大数据软件的机遇与挑战

孙家广,中国工程院院士,清华大学教授,现任大数据系统软件国家工程实验室主任、中国图学学会理事长。主要研究方向为计算机图形学、计算机辅助设计、软件系统及工程。

2019年,大数据、数据科学、机器学习、人工智能领域的研究与应用持续发展。物联网与传感设备的普及带来数据的爆炸性增长。大数据作为产业发展的创新要素,在数据科学与技术、商业模式、产业格局、生态价值与教育层面,均带来了新理念和新思维。大数据与人工智能的快速普及应用除了受数据量激增因素影响外,还有另外两方面因素影响:一是深度神经网络算法处理大规模非结构化数据集的能力越来越强;二是算力的飞跃。随着光刻技术进一步发展,终端设备和边缘设备的数据处理能力持续提高,云、端与边缘计算结合,实现低成本海量可用计算资源。

2019年大数据与人工智能生态圈中,最活跃的领域包括:大数据基础设施服务、大数据分析服务、数据资源服务、数据源管理及API服务、跨平台数据存储及分析服务,开源大数据软件工具,以及在各个垂直领域的产业大数据应用和企业大数据应用(图1)。大数据、人工智能与产业深度融合,在交通运输、电子商务、金融服务、医疗健康、科学研究等领域展现出广阔的应用前景。"数字化转型"是大数据技术应用的驱动力,是要让企业真正成为"数据驱动"的企业,使得企业生产更加绿色、智能。大数据已经逐渐成为企业升级转型发展的有力引擎,在提升产业竞争力和推动商业模式创新方面发挥越来越重要的作用。

同时,大数据也开辟了国家治理的新路径,国家社会管理现代化面临着由碎片型向整体型、由应急型向预防型、由管控型向参与型、由粗放型向精细型、由静态型向动态型转变的"五位一体"的全面变革。物联网推动互联网应用从消费领域向生产领域扩展,并逐步深入城市管理各个环节。通过对海量、动态、高增长、多元化、多样化数据的高速处理,人们快速获得有价值信息,提高公共决策能力,从而逐步改变国家治理架构和模式。

图1　大数据与人工智能生态圈顶层分类

目前最重要的大数据技术领域主要包括以下4个方面。

1)生态系统的建设。提及大数据,就无法避免提及Apache Hadoop。多年来,Hadoop已经发展到包含整个相关软件生态系统,许多商业大数据解决方案都基于Hadoop,基于Hadoop的产品和服务市场持续增长;其次,大数据处理引擎的研发,Apache Spark是Hadoop生态的重要组成部分,已经在生产环境中广泛部署,也吸引了大量的项目开发者;此外,处理和统计数据的编程语言和软件环境,例如开源项目R语言得到数据科学家的广泛应用,许多流行的集成开发环境(IDE),包括Eclipse和Visual Studio,都支持R语言,R已经成为世界上最流行的用于大数据项目的高级语言之一。

2)海量数据存储方案,例如数据湖(data lake)。许多企业正在建立数据湖(存储来自许多不同的数据源的数据并按原态

存储),当企业想要存储数据但尚不确定如何使用数据时,数据湖尤其具有吸引力。物联网(IoT)数据的爆发正在影响数据湖应用的增长。

3)NoSQL数据库的发展。为适应非结构化数据的存储与高性能需求,以及相对不那么严苛的数据一致性的要求,MongoDB、Redis、Cassandra、Couchbase等NoSQL数据库流行。随着大数据趋势的增长,NoSQL数据库变得越来越流行。

4)数据的预测分析。预测分析是大数据分析的子集,是根据历史数据预测未来事件或行为。通过数据挖掘、建模和机器学习技术,获取对未来趋势的洞察。

在大数据时代,机遇与挑战并存。大数据技术研究者在迎接数据与智能技术带来无限可能的同时,也不得不面对其所蕴藏的风险。随着公民个人和企业组织所有的行为均被数字化,海量数据的实时处理与分析技术更加成熟,大数据在带来奇迹的同时也引入滥用和误用的风险。大数据安全保护技术与数据权责管理成为大数据领域最重要的主题,任何组织都无法回避谁拥有影响未来的数据权的问题。

互联网的早期阶段,数据隐私更多是要保护用户在线行为的隐私,这只占人民日常生活的一小部分,因此得到的关注是非常有限的。随着个人生活和工作的全部活动都通过网络和互联设备来完成,海量数据融合的能力、人脸识别的能力、结果预测的能力、异常分析的能力整合在一起将带来严重的数据隐私风险。

孙家广

(清华大学大数据系统软件国家工程实验室,北京 100084)