网络信息空间的大数据计算

特邀编辑：胡春明　马　帅　怀进鹏
北京航空航天大学

关键词：网络信息空间　大数据计算

信息社会的快速发展引发了数据规模的爆炸式增长，使网络信息空间大数据成为继人力、资本之后一种新的非物质生产要素，甚至被认为是关系国家经济发展、社会安全和科技进步的重要战略资源，蕴含巨大价值。

自2012年美国宣布投入2亿美元启动“大数据研发计划”(Big Data R&D Initiative)以来，英国、欧盟、澳大利亚、日本等发达国家和地区均提出了国家层面的大数据发展规划。我国也充分认识到大数据时代带来的重大机遇，2012年3月科技部发布的《“十二五”国家科技计划信息技术领域项目》就将大数据研究列在首位。2015年9月，国务院印发《促进大数据发展行动纲要》，系统部署大数据发展工作，首次明确提出建设数据强国。2016年发布的《中华人民共和国国民经济和社会发展第十三个五年规划纲要》提出“实施国家大数据战略”，把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新。今天，大数据已经成为推动行业生产效率提升、促进企业和社会管理变革的利器，并形成规模巨大的产业生态。到2020年，我国大数据相关产品和服务业务收入预计将突破1万亿元，年均复合增长率保持30%[1]。同时，也将带动IT硬件、云计算、数据服务等相关产业的发展，并对“互联网+”“中国制造2025”等国家战略起到促进作用。

通常认为大数据具有“4V”特征，即规模庞大(Volume)、种类繁多(Variety)、变化频繁(Velocity)、价值巨大但价值密度低(Value)。这些特征对发现事实、揭示规律并预测未来提出了新的挑战，并将对已有计算模式、理论和方法产生深远的影响。首先，网络信息空间大数据数量庞大，数据的统计特征分布不均匀。在传统的采样方法中，样本选取的差异在减少计算量的同时可能会引入结果的不确定性，采样的质量和精确性都会对计算结果产生影响。但是，在大数据的计算中，对单一数据项和分析算法的精确性要求不再苛刻，通过对大量数据的分析处理能够有效弥补传统抽样方法的局限。其次，大数据种类繁多，变化频繁。已有的计算模式往往通过预先确定的分类方法降低问题的难度和规模，提高预测的准确性。而在大数据计算中，数据的持续更新可能难以形成稳定的分类，不仅要考虑可分类条件下的精确算法，还要考虑动态数据下的增量算法。最后，大数据研究不同于传统的逻辑推理研究，是对巨大的数据做统计分析和归纳。传统的确定性问题往往通过自顶向下的还原方法，逐步分解并加以研究，而对多源异构大数据相关问题的研究不仅需要还原方法，还需要自底向上的归纳方法，通过关联关系补充因果关系的不足，实现多源数据和多种计算方法的有效融合。

综上所述，大数据计算具有“近似处理、增量计算、多源归纳”的计算属性，并可进一步归纳为

[1] 数据来自工信部印发的《大数据产业发展规划(2016 – 2020年)》。

大数据计算的**“3I”特征**，即近似性 (**I**nexact)、增量性 (**I**ncremental) 和归纳性 (**I**nductive)，分别在数据层面、算法层面和系统层面给大数据计算带来了“可表示”“可计算”和“可操作”三大问题。

近似性：网络信息空间大数据计算通常面对近乎全量的大数据集，传统计算复杂性理论认为的易解问题在大数据时代下已成为难解问题。由于数据本身的异构和噪声，很难按照传统精确处理的思路来进行大数据的挖掘。此外，许多应用需求旨在寻找数据间的潜在关联关系和宏观趋势特征，允许解的质量在一定区间内近似。例如，在微博突发事件分析与预警中，突发事件本身会受到普遍而强烈的噪声数据干扰，热点事件及宏观态势的判断也有很强的时效性要求，以时间消耗为代价的精确计算不再适用。**因此，从数据层面，需要综合考虑数据的语义特征、结构特征与质量特征，理解并量化度量数据的价值分布；从算法理论层面，需要建立大数据下的算法复杂性理论及近似算法理论，识别数据量对算法质量的关联关系；从系统层面，需要设计满足用户需求的非精确计算架构，达到用户需求与计算效能的均衡。**

增量性：网络信息空间大数据动态持续产生，不断更新，很难形成大数据的统一视图。此外，许多大数据处理对实时性要求越来越高，全量式的批处理和迭代处理方式在时间上难以满足需求，增量式处理成为一种解决办法。例如，百度智能搜索涉及近万亿的网页，大量网页频繁更新，在构建搜索索引和获取用户查询结果时，很难及时对近千PB(1PB=10^{15}B) 网页数据进行全量计算；再如突发事件预警需要业务用户对数据进行长期、频繁地探索，并根据不断更新的结果对数据源、分析方法和计算过程等要素进行调整，以获得更准确及时的结果。**因此，从数据层面，需要量化度量数据的动态复杂性；从算法理论层面，需要考虑数据动态性及其对解的质量的影响，并设计增量式处理算法；从系统层面，需要设计支持增量计算的存储和处理架构及相关机制。**

归纳性：大数据的多源异构特征对网络信息空间数据挖掘提出新挑战并带来机遇。寻找同一实体在多源数据之间的潜在关联性，有助于进一步规避数据中的噪声干扰，并通过多源数据处理的智能归纳融合，修正非精确数据处理引入的偏差，同时获得比单一数据源更好的处理效果。例如，百度根据用户的搜索日志及其在“百度贴吧”和“百度知道”等不同产品线中提交的数据进行归纳融合，建立用户行为模型，可提供更为准确的个性化搜索结果。**因此，从数据层面，一方面要研究多源异构数据的表示、度量与语义理解方法，努力减少多源异构数据带来的难题，另一方面需要关注多源数据间的潜在关联性和融合方法；从算法层面，需要寻找新的多源数据处理和智能归纳融合算法，并提高算法精度及效率；从系统层面，需要研究多源数据间可迁移学习的数据挖掘新方法，探索融合机器挖掘和人群分析的多种数据处理机制。**

本期专题主要针对网络信息空间大数据计算在数据层面的“可表示”问题，算法理论层面的“可计算”问题，系统层面的“可操作”问题，特别邀请了相关领域专家学者对大数据计算的最新研究进展和发展趋势进行了深入探讨。

华为诺亚方舟实验室的尚利峰博士等撰写的《自然语言对话关键技术及系统》一文，针对数据层面的“可表示”问题，回顾了自然语言对话系统近二三十年来的研究工作：从传统的基于概率决策过程的多轮对话系统，到近年来提出的基于深度学习的生成式对话系统，再到将深度学习和符号处理相融合的神经符号对话系统。自然语言对话系统研究的目的是希望机器人能够理解人类的自然语言，同时进行个性化的情感表达、知识推理、信息汇总。作者指出更好地进行自然语言理解、知识表示和推理是整个对话系统发展的核心推动力。

北京航空航天大学、英国爱丁堡大学的樊文飞教授等撰写的《资源有限下的大数据有界查询处理》一文，针对算法理论层面的“可计算”问题，回答了如何描述大数据计算的复杂性，什么是大数据计算中的易解问题；对于非易解问题，如何在资源受

ADL

CCF 学科前沿讲习班

第95期

大数据环境下内存存储与内存计算

2018年9月17~18日 北京

学术主任兼讲者

舒继武

清华大学

廖小飞

华中科技大学

特邀讲者

李　涛　佛罗里达大学

吴结生　阿里巴巴云计算公司

肖　侬　国防科技大学

周礼栋　微软亚洲研究院

陈海波　上海交通大学

查　伟　华为公司

限的情况下设计和实现高效算法。文章介绍了在大数据易解性复杂性理论、资源受限下的大数据查询评估的理论模型、算法与框架大数据复杂度建模理论和计算模型等方面的最新研究成果。

北京航空航天大学李昂生教授撰写的《结构信息度量》一文，针对数据层面的“可表示”问题和算法理论层面的“可计算”问题，阐述了大数据处理的根本任务就是从大规模噪音结构中解码出规律，并提出了一个新理论——结构信息论。该理论可以度量嵌入在复杂系统中的信息，区分网络空间大数据的规律与噪音，解码出嵌入在大规模噪音结构中的规律，从而解决对高效算法的高精度定义这一难题。

国防科技大学张一鸣教授等撰写的《面向存算联动的大规模网络内存存储系统》一文，针对系统层面的“可操作”问题，根据网络搜索、电子商务、社交网络等应用的存算联动特点，以及低延迟、高带宽、高吞吐率等需求，介绍了面向在线数据密集型(Online Data Intensive, OLDI)应用的大规模内存存储系统的研究进展。

香港科技大学杨强教授等撰写的《迁移学习：回顾与进展》一文，针对系统层面的“可操作”问题，介绍了迁移学习的起源和现状，梳理迁移学习技术的发展脉络，及其在研究领域的应用，呈现出迁移学习完整的发展历程。作者通过不同的实例介绍了迁移学习如何“举一反三”，通过将某个领域的知识迁移到另一个领域的学习中，打破计算资源和数据资源的桎梏。

以上五篇文章从模型到算法、从原理到系统介绍了如何解决大数据计算的“可表示”“可计算”和“可操作”三个问题，提出了许多新的理念和研究方向，希望能鼓舞更多的学者参与到大数据计算的基础理论、算法、系统和应用研究中，开拓新的研究方向和领域。 ■

致谢：

此专题内容得到了国家重点基础研究发展计划（973计划）项目（2014CB340300）的支持。

胡春明

CCF专业会员、CCCF译文编委。北京航空航天大学副教授。主要研究方向为计算机软件与理论、分布式系统、数据中心资源管理与调度。

hucm@buaa.edu.cn

马　帅

CCF高级会员。北京航空航天大学教授。主要研究方向为大数据、数据库理论与系统等。

mashuai@buaa.edu.cn

怀进鹏

中国科学院院士，北京航空航天大学计算机学院教授。长期从事计算机软件与理论方向研究工作。

huaijp@buaa.edu.cn