

大数据战略：从数据大国到数据强国

王建民

清华大学软件学院,大数据系统软件国家工程实验室,工业大数据系统与应用北京市重点实验室,
北京 100084

大数据时代,数据成为与物质资产和人力资源同样重要的基础生产要素。海量数据的空前集聚和计算力的指数级增长释放了深度学习算法的生产力,推动人工智能应用的大发展。大数据技术和大数据产业的蓬勃发展为线上互联网和线下实体经济带来了新的增长点。

为鼓励大数据领域的技术创新与应用创新,国务院于2015年印发《促进大数据发展行动纲要》,鼓励加快经济社会各领域的大数据开发与利用,催生出更多的新产业、新业态、新模式,推动国家、行业、企业在数据的应用需求和发展水平方面进入新的阶段。在内部技术条件成熟、外部政策因素推动的激励下,中国涌现出一批从传统业务扩展并转型到大数据业务的企业,通过对企业内部生产数据、客户、用户数据的分析,帮助企业实现智能决策,提高运行效率和风险管理能力。新应用新模式层出不穷,大数据产业呈现出蓬勃发展的态势。

毋庸置疑,大数据已成为当代科技前沿热点,是重要的基础性战略资源,其发展为科学研究范式带来了深刻改变,为人类世界创造了更多可能。以此为契机,《科技导报》策划出版“大数据战略:从数据大国到数据强国”专题,围绕以清华大学大数据研究中心为代表的大数据创新研究平台,总结在大数据基础理论、核心技术与系统、关键领域应用层面取得的研究成果和最新趋势,探讨大数据技术与各行业领域的结合方式,指出存在问题和发展建议。

本专题是对大数据技术研究与发展不同主题的介绍,包括:(1)从理论基础的角度,提出了一个

大数据系统的代数模型和自顶向下的分析方法,是对大数据系统复杂性和行为的高度抽象和全新认知视角;(2)提出了一组大数据系统数学性质的定义,探讨了大数据的认知基础及其分类模型,给出了科学、工程和社交领域的大数据表示模式;(3)概述了大数据产品的发展脉络,讨论大数据产品研发的机遇和挑战;(4)分析了大数据安全的内涵、关键技术及保障机制,指出了大数据产业面临的安全挑战与风险;(5)分析了数据可视化的通用需求和关键研究问题,给出了一个交互式数据可视化工具的设计方案;(6)介绍了Aache开源社区中,大数据管理系统的关键技术和系统优缺点。

数据作为计算的处理对象,与软件密不可分。数据的获取与清洗、集成与分析、呈现与应用等环节都离不开软件的支持。在大数据时代,系统与软件工程面临2个主要挑战:系统与软件工程要针对大数据处理与分析需求,开发支持大数据处理各个环节的软件技术与系统;系统与软件工程实施过程中,会涉及大量具有大数据特征的系统运行过程数据,要对这些多维数据进行充分关联挖掘和机器学习,发现数据驱动的软件开发和运行规律,形成基于大数据的软件方法学。大数据应用系统覆盖数据的获取、清洗、集成、分析与可视化等大数据全生命周期的多个处理环节,而每个环节都存在着多款软件工具,它们以开源软件构件形式在大数据生态系统中“野蛮生长”,给面向领域的大数据应用系统构建、运行与优化带来了挑战。

(责任编辑 刘志远)