的快速发展具有重要的现实意义和应用价值。通过阅读此书，读者可以了解到基于粒计算理论的大数据挖掘方法和其相关领域的知识背景以及获取此研究领域最前沿的信息。本书很适合作为参考资料或者研究生的课程教材。

潘　毅
美国佐治亚州立大学
2016 年 5 月 16 日

前　言

随着新兴信息技术和应用模式的快速产生与发展，现代信息社会已经迈入大数据时代。大数据的分析、挖掘与应用也已经渗透到国家治理、经济运行、文化建设和社会管理的方方面面。大数据具有体量巨大、类型繁多、价值密度低、变化速度快和数据真实性等特点，它给人们带来了前所未有的挑战。如何有效地从动态变化，结构化、半结构化和非结构化等多模态数据共存的大数据中进行高效实时的数据挖掘并发现有价值的知识已成为当前信息科学领域亟待解决的科学问题。

大数据中提炼出的知识将在更高的层面、更广的视角、更大的范围帮助人们提高洞察力，提升决策力，将为人类社会创造前所未有的重大价值。而在大数据环境中，由于数据采集手段的不足、测量产生的误差和人为因素等导致数据的非真实性特征更加鲜明、不确定性更加显著，因此不确定性问题处理成为从大数据中发现有用知识极其困难的挑战。如何在数据分析与挖掘阶段对大数据的不确定性问题进行有效的处理，已成为大数据知识获取的一个重要研究课题。粒计算作为一种新的计算范式，为我们提供了一套基于信息粒化的复杂问题求解理论框架，是当前计算智能领域中模拟人类思维和解决复杂问题的核心技术之一。通过对复杂问题的抽象、划分从而转化为若干较为简单的问题，粒计算可以从不同粒度层次的角度对复杂问题进行多层次、多视角的简化分析与处理，并通过忽略不必要的求解细节来提高问题处理的计算效率。粗糙集理论是不确定信息近似处理的一种重要粒计算模型，其利用信息的已知划分，通过上下近似集对不精确或不确定的目标概念进行近似刻画，从而不需要待处理数据之外的任何先验信息。粒计算和粗糙集理论所具有的复杂问题求解优势将为大数据环境中不确定性问题的近似求解、处理与解决提供重要的理论依据。

采集、分析大数据是一个持续更新、不断优化的升级过程。“大数据”由“小数据”发展而来，数据随着时间的推移，产生得快，变化得快，折旧得也快，数据快速增长化成为大数据的另一个重要特征，数据的激增使得大数据环境中信息处理的时效性要求越来越高。如何分析、设计动态高效的知识获取方法来应对大数据环境中数据处理的时效性需求也已成为当前信息科学领域亟待解决的挑战性任务。传统的批量式知识获取方法在面对不断变化的动态数据环境，随着问题求解规模的不断增大，对时间和空间的需求也会迅速增大，从而导致知识学习的速度远不

及数据更新的速度。增量更新技术模拟人类的认知机理，能够在不断变化的动态数据环境中实现基于增量数据的渐进性知识修正、加强、更新和维护，为我们降低了数据快速增长时知识获取方法对时间和空间的需求，对于提高大数据挖掘和分析的实时处理能力，实现从复杂海量数据到潜在有用知识的高效转化具有重要的借鉴作用。

随着数据量的不断增加和问题求解规模的不断扩张，传统的基于串行计算技术的数据挖掘模型及算法无法满足大数据环境中人们对响应时间、吞吐量的可伸缩性要求。并行计算是提高计算机系统计算速度和处理能力的一种有效手段。云计算是由并行计算、网格计算、分布式计算和效用计算等发展而来的一种基于互联网的新兴的计算模式。它可为人们提供各种不同层次、各种不同需求的低成本、高效率的智能化服务及信息服务模式的改变。云计算中并行技术可最大限度地整合计算存储资源，能够有效应对多源异构动态数据挖掘时信息数据异构分布、计算资源利用不足的效率瓶颈。基于云计算技术提高大数据处理效率是符合当前智能信息处理的发展趋势。因此，充分应用云计算并行技术来优化基于粒计算和粗糙集理论的大数据学习算法，以突破粒计算和粗糙集理论应用于大规模复杂动态数据中实时处理的效率瓶颈问题，推动大数据分析处理理论、方法及其算法的发展与完善。

本书旨在基于粒计算和粗糙集理论，利用增量学习技术和并行计算模型，以大数据环境中的不确定性问题和实时分析处理为研究目的，开发高效实用的大数据挖掘与学习算法。本书的研究工作不仅拓展了粒计算与粗糙集理论及应用的研究范畴，为大数据环境中的数据挖掘与知识发现问题提供了新的处理技巧和研究视角，而且可以促进大数据产业的快速发展，加快实现数据增值服务，具有重要的理论意义和实际应用价值。

本书共 7 章，第 1 章综述大数据挖掘、粒计算与粗糙集理论的研究现状；第 2 章给出本书的预备知识，包括经典粗糙集理论与扩展粗糙集模型，以及基于粗糙集理论的属性约简方法和粒度度量等；第 3 章介绍面向大数据的并行大规模特征选择方法；第 4 章介绍多维粒度动态变化下粗糙近似集的增量更新方法；第 5 章介绍信息系统中属性值粗化细化时决策规则动态更新方法；第 6 章介绍动态不完备数据环境中概率粗糙集模型及其近似集的高效求解方法；第 7 章介绍复杂数据融合与高效学习方法。

本书的工作得到了很多专家和同行的帮助，包括比利时国家核能研究中心阮达研究员，加拿大里贾那大学姚一豫教授和姚静涛教授，美国佐治亚州立大学潘毅教授，台湾科技大学洪西进教授，西南交通大学徐扬教授、秦克云教授和刘盾副教授，重庆邮电大学王国胤教授，同济大学苗夺谦教授，山西大学梁吉业教授、李德

玉教授和钱宇华教授，南京大学周献忠教授和商琳副教授，天津大学胡清华教授和代建华教授，浙江海洋学院吴伟志教授，河北师范大学米据生教授，河南师范大学徐久成教授，闽南师范大学李进金教授和祝峰教授等。

本书的出版受到了国家自然科学基金项目 (No. 61573292, 61572406) 的资助，在此表示衷心感谢。另外，由于作者水平有限，书中不足之处在所难免，敬请读者指正 (联系方式：trli@swjtu.edu.cn)。

作　者

2016 年 4 月

目　录

第 1 章　绪论 ····· 1
1.1　大数据及其挖掘技术 ····· 1
1.2　粒计算理论 ····· 6
1.3　粗糙集理论 ····· 8
1.4　基于粒计算与粗糙集的数据挖掘 ····· 11
1.4.1　面向海量数据的数据挖掘 ····· 11
1.4.2　面向动态数据的数据挖掘 ····· 12
1.4.3　面向复杂数据的数据挖掘 ····· 14
1.5　本章小结 ····· 15
第 2 章　预备知识 ····· 16
2.1　经典粗糙集模型 ····· 16
2.2　面向复杂数据的扩展粗糙集模型 ····· 18
2.2.1　邻域粗糙集模型 ····· 18
2.2.2　集值粗糙集模型 ····· 20
2.2.3　不完备粗糙集模型 ····· 20
2.3　面向有噪声数据的概率粗糙集模型 ····· 22
2.3.1　0.5 概率粗糙集模型 ····· 22
2.3.2　决策粗糙集模型 ····· 23
2.3.3　变精度粗糙集模型 ····· 23
2.4　属性约简 ····· 24
2.4.1　属性约简的基本框架 ····· 24
2.4.2　启发式属性约简 ····· 25
2.4.3　不协调属性约简 ····· 29
2.5　粒度度量 ····· 31
2.6　本章小结 ····· 32
第 3 章　并行大规模特征选择 ····· 33
3.1　并行特征提取方法 ····· 33
3.1.1　模型并行方法 ····· 33
3.1.2　数据并行方法 ····· 34
3.1.3　模型-数据并行方法 ····· 35

3.2 并行特征提取算法 ····· 35
3.2.1 评价函数的统一表示 ····· 36
3.2.2 评价函数的分治方法 ····· 36
3.2.3 基于 MapReduce 的并行属性约简算法 ····· 39
3.2.4 基于 Spark 的并行属性约简算法 ····· 43
3.2.5 基于粒计算的并行属性约简加速算法 ····· 44
3.3 实验分析 ····· 46
3.3.1 数据集和实验平台 ····· 47
3.3.2 与串行算法的对比 ····· 47
3.3.3 不同并行算法的对比 ····· 51
3.3.4 高维数据上的表现 ····· 52
3.3.5 天文大数据上的应用 ····· 53
3.4 本章小结 ····· 54
第 4 章 近似集动态更新 ····· 55
4.1 知识粒度的变化性质 ····· 55
4.2 基于粒的近似集增量更新方法 ····· 57
4.2.1 等价类特征矩阵 ····· 58
4.2.2 等价类特征矩阵更新原理 ····· 59
4.2.3 基于粒的近似集更新算法 ····· 61
4.3 算例 ····· 64
4.4 算法复杂度分析 ····· 68
4.5 实验方案及性能分析 ····· 68
4.5.1 实验方案 ····· 68
4.5.2 实验结果 ····· 69
4.6 本章小结 ····· 74
第 5 章 规则动态更新 ····· 76
5.1 等价类的向量和矩阵表示 ····· 76
5.2 等价类的泛化决策性质 ····· 77
5.3 基于最小辨识属性集的约简生成 ····· 78
5.3.1 最小辨识属性集及其生成算法 ····· 78
5.3.2 属性重要度矩阵 ····· 80
5.3.3 约简的生成 ····· 81
5.3.4 算例 ····· 82
5.4 属性值粗化细化的定义及性质 ····· 84
5.4.1 属性值粗化细化的定义 ····· 84

5.4.2 决策信息系统的动态性质……85
5.4.3 决策规则的动态性质……87
5.5 属性值粗化时规则更新原理及算法……92
5.5.1 属性值粗化时规则更新原理……92
5.5.2 属性值粗化时规则更新算法……95
5.5.3 算例……97
5.6 属性值细化时规则更新原理及算法……99
5.6.1 属性值细化时规则更新原理……99
5.6.2 属性值细化时规则更新算法……102
5.6.3 算例……104
5.7 算法复杂度分析……106
5.8 实验方案及性能比较……109
5.8.1 实验方案……109
5.8.2 实验结果……110
5.9 本章小结……113
第 6 章 面向缺失数据的动态概率粗糙集方法……114
6.1 面向缺失数据的概率粗糙集模型……115
6.2 面向对象更新的动态概率粗糙集方法……115
6.2.1 条件概率的增量估计策略……115
6.2.2 概率粗糙近似集的增量更新方法……121
6.3 算法设计与分析……123
6.4 算例……128
6.5 实验方案与性能分析……130
6.5.1 实验方案……131
6.5.2 性能分析……132
6.6 本章小结……139
第 7 章 复杂数据融合与高效学习算法……141
7.1 复合粗糙集模型……141
7.2 近似集的矩阵表示方法……144
7.2.1 基于矩阵运算的近似集构造方法……144
7.2.2 基于布尔矩阵的近似集表示方法……146
7.2.3 基于布尔矩阵的近似集计算方法……147
7.3 算法设计与复杂度分析……148
7.3.1 基于布尔矩阵的近似集计算算法……148
7.3.2 基于矩阵的近似集计算的批处理算法……149

7.4 近似集的多核并行计算方法……150
7.4.1 近似集的并行计算方法……151
7.4.2 GPU 架构与 CUDA……152
7.4.3 基于 Single-GPU 的近似集计算算法……154
7.4.4 基于 Multi-GPU 的近似集计算算法……155
7.5 实验分析……157
7.5.1 实验设置……157
7.5.2 批处理算法的性能……158
7.5.3 GPU 算法的性能……159
7.5.4 Multi-GPU 算法的性能……161
7.6 本章小结……163
参考文献……164

第1章 绪 论

1.1 大数据及其挖掘技术

近年来，随着互联网、物联网、云计算和三网融合等信息与通信技术的迅猛发展，数据的快速增长成了许多行业共同面对的严峻挑战和宝贵机遇，信息社会已经进入了“大数据”时代。“大数据”是一个抽象的概念，若仅从字面来看，它是指数据规模巨大。但是光凭体量巨大这一点显然无法区别大数据与以往的“海量数据”和“超大规模数据”等概念。人们目前对于大数据还没有一个公认的定义，现有的一些大数据定义基本上都是从其特征出发来刻画。其中，维基百科将大数据定义为“所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理并整理成为人类所能解读的信息”；百度百科中大数据的概念是“所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策更积极目的的资讯”。简而言之，大数据与一般数据的区别在于：大数据是指不能用传统存储技术和算法在合理的时间内进行分析与处理。

当前大数据在各个领域中开始崭露头角，取得了令人瞩目的成就。例如，在社会民生方面，2015 年中国春运大军已经增长到 36 亿人次，人们很关心这 36 亿人次在这么短的时间内是如何迁徙的，央视借助百度迁移 (用手机中基于位置服务的定位功能和大数据可视化技术等) 把春运大军的迁徙状况形象地呈现在电视屏幕上，给每一位观众带来最直观的感受，也为运输部门的决策提供了重要参考依据。阿里金融的阿里小贷业务也堪称大数据应用中的典型案例，其目的是为阿里巴巴 B2B 业务、淘宝和天猫三个平台的商家提供订单贷款和信用贷款。阿里巴巴利用该集团中庞大的客户资源大数据和信息流，通过分析淘宝、天猫、支付宝和 B2B 上商家的各种类型数据，对商家进行信用评级，商家凭借这个信用评级，不用提交任何担保、抵押，就可以申请阿里金融旗下的信贷产品。与现有银行相比，这种创新的金融信贷审批模式极大地提高了贷款效率和企业竞争力。

大数据的不断迅猛发展也呈现出其独特的特性，可概括为以下五方面，也称为“5V”。

(1) 数据量大 (volume)。数据集的规模不断扩大，已从 GB 到 TB 再到 PB 级，甚至已经开始以 EB 和 ZB 来计数。截至目前，人类生产的所有印刷材料的数据量是 200PB，而历史上全人类说过的所有话的数据量大约是 5EB。根据 IDC 的“数

字宇宙”的报告，预计到 2020 年，全球数据使用量将达到 40ZB[1]。

(2) 种类繁多 (variety)。相对于以往便于存储的以文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片和地理位置信息等，现代互联网上半结构化和非结构化数据占有比例将达到整个数据量的 95% 以上，这些多类型的数据对数据的处理能力提出了更高的要求[2]。

(3) 速度快 (velocity)。大数据区别于传统数据挖掘的最显著特征是它往往以数据流的形式动态、快速地产生，具有很强的时效性，用户只有把握好对数据流的掌控才能有效利用这些数据[3]。

(4) 价值密度低 (value)。基于传统思维与技术让人们在实际环境中往往面临信息泛滥而知识匮乏的窘态，呈现出价值密度的高低与数据总量的大小成反比的情况。但对于众多潜在的应用而言，大数据整体往往蕴藏着巨大的价值[2]。

(5) 真实性 (veracity)。现实世界中的数据普遍存在模糊性、不一致性或含有噪声，例如，当传感器受到外界干扰时，将导致所获得的数据存在误差等。

大数据的涌现不仅改变着人们的生活与工作方式、政府的管理方法和企业的运作模式，甚至还引起科学研究模式的根本性改变。大数据是与自然资源、人力资源一样重要的战略资源，隐含着巨大的社会和经济等价值，已引起了各行各业的高度重视[4]。近几年，*Nature* 和 *Science* 等国际顶级学术刊物相继出版专刊来专门探讨大数据的研究。其中，2008 年 *Nature* 出版的专刊从互联网技术、网络经济学、超级计算、环境科学、生物医药等多方面介绍了大数据带来的技术问题和挑战[5]。2011 年 *Science* 推出的专刊讨论了数据洪流所带来的挑战，特别指出，倘若能够更有效地组织和使用这些大数据，人们将得到更多的机会发挥科学技术对社会发展的巨大推动作用[6]。2012 年 4 月欧洲信息学与数学研究协会会刊 *ERCIM News* 也出版了专刊，讨论大数据时代的数据管理、数据密集型研究的创新技术等问题，并介绍了欧洲科研机构开展的研究活动和取得的创新性进展[7]。IEEE 计算机学会决定，从 2013 年开始，每年举办一次 IEEE Big Data 国际学术会议，并创办了 *IEEE Transactions on Big Data* 等学术期刊。中国计算机学会于 2012 年成立了大数据专家委员会，其宗旨是探讨大数据的核心科学与技术问题，推动大数据学科方向的建设与发展，构建面向大数据产学研用的学术交流、技术合作与数据共享平台，并为相关政府部门提供战略性的意见与建议，已连续发布了《中国大数据技术与产业发展报告》和《大数据发展趋势预测报告》等。Elsevier、Springer 等科技出版社也于近年来相继创刊了大数据方面的国际期刊。上述情况表明，大数据已成为一门新兴科学并已受到科技界的广泛重视[2]。不仅如此，许多国家政府对大数据技术与应用研究给予了高度的重视和关注。2012 年 3 月，美国政府宣布投资 2 亿美元启动“大数据研究和发展计划”，认为大数据是“未来的新石油”，将“大数据研究”上升为国家意志，对未来的科技与经济发展必将产生深远影响[8]。同年日本

政府推出了新的综合战略“活力 ICT 日本”，重点关注大数据应用所需的云计算、传感器和社会化媒体等智能技术开发[2]。2013 年，英国政府宣布投资 6 亿英镑发展大数据等 8 类高新技术，其中信息行业新兴的大数据技术将获得 1.89 亿英镑，占据总投资的近 1/3。同年，澳大利亚政府也出台了其大数据战略规划方案。我国科技界与信息技术密切相关的产业界对大数据技术与应用的关注程度正在逐渐增强，并引起了政府相关部门的重视。2015 年我国也发布了《促进大数据发展行动纲要》，这是指导我国大数据发展的国家顶层设计和总体部署。还有，中国科学院先后于 2012 年 5 月、2013 年 5 月、2014 年 10 月和 2015 年 10 月组织召开了题为“大数据科学与工程”、“数据科学与大数据的科学原理及发展前景”、“科学大数据的前沿问题”和“健康科学大数据与精准医学”香山科学会议。国家发改与地方政府主导的“智慧城市”计划已开始实施，部分省份已经建成或正在建设一批大数据中心。科技部已经部署了若干大数据或与大数据密切相关的 973 计划和专项研究计划[2]。

Big data is worth nothing without big science. As with gold or oil, data has no intrinsic value. Big science, which bridges the gap between knowledge and insight, is where the real value is.

—— Webtrends CEO Alex Yoder①

大数据的出现是前所未有的挑战，也是千载难逢的机遇。数据的复杂性本身有可能隐含更多有价值的信息，如何有效挖掘大数据中蕴涵的知识，使之服务于社会生活的方方面面，是科研工作者、工程技术人员、管理者所共同关注的焦点。中国工程院院士李国杰列出了以下几个值得高度重视的问题[8]：①数据的去冗余和高效率低成本的数据存储；②新的数据表示方法；③数据融合；④高效处理非结构化和半结构化数据；⑤适合不同行业的大数据挖掘分析工具和开发环境；⑥大幅度降低数据处理、存储和通信能耗的新技术。近年来，众多学者针对以上问题，围绕大数据体量大、动态性强、不确定性和多源异构等特点展开了深入的分析和探讨。

1. 针对大数据体量大特点的研究

数据的快速增长给互联网公司带来了极大的挑战，为应对海量数据，Google 公司于 2003 年开始依次公布了 GFS(Google File System)[9]、MapReduce[10] 与 Bigtable[11] 三篇技术论文，为大数据存储和计算等问题提供了一个全新的解决思路。受到这些技术的启发，Apache Software Foundation 公司开发了分布式并行计算系统 Hadoop②。Hadoop 具有吞吐量大、自动容错等优点，在海量数据处理上得到了广泛的使用，现已成为当今世界上最热门的大数据处理平台之一。近年来，随

① http://www.cnet.com/news/big-data-is-worth-nothing-without-big-science.

② http://hadoop.apache.org.