大数据的若干基础研究方向

朱扬勇[1,2]，熊赟[1,2]
1. 复旦大学计算机科学技术学院，上海 200433；
2. 上海市数据科学重点实验室，上海 200433

摘要
大数据问题的关键技术挑战在于：找到隐含在低价值密度数据中的价值；在希望的时间内完成。指出前者需要将领域知识和数据技术结合，这种结合的理论和新型算法构成大数据的分析基础和应用基础；后者需要设计新的计算机、集群体系、计算框架、存储体系和数据管理方法，这些构成大数据的计算基础和数据基础。另外，这两个挑战都涉及数学理论，这是大数据的数学基础。系统地分析了大数据的数学基础、计算基础、数据基础、分析基础和应用基础等基础研究方向。

关键词
大数据；数据科学；大数据基础；数据资源

中图分类号：TP311 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017023

Foundation issues for big data research

ZHU Yangyong[1,2], XIONG Yun[1,2]
1. School of Computer Science, Fudan University, Shanghai 200433, China
2. Shanghai Key Lab of Data Science, Shanghai 200433, China

Abstract
The key technical challenges for big data lie in how to discover the value of the low-value-density data and how to complete the task in the desired time. The ways to take up these challenges from three aspects were discussed. First is that the former challenge requires the combination of domain knowledge and data technology. This combination of theory and new algorithms forms the basis of application and analysis of big data. Second is that the latter challenge needs to design new types of computer, cluster system, computing framework, storage system and data management method, which forms the basis of computing and data of big data. Thirdly, both challenges relate to mathematical theory, which is the basis of mathematics of big data. In conclusion, several foundation issues for big data research including the basis of mathematics, computing, data, analysis and application of big data were analyzed.

Key words
big data, data science, foundation of big data, data resource

1 引言

数据资源开发利用的目的是实现大数据价值。大数据是指为决策问题提供服务的大数据集、大数据技术和大数据应用的总称[1]；大数据问题是指不能用当前技术在决策希望的时间内处理分析的数据资源开发利用问题[1]。就当下而言，现有技术难以处理PB级别以上的数据。PB级别的数据规模已经超过了大部分企业和机构自身积累的业务数据规模，这说明PB级别数据规模的应用涉及了企业和机构外部的数据，这是质的转变。在PB级别数据规模下，企业和机构可以用历史数据、跨界数据等足够多的数据来做决策。这将对科学研究、政府治理、民生改善、产业发展发挥革命性的作用。PB级别数据规模带来的技术挑战也是非常明显的，即“大数据问题”。从技术角度看，PB级别的数据量给计算机的硬件和软件都带来了挑战：机械硬盘、盘阵、体系结构、计算框架、数据移动、多地计算/异地计算、数据管理、数据分析等。随着计算技术的进步，这些技术问题是否就会得到解决？回答是否定的。摩尔定律（Moore’s Law）不能解决大数据问题，摩尔定律是由英特尔（Intel）创始人之一戈登·摩尔（Gordon Moore）提出来的[2]。其内容为：当价格不变时，集成电路上可容纳的元器件的数目，每隔18~24个月便会增加一倍，性能也将提升一倍。换言之，每一美元所能买到的电脑性能，将每隔18~24个月翻一倍以上。这一定律揭示了信息技术进步的速度。自从PC时代以来，每次技术进步生产的计算机以亿万台计，每台计算机都在生产数据。现在想将很多台计算机生产的数据集中起来处理和分析，显然是现有技术不能实现的。这就是大数据问题产生的根源。大数据问题的关键技术挑战在于：找到隐含在低价值密度数据资源中的价值；在希望的时间内完成所有的任务。前者需要领域业务知识和数据技术相结合，这种结合的理论和新型算法构成了大数据的分析基础和应用基础；后者需要设计新的计算机、集群体系、计算框架、存储体系和数据管理方法，这些构成了大数据的计算基础和数据基础。另外，这两个挑战都涉及了一些抽象的数学理论，这是大数据的数学基础。面对大数据问题的挑战，需要研究如何发展大数据，如何从实践总结经验，抓住核心问题重点突破，实现原始创新，从基础做起，掌握核心技术，探索大数据的基础理论，为解决大数据如何用、如何管、如何算等关键问题提供理论指导。

2 大数据的基础

信息化和大数据二者的关系是“不混淆、不冲突”。

- 不混淆是指信息化和大数据是不同的，不能混为一谈，信息化是生产数据的，大数据是开发数据的。从技术角度来看，信息化是技术进步促进数据增长；大数据是数据增长促进技术进步。
- 不冲突是指大数据不取代信息化，信息化不包含大数据。这样，信息化工作照样做，并且信息化仍然将快速发展。但大数据已经从信息化工作中独立出来，如果说信息化对应的技术叫信息技术（information technology，IT），那么大数据对应的技术可以叫数据技术（data technology，DT）。

核心点是数据增长促进技术进步。例如，当前10 PB规模的数据无论是从上海到北京，或者是从一个数据中心到另外一

个数据中心，甚至从一个存储集群到另一个存储集群，都是困难的、难以忍受的。既然数据大到难以移动，那么如何集成多个数据资源进行计算，如何设计数据不移动、软件移动的计算模型，实现多地计算、异地计算。又如，如何开发利用数据资源，如何确保数据安全和数据主权，这些都是数据量增长对技术进步的要求。

信息化的技术和大数据的技术是不同的，参考文献[1]给出了二者技术的对比。这样，信息化的基础研究和大数据的基础研究也是不一样的。大数据的基础包括：应用基础、分析基础、数据基础、计算基础和数学基础5个方面。

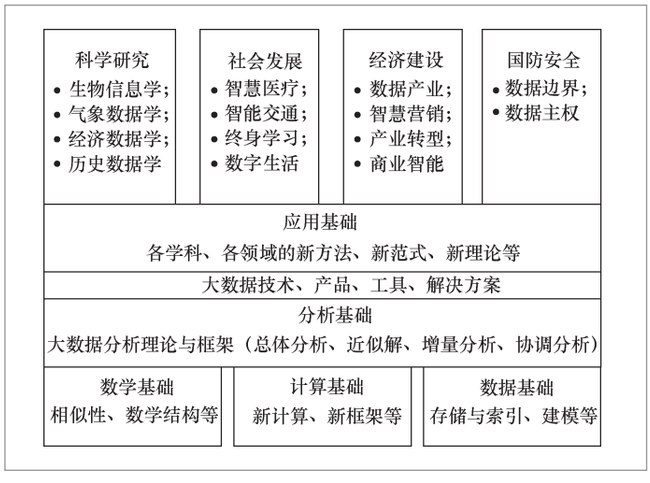
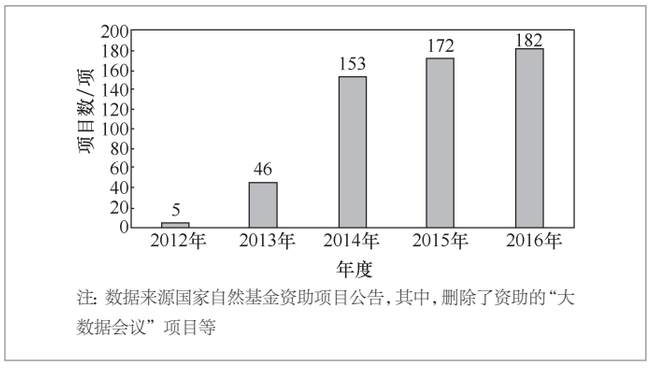


图1 大数据基础逻辑关系



注：数据来源国家自然基金资助项目公告，其中，删除了资助的“大数据会议”项目等

图2 2012—2016年国家自然科学基金资助的以“大数据”为主题词的项目数

大数据的应用基础包括各学科、各领域的基于数据的新方法、新范式、新理论等，用于支撑基于大数据的科学研究方法、社会发展方式、经济建设模式和国防安全手段。大数据的应用基础是建立在大数据技术、产品、工具和解决方案之上的，而这些产品和工具的开发需要大数据的分析基础。大数据的分析基础包括大数据分析理论与框架、大数据分析方法和算法、业务驱动的分析理论和方法等，大数据分析方法和算法的实现和实施需要大数据的数据基础、计算基础和数学基础。大数据的数据基础包括大数据的治理和管理、存储理论和模型、可视化等；大数据的计算基础包括多地计算/异地计算、计算框架、硬件设备、网络设备等；大数据的数学基础包括数据的数学结构、数据代数、数据相似性等。**图1**给出了大数据基础的逻辑关系。

自2012年起，国家自然科学基金委员会对大数据研究开始立项，总体资助情况分布如**图2**所示。

从**图2**可以看出，在这5年期间，大数据相关项目资助数持续增长，尤其是在2014年，相对于2013年有较大幅度的增长。2012年，大数据概念刚刚兴起，国家自然科学基金项目申请时间点已过，所以2012年几乎没有什么项目。2013年，各学部开始支持大数据项目，全社会掀起大数据热，各行各业都在讨论大数据。于是，到2014年，大数据项目大幅度增长，主要是应用基础的研究项目大幅度增长，几乎每个学部都支持了大数据的研究项目（如**图3**所示）。受资助的大数据相关项目在应用基础、分析基础、数据基础、计算基础和数学基础各方面的项目资助占比情况如**图4**所示。总体来看，科研项目关注最多的是对各个领域大数据应用问题的研究（应用基础占62%），其次主要集中在大数据挖掘

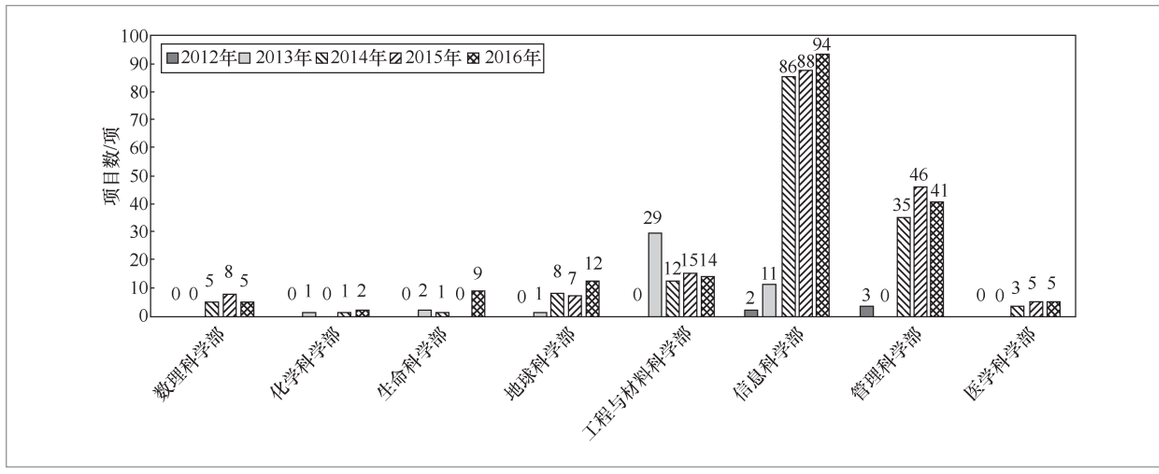


图 3 2012—2016 年国家自然科学基金委员会各学部资助项目数分布

方法研究(即分析基础占18%),其他依次为:数据基础(占11%)、计算基础(占7%)和数学基础(占2%)。

从2012—2016年每年的占比情况来看(如**图5**所示),应用基础有明显的上升趋势;分析基础在2013年有明显的上升,随后占比较为平稳;数据基础历年占比都较为平稳,且所占比重也并不大;计算基础所占比例呈现出逐年下降的趋势(2016年稍有回升);数学基础被关注的比例历年很低,并维持稳定(低比例)。

从**图5**可以看出,在2012年大数据发展初期,计算框架和计算能力是推动大数据发展最急需的基础,而之后随着开源计

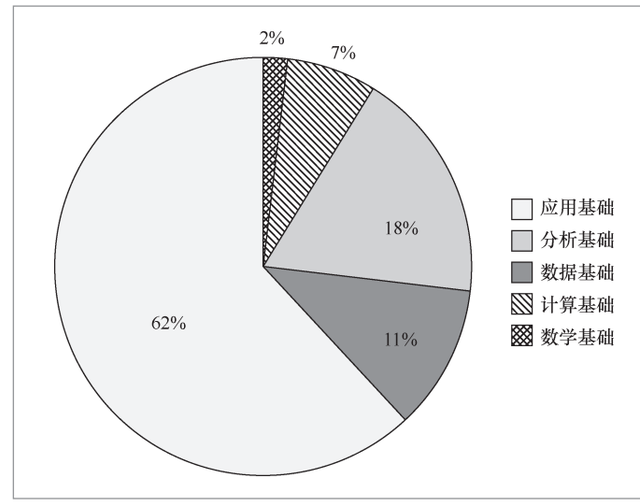


图 4 2012—2016 年在五大基础方面的项目数总占比

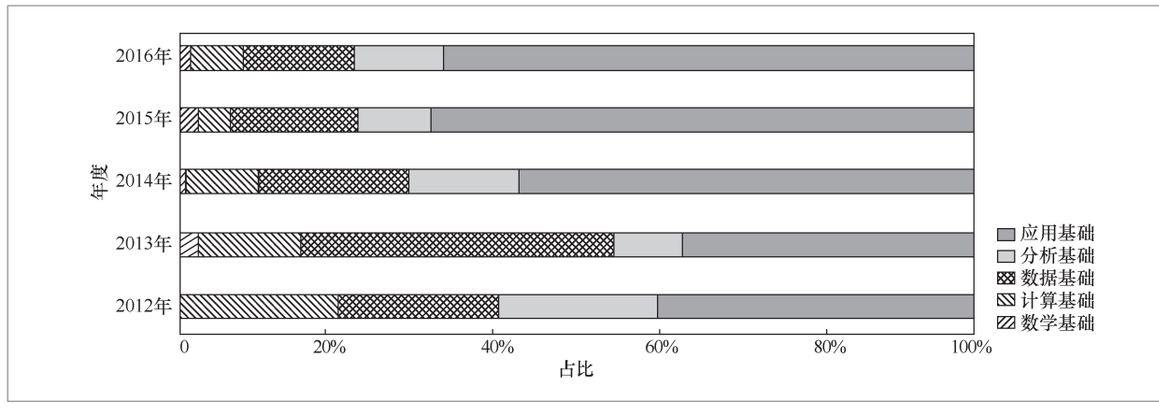


图 5 2012—2016 年在五大基础方面的项目数年度占比

算框架的出现，计算基础的比例又开始下降。然而，2016年，数据开放共享成为趋势和重点，数据迁移、异地交换的需求又促进研究者探索新的计算框架。并且，分析基础在2013年的突增也说明当时对大数据方法需求的增长，随后相对稳定。在计算基础下降的过程中，应用基础占比逐渐上升，这说明越来越多的领域参与到大数据的研究中来。

图6为2012—2016年国家自然科学基金项目在五大基础方面的项目数的变化情况。可以看出，大数据研究过程中数学基础方面还需要更多的研究者参与和关注。在应用基础研究方面目前已经涉及众多领域，应该总结共性技术和理论，减少重复研究。数据的资源性迫切需要开放共享，数据基础的新问题、新技术、新理论的研究应该建立起来。

3 大数据的应用基础

大数据的应用渗透到越来越多的领域，各领域大数据理论和方法的研究将为创新大数据应用、提升大数据价值奠定基础，创造出基于大数据的新型科学研究、管理决策、社会发展、经济建设方法和模式等。大数据的应用基础主要表现在各个学科基于大数据的创新，以科学研究的第四范式为代表[3]，包括对人文社会科学的研究、管理决策新方法、外部事件驱动的管理决策方法、基于微观数据的宏观经济学等。

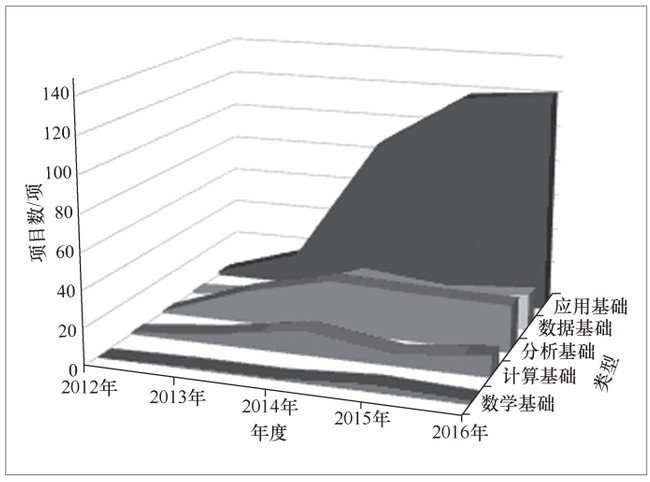


图 6 2012—2016 年在五大基础方面的项目数年度变化情况

GRAY J指出[3]：几千年前，科学研究是用实验解释自然现象的；几百年前，科学研究用理论模型探索科学规律，用实验验证理论；几十年前，科学研究用计算机模拟复杂现象，探索其中的奥秘；现在，科学研究是基于对数据的探索。科学的目的是认识宇宙、认识物质、认识生命、认识社会。

- 在认识宇宙方面：人们用了很多方法，早期科学家用肉眼观测天空，后来用望远镜，现在用射电望远镜。这些望远镜得到的结果是各种各样的宇宙图像，天文学家通过分析这些图像来研究宇宙。
- 在认识物质方面：人类还在研究物质的构成，为此于2008年建成并运行了欧洲强子对撞机装置。每一次正负电子的对撞，都产生了巨量的数据。科学家们经过不懈的努力，用了150个计算站点对200 PB数据用了3年时间分析，继而科学家发现上帝粒子[4]。
- 在认识生命方面：自从DNA被发现，人类对生命的认识进入了全新的阶段，人类似乎找到生命的本质、遗传的本质。DNA可以用A（腺嘌呤）、C（胞嘧啶）、G（鸟嘌呤）、T（胸腺嘧啶）4个字母的字符串表示，于是DNA变成了可以用计算机计算的数据，生命科学研究就出现计算生物学的分支，并且迅速发展。生命科学家开始分析数据，或者通过分析数据来研究生命。
- 在认识社会方面：信息化极大地推

动了社会的发展和进步，社会的抽样调查、问卷之类的研究方法将被淘汰，国家统计局已经和百度、阿里巴巴等数十家互联网公司合作进行数据收集和统计工作，共同推进大数据在政府统计中的应用，不断增强政府统计的科学性和及时性。

从上述分析可知，不论是自然科学还是社会科学，先进的研究方法是在数据上开展研究，这也说明，认识数据先于认识宇宙、认识物质、认识生命和认识社会。

大数据应用基础的主要研究方向如下。

- 各学科基于大数据的新方法、新范式、新理论等，包括生命科学、物理、化学、天文、历史、社会、管理、经济等学科的大数据方法和模型研究与探索。
- 各领域基于数据的新方法、新范式、新理论等，包括医疗、金融、交通、环保、商业等领域的大数据创新模式、智能决策方法和模型研究与探索。
- 用于支撑基于大数据的科学研究方法、社会发展方式、经济建设模式和国防安全手段。

国家自然科学基金委员会也已对上述各个研究方向开展资助，例如在生命科学的大数据方法研究项目有："利用大数据信息挖掘和基因进化方法研究禽流感病毒的跨地域传播""基于大数据整合挖掘的肾细胞癌分子进化机制研究"等；商业、交通、环保等领域的大数据方法研究项目有："大数据背景下的商业模式创新机制研究""大数据驱动的产品精确设计理论、方法及其应用研究""大数据环境下的复杂城市交通系统预测与控制""数据驱动的我国PM2.5污染规律模型智能构建方法研究""大数据驱动的我国典型重点流域水污染防控决策研究"以及"数据驱动的军事复杂系统风险决策分析方法及其应用研究""面向军事情报的多媒体大数据分析与展示"等项目。

4 大数据的分析基础

开发数据的核心是数据分析，也就是说大数据技术的核心是数据分析技术。目前，大数据分析技术主要在传统方法上延伸拓展，还没有从本质上解决大数据利用面临的挑战。这需要探索大数据分析技术的共性问题，主要研究方向如下。

（1）传统数据分析算法的改进原理

现有的大数据分析理论与方法大多从传统的统计分析、数据挖掘、机器学习、数据融合等领域派生出来，例如*K*-means++[5]、*K*-means Ⅱ[6]等聚类算法对经典*K*-means算法进行了改进，实现了大规模数据高效聚类。大数据的特点使现有方法超出了其使用条件和范围。因此，如何在拓展原有方法的基础上，研究适用于大数据特征的数据分析方法成为大数据时代的挑战，包括研究扩展传统的数据挖掘、机器学习、数据融合算法的原理。

（2）新型数据挖掘算法

大数据挖掘是从大数据中寻找其规律的技术[7]。大数据具有高价值、低密度的特性，"寻找"变得更具挑战性。分类分析需要有标签的训练集指导建模，但是大数据集中大多是没有经过专家打好标签的数据，需探索新的分类方法，以利用较少的有标签样本和较多的无标签样本进行学习。此外，面向高价值低密度的大数据集，存在这样一类数据挖掘需求：发现给定大数据集里面少数相似的数据对象组成的、表现出相异于大多数数据对象而形成异常的群组，被称为特异群组挖掘[8-10]。新的大数据挖掘方法研究包括特异群组挖掘方法、面向海量数据查询的相似性计算方法、大规模带时序可信知识图谱自动构建方法、动态大图分析方法等。

（3）高维数据分析方法

通过对大数据本身的压缩来适应有限存储和计算资源，除了研发计算能力更强、存储量更大的计算机之外，维规约技术（包括选维、降维、维度子空间等）是一类有效的方法，但也具备技术挑战。需要面向不同类型的数据研究语义保持下的大数据维规约技术（包括特征分析、特征选择、降维、子空间等），形成新的高维大数据分析方法和理论。

（4）深度学习方法

深度学习[11]已成为机器学习研究中的一个热点领域，通过构建具有多隐层的学习模型和大量训练数据学习特征。对深度学习的研究不仅仅聚集在视觉、图像、语音、文本等分析研究中，也需要拓展到时间序列、复杂网络等数据集上。深度学习模型优化、计算能力等方面仍然是当前深度学习的难点，需要形成理论体系。

国家自然科学基金委员会在大数据分析基础方面资助的项目有："大数据机器学习分布式算法的可行性理论""基于知识指导和模糊信息粒化的时序大数据分析和挖掘""RADIUS *K*-means算法及其拓展问题的研究""基于多源异构不确定数据的高效用信息挖掘的研究""面向图像序列的深度学习理论与方法""面向大数据的快速关联分析关键技术研究""面向大数据分析的自学习网络关键技术研究""基于认知计算的大数据挖掘理论与技术"项目等。

5 大数据的数据基础

数据是基础资源和战略资源，对资源竞争就会产生很多问题。国家之间的竞争涉及数据主权和数据安全，还有全球数据治理；一个国家内部涉及政府和公共数据资源的开放共享、交易流通，涉及数据权属和隐私安全保障，还涉及数据管理、数据质量、可视化等数据可用性问题。大数据研究需要有各种各样的数据环境，数据环境要与来源多样、类型多样的数据结合，并且数据规模要足够大。需要从数据质量保障、数据存储和数据索引等方面，研究大数据管理和治理的基础方法，包括以下几个方面。

（1）大数据治理

确保数据稀缺性不丧失和隐私不泄露是推动和实现数据开放共享的关键，有必要探索数据隐私保护机制及模型、大数据权属认定与保障理论及体系、区块链技术，构建数据自治开放理论体系，推动大数据交易。

（2）外部数据的质量保障机制

大数据来源广泛，也导致了数据的可靠性下降，如果不经过处理而将非确定的大数据用在后续分析中，必然会导致最终结果的不确定或不准确。因而要研究数据清洁的收益和代价计算，高效量化出单个数据清洗对全局结果判决质量的提升以及清洗单个数据耗费的开销，从而选择最高性价比的数据进行清洗。研究多数据源融合的清洗问题，梳理融合中多数据源之间的依赖关系，确定数据清洗的时机与策略，达到最优的清洗性价比。研究大数据可用性的定义与表达、大数据的错误传播与可用性相关模型，大数据可用性中的可计算与计算复杂性度量模型，基于群智计算的迭代数据清洗机制等基础理论和框架。

（3）大数据建模

关系模型、面向对象模型在以前的数据管理技术中发挥了核心作用。但原有的数学模型多是针对一种类型的数据，而大数据中包含结构化数据、半结构化数据和非结构化数据，因此需要研究相应的建模方法，将不同类型的数据从语义上关联在

一起，以复杂关联网络等技术为基础研究连接各种不同类型数据的数据描述机制，支撑对大数据的管理。

（4）大数据索引

传统索引结构常用于结构化数据库系统，能够提高小规模数据检索速度和查询表连接效率。然而，大数据环境下，传统索引结构存在冗余、存储空间过大、更新困难以及不适用于分布式存储环境等缺陷，这需要针对大数据的存储与数据特性研究大数据索引，包括非结构化数据索引结构、基于分布式存储的数据索引结构、高维与多目标需求下的数据索引结构等多种索引模型和索引性能评估模型。

（5）大数据可视化机理和方法

可视化是刻画数据之间关联性的有效手段，可视化分析贯穿于大数据全过程，如何实现可视化方法协助分析大规模、高维度、多来源、动态演化的数据，是当前可视化面临的挑战。因此，需要探索大数据全局视图可视化机理、时空相关的交互可视化机理、超大规模节点的高速数据渲染方法。

（6）知识图谱

知识图谱用于刻画实体或概念及其之间的关系，在大数据环境下，知识图谱更新和复杂性都急剧增加，为构建高质量知识图谱和实现有效推理，需要研究复杂知识图谱的语义描述方法、不确定知识图谱的构建与管理、基于知识图谱的多种类型数据表示模型、跨结构数据的存取机制和语义表示等。

国家自然科学基金委员会对数据基础研究方向的资助项目有："大数据协同计算及查询服务的隐私保护""大数据环境下的首席数据官、数据治理及组织绩效关系研究""高质量大数据集成关键技术的研究""大数据一致性错误管理理论与关键技术""大数据集背景下概念格的多粒度构造和存储研究""分布式不确定数据查询处理关键技术研究""面向大数据的信息可视化设计方法研究""高维大数据可视化的散度模型、算法及评价""基于外存的海量知识图谱数据的查询处理"等。

6 大数据的计算基础

大数据通常需要有能够分析处理这些数据的软硬件环境，即需要具有足够计算能力的计算环境。以深度学习为例，Hinton G E于2006年在《Science》上发表的论文《Reducing the dimensionality of data with neural networks》[11]提出了数据降维方法—— deep autoencoder，这成为深度学习开创性标志算法之一。然而，却并没有作为被广泛关注和使用的方法，而是随着数年后计算条件和计算能力的提升，深度学习方法才开始发挥更为重要的应用价值。这意味着需要探索新型高效能系统结构、性能导向的大数据计算框架和多地计算/异地计算理论与方法。

（1）新型高效能系统结构

当前，计算机系统的计算部件、存储部件、通信部件的功能和性能已朝着高速、高容量、高带宽的方向发展，并具有可编程、可定制等特点。如何利用这些部件构建新型高效能计算机系统满足大数据处理需求，是一个迫切需要解决的问题。这需要探索可重构、高度可配置的新型高效能系统结构；研究计算、存储和通信部件的有机结合、按需配置、弹性伸缩的方法；研究可变结构、软硬件结合的拟态计算机系统结构；研究高效能分布式存储系统的构建原理。

（2）性能导向的大数据计算框架

大数据的规模、计算时效性以及异构数据分布存储的特征，对计算机系统

的高通量、高时效和高并行提出了挑战。性能导向的并行计算框架是应对大数据挑战的关键和基础。这需要研究分析大数据应用的计算特征、通信特征和存储特征；研究并行计算系统的高通量、高时效计算技术，包括实时分布式内存系统、内存计算系统、异构多核平台的性能加速技术等；研究性能可预测的并行计算模型。

（3）多地计算/异地计算理论与方法

数据大的难以移动、数据重要的不愿移动，在此背景下，如何求解一个全局问题是一大挑战。通常在大数据所属地计算局部解，即大数据应用具有数据存储的分布性问题，在数据所在地进行计算，产生的部分计算结果可能出现不一致、相互背离等现象，需要通过不同方式的计算进行验证，这给求解全局问题带来挑战。因此，需要研究大数据多地计算/异地计算基础问题，包括异地计算行为建模；研究局部解的局限性评估机制、局部解发送接收的身份验证机制、局部解的优化融合策略；研究全局解的最优性评估机制、提高全局解最优率的异地选取策略等。

国家自然科学基金委员会对上述各个研究方向开展资助的项目有："云计算环境下大数据本地化技术研究""结合逻辑与物理I/O访问信息的存储系统优化策略的研究""互联云环境下面向数据中心的服务资源分配与调度机制研究""面向大数据的高时效并行计算机系统结构与技术""大数据高效能存储与管理方法研究""面向大数据的计算与存储融合CPU体系结构研究""分布式计算与系统""面向大数据内存计算的计算机体系结构""大数据环境下增量式迭代计算框架的研究与优化""云计算环境下面向大数据的在线聚集并行优化机制研究"等。

7 大数据的数学基础

现有的大数据研究大多关注将传统方法应用于大数据的实际场景，而没有从大数据的内在结构、本质特征出发，这制约了大数据研究的深度发展，使得大数据技术面临瓶颈。因此，研究大数据需要从数据本身出发，这离不开数学方法的支持。大数据的表示和度量是大数据研究的关键和基础，相应地，与之密切相关的大数据代数系统、大数据内在数学结构和大数据相似性度量成为大数据的数学基础的重要内容。

（1）大数据的代数系统

关系代数为关系型计算提供理论依据。然而，高扩展性是大数据分析的重要需求，传统的关系数据模型难以胜任当前存在的非结构化数据（如文本数据、序列数据、流式数据等）的处理。近年来，已出现一些非关系型数据库（如HBase、MongoDB等），在非结构化数据上的复杂数据分析能力有所提高，并得到广泛应用。但是，目前缺少对非关系型数据库的数据代数的研究。对于非关系型数据，定义由数据集构成的集合上的度量方法和运算，形成一定论域上的数据代数等，这些都将在数学基础上对非关系型数据提供理论支持，有望突破现有技术瓶颈。

（2）大数据内在数学结构

数据有复杂的拓扑、网络等不同结构，在大数据问题中，数据本身往往具有更为复杂的内在数学结构，例如，高维数据空间中因为具有一定的约束条件而具有流形的数据结构；又如，在图像等非结构化数据中，先天性地具有低秩的数学性质。在深刻理解和挖掘内在相应结构的基础上，才能有效建立分析模型。针对大数

据集的流形或复形等复杂数学结构和稀疏、低秩等数学性质，设计合理描述的数据结构，构建相应的度量，选取多尺度自适应的基底表示，为构建分析模型、形成反映内在结构参数的分析算法提供理论支撑，并通过数学结构的性质，保证算法的适用性。

（3）大数据的相似性度量

相似性是数据挖掘分析任务的核心。简单数据类型的相似性度量支撑传统数据分析模型，然而，针对复杂数据类型，这些相似性度量难以真实反映数据之间的关系。针对大数据复杂性特征，定义空间非刚性结构的相似性度量和超高维、多类型的大数据相似性度量，发展非线性降维方法、核理论以及相应的高效算法和稳定性分析。

如前所述，大数据研究过程中数学基础方面还需要更多的研究者参与和关注，目前，国家自然科学基金用在大数据的数学基础方面的相对较少，但也有对上述研究方向的涉及，例如："大数据环境下服务的存储范式理论及操作代数研究""大数据中的广义稀疏几何结构学习方法研究""适应大数据处理的概率相似度学习研究"等。

8 结束语

数据作为一种基础性资源已经得到广泛认可，大数据开始在各行各业发挥作用，同时也遇到了很多难题，例如数据质量、数据权属、数据开放共享、数据快速分析、异地数据计算等，这些问题可以统称为大数据问题。由于大数据的广泛性，大数据问题是多层次、多样的，需要各领域的科学家共同努力。本文将大数据的基础问题归类为应用基础、分析基础、数据基础、计算基础和数学基础五大类别。特别地，将各学科基于大数据的创新研究也归类为大数据的应用基础。这是因为科学研究进入了所谓的第四范式，第四范式的科学研究需要先行积累数据、研究数据、分析数据、观察数据，然后才进行业务研究，或者说，认识宇宙、认识物质、认识生命、认识社会，需要先行认识相应的数据，即认识数据先于认识宇宙、物质、生命和社会。建议国家层面研究基金全面部署，探究大数据各层面的基础问题，实现大数据原始创新，推动数据科学发展，有利于形成大数据的核心竞争力。

参考文献：

[1] 朱扬勇，熊赟. 大数据是数据、技术，还是应用[J]. 大数据，2015007.
ZHU Y Y, XIONG Y. Defining big data[J]. Big Data Research, 2015007.

[2] MOORE G E. The microprocessor:engine of the technology revolution[J]. Communications of the ACM, 1997, 40(2): 112.

[3] HEY T, STEWART T, KRISTIN T. The forth paradigm: data-intensive scientific discovery[M]. Beijing: Microsoft Research, 2009.

[4] CARMI D, FALKOWSKI A, KUFLIK E, et al. Higgs after the discovery: a status report[J]. Journal of High Energy Physics, 2012, arXiv:1207.1718.

[5] BAHMANI B, MOSELEY B, VATTANI A, et al. Scalable k-means++[J]. Proceedings of the VLDB Endowment, 2012,5(7): 622-633.

[6] ARTHUR D, VASSILVITSKII S. K-means++: the advantages of careful seeding[C]// 18th ACM-SIAM Symposium on Discrete Algorithms, January 7 - 9, 2007, New Orleans, Louisiana, USA. New York: ACM Press, 2007: 1027-1035.

[7] 熊赟，朱扬勇，陈志渊. 大数据挖掘[M]. 上海：上海科学技术出版社，2016.
XIONG Y, ZHU Y Y, CHEN Z Y. Big data

mining[M]. Shanghai: Shanghai Scientific & Technical Publishers, 2016.

[8] 熊赟, 朱扬勇. 特异群组挖掘: 框架与应用[J]. 大数据, 2015020.
XIONG Y, ZHU Y Y. Abnormal group mining: framework and applications[J]. Big Data Research, 2015020.

[9] XIONG Y, ZHU Y Y, YU P S, et al. Towards cohesive anomaly mining[C]// 27th AAAI Conference on Artificial Intelligence (AAAI), July 14 - 18, 2013, Bellevue, Washington, USA. San Francisco: AAAI Press, 2013: 984-990.

[10] XIONG Y, ZHU Y Y. Mining peculiarity groups in day-by-day behavioral datasets[C]//IEEE International Conference on Data Mining (ICDM), December 6 - 9, 2009, Miami, Florida, USA. New Jersey: IEEE Press, 2009: 578-587.

[11] HINTON G E, SALAKHUDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

作者简介

朱扬勇(1963-), 男, 博士, 复旦大学计算机科学技术学院教授、学术委员会主任, 上海市数据科学重点实验室主任。1989年起从事数据领域研究, 2008年提出数据资源保护和利用, 2009年发表了数据科学论文《Data explosion, data nature and dataology》, 并出版专著《数据学》, 对数据科学进行了系统探讨和描述。2010年创办了“International Workshop on Dataology and Data Science”, 2014年和石勇、张成奇共同创办了“International Conference on Data Science”。担任第462次香山科学会议“数据科学与大数据的理论问题探索”的执行主席、《大数据技术与应用丛书》主编。主要研究方向为数据科学、大数据。

熊赟(1980-), 女, 博士, 复旦大学计算机科学技术学院教授。2004年起从事数据领域方面的研究工作, 作为项目负责人主持国家自然科学基金、上海市科委发展基金以及企业合作项目。相关研究成果在本领域国际权威期刊或会议发表论文40余篇, 出版著作3本。主要研究方向为数据科学、大数据。

收稿日期: 2017-01-06
基金项目: 国家自然科学基金资助项目(No. 91546105); 上海市科技发展基金资助项目(No. 16JC1400801)
Foundation Items: The National Natural Science Foundation of China (No. 91546105), Shanghai Science and Technology Development Fund (No. 16JC1400801)