大数据技术专题

序言：大数据及其应用方兴未艾

近年来，随着物联网、云计算、移动互联网、车联网等技术的成熟和迅速普及，人类社会正在以更快的速度产生不同类别如图像、视音频、健康档案等海量数据。据国际数据公司 IDC 预计，到 2025 年全球数据量将达到 175 ZB（约 1 750 亿 TB），这意味着在人类文明的所有数据中，超过 99% 是近几年产生的。毫无疑问，历经机械时代、信息时代后，人类正步入一个崭新且充满挑战的新时期——大数据智能时代。大数据是新时代的宝贵资源，但并非数据量大即可称为大数据，一般来说，大数据具有 4V 特征，即 Volume（量大）、Velocity（高速）、Variety（多样）和 Veracity（真实）。利用这些信息资产，通过新型的处理技术挖掘隐藏的丰富信息，从而促成更强的决策能力、洞察力与最优化处理，是大数据技术的本质与出发点。具体而言，大数据技术包含数据获取与预处理、数据存储、计算分析挖掘及实际应用等多个方面。作为近年的热门方向，无论工业界还是学术界均对大数据技术的研究方兴未艾，Google 的 MapReduce 编程模型、Facebook 的 Hive、伯克利 AMPLab 的 Apache Spark 框架都是典型的大数据技术。本刊从宽广的范围内组织了一期大数据技术专题，报道国内学者在大数据技术方面的研究成果，包括大数据存储、数据挖掘算法、大数据平台、视觉大数据处理芯片体系结构和超高分辨率图像大数据处理框架等方面。

在数据存储领域，深信服科技有限公司和中国科学院深圳先进技术研究院（以下简称“中科院深圳先进院”）异构智能中心提出了一种基于 Glusterfs 的端到端校验方案，来解决 Glusterfs 文件系统中存在的数据完整性风险。在

数据处理领域，针对数据降维问题，中科院深圳先进院医学信息中心李烨课题组提出一种非负子空间聚类算法来发掘数据的子空间结构信息；而云南师范大学杨昆课题组则通过计算 IC 卡和手机两种不同设备历史轨迹的时空相似性，重现乘客在地铁网络里的完整轨迹。在大数据平台构建方面，上海交通大学李超课题组和卫宁健康科技股份有限公司人工智能实验室利用大数据技术构建一个面向医疗临床科研的大数据平台，解决电子病历临床数据结构多样、信息基础设施不一致的问题。在生物医学领域，结合网络嵌入算法，中科院深圳先进院医学信息中心蔡云鹏课题组利用肠道微生物组学大数据挖掘了潜在的致病基因并探究了肠道菌群分布的关联模式。殷鹏课题组则提出了一种全新的基于多步筛选法的全基因组关联分析方法，优化了关联 SNP（每个单核苷酸多态性）位点的挖掘。在数据处理框架方面，国防科技大学沈立课题组成功研制出高效的多模型图像超分辨率框架。在更为基础的大数据处理芯片体系结构方面，国防科技大学王蕾课题组研究了用于加速视觉识别的硬件架构。

目前，大数据技术已广泛应用于多个领域并取得了骄人成绩，如商业、医疗、教育、农业等，为社会的发展注入了全新的活力。但仍然存在由数据、计算架构与系统复杂性所造成的各种挑战与困难。同时，在国家层面，大数据是重要战略资产，是未来的新石油，拥有数据的规模和运用数据的能力是衡量综合国力的重要因素。特别在国际局势日趋复杂，贸易争端持续升级的今天，只有把握住大数据智能时代的发展脉络，大力发展新一代大数据、人工智能技术，掌握数据信息领域的主动权，中国才能实现弯道超车，彻底突破美国施加的桎梏。

喻之斌 研究员　殷鹏 副研究员
中国科学院深圳先进技术研究院
2019 年 9 月