# 大数据分析系统创新平台与生态建设

**王元卓[1,2]，靳小龙[1,2]，程学旗[1,2]**
1. 中国科学院计算技术研究所，北京 100190；2. 大数据分析系统国家工程实验室，北京 100190

**摘要**
针对大数据分析处理深度与智能化程度不足，大数据分析计算不贯通、数据模型参数复杂，大数据算法易用性差、分析处理链路缺失以及大数据发挥协同效应条件要求高等共性技术与工程化瓶颈，系统地介绍了大数据分析系统国家工程验室研制的大数据分析系统创新平台的技术架构。该平台将重点支持科学计算、政府治理和社会安全等方面的示范应用，通过开放生态向各领域与全行业辐射，推动创新创业发展。

**关键词**
大数据分析系统；大数据开放生态；国家工程实验室

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018010

# *Innovation platform and ecology construction of big data analysis system*

**WANG Yuanzhuo[1,2], JIN Xiaolong[1,2], CHENG Xueqi[1,2]**
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
2. National Engineering Laboratory for Big Data Analysis System, Beijing 100190, China

***Abstract***
In view of bottlenecks in common technology and engineering faced by big data analysis in our country, such as low level of processing depth and intelligence degree in big data analysis, incoherence of big data analysis and calculation, complexity of model parameters, poor usability of big data algorithms, deficiency of analysis process, and high requirements of synergistic effect played by big data, the innovation platform of big data analysis system developed by the National Engineering Laboratory for Big Data Analysis System (BDASlab) was systematically introduced. The BDASlab will focus on supporting demonstration applications of scientific computing, government governance, social security, and other aspects. The BDASlab will radiate various fields and entire industry through an open ecological-system way, promoting development of innovation and entrepreneurship.

***Key words***
big data analysis system, big data open ecology, National Engineering Laboratory

## 1 引言

近几年，科技界和企业界甚至世界各国政府都把大数据作为发展的重点。美国政府更是把大数据看作“未来的新石油”，拥有数据的规模和运用数据的能力将成为衡量一个国家综合国力的重要组成部分，国家间和企业间对数据的占有和控制将成为新的争夺焦点。大数据分析技术是使用大数据的核心关键技术，对数据的深度分析和利用将在推动经济持续增长、提升国家和企业竞争力方面起到至关重要的作用。大数据将成为未来新的经济增长点，企业将向分析即服务升级转型，并进一步改变行业的生态。在这样的背景下，IBM、Oracle、Google、Microsoft等全球IT界巨头都已开始了在大数据方向的技术布局。在国家层面，大数据时代对数据的积累、处理和价值利用能力将成为一个国家国力的新标志[1,2]。

目前，对大数据的分析与利用已经给国外各行各业带来了巨大的变革性的机会，美国麦肯锡咨询机构发布的一份关于大数据的研究报告中指出，大数据已经对金融业、电信行业、政府公共管理、健康医疗等行业产生了重大的影响。其中，在公共管理领域，每年产生约2 500亿美元的潜在价值。在政府公共管理领域，英国通过大数据分析技术的使用，优化政府部门的日常运行，刺激公共机构的生产力，为英国政府节省了130亿~220亿英镑；减少福利系统中的诈骗行为和错误数据量，为英国政府节省了20亿~80亿英镑。在电信行业，发达国家的电信运营商（如法国电信、英国O2公司和沃达丰）针对用户消费的大数据进行分析评估，借助大数据改善服务水平，提升了用户体验。当前，大数据分析已经成为全球信息技术产业发展的重点，对大数据的分析利用不断地深入各国各行业的创新应用，推动了传统产业新的发展模式，促进了国民社会的发展和经济的提升。

在我国，各行各业也都在积极布局大数据的技术研发与产业应用。以阿里巴巴集团、百度公司、腾讯公司为代表的互联网与IT企业依靠自身的数据优势，将大数据作为公司的重要战略，引领我国大数据技术的快速发展。阿里巴巴集团在其拥有的大规模数据和云平台基础上，向移动开发者提供技术、数据、平台等基础设施服务，帮助开发者实现数据的精准分析挖掘与个性化推送。“百度网盟推广”基于大规模日志数据的学习与预测，大幅提高网站的客流量和广告收入。中国移动通信集团有限公司提出了大数据时代全新的移动互联网战略，包括构筑“智能管道”、搭建“开放平台”、打造“特色业务”与提供“友好界面”，构建了大云产业联盟，在核心模块合作、授权技术服务、应用开发技术攻关等不同层面展开产业合作。京东集团优先开放“电商云”和“数据云”，并进行人工智能和机器学习的方法研究，从海量数据中挖掘商业价值，提高了商业运营效率。

## 2 大数据分析的共性技术与工程化瓶颈

尽管各行各业都在积极开展大数据分析的技术创新与布局，但我国大数据分析系统的发展应用仍然面临众多挑战[3-5]。

（1）大数据分析处理的深度与智能化程度不足，导致分析程度浅，利用率低

随着互联网时代的到来，数据产生的途径也在不断增加，数据类型的数量也在持续增长，这就给传统的数据分析与挖掘技术（如检索、主题发现、语义和情感分析等）带来很大的困难。传统的数据处理方法面

临着不同数据类型有机融合的新挑战。与此同时，随着数据生成方式愈来愈多样化，半结构化与非结构化数据逐渐成为大数据的主流数据源，大规模的非结构化数据同样也给传统的数据处理方法带来了巨大的挑战。另外，随着数据规模的增大，需要刻画的数据特征数量必然在增加，而数据特征组成的数据内在模式则会呈指数性增长。数据类型的多样化导致了数据模式的复杂化、多样化，这使得网络数据通常蕴含更多的无用信息和噪音，因此需要高效、顽健的方法实现去冗存精。一般来说，大数据有维度高的特征，这就会引发数据高度稀疏以及“维度灾难”等问题。

（2）大数据分析计算不贯通，数据模型参数复杂，导致处理效率低，时效性差

传统数据计算模式缺乏对数据全生命周期的计算，使得很多数据分析的任务无法深入、融合计算。大数据分析面临模型复杂度高、任务周期长、实时性要求强等难题。对大数据分析系统的系统架构、计算框架、处理方法提出了新的挑战，更对大数据处理系统的运行效率提出了要求。

（3）大数据算法易用性差、分析处理链路缺失，导致使用门槛高，工程化难

大数据技术起源于互联网行业，目前最成功的应用也在互联网行业，在其他行业的应用还处于初级阶段。我国的各级政府机关和各类传统行业在日常管理和业务运行中也积累了大量的数据。大数据的真正价值所在是深度价值发现和行业应用，如何推进大数据技术应用，唤醒这些沉睡的大数据资源，实现管理上的科学决策，开创新的业务模式，是这些数据拥有者关心的问题。

（4）大数据发挥协同效应条件要求高，需要产业链各个环节的企业达成竞争与合作的平衡

大数据对产业链中的企业提出了更多的合作要求。如果没有对整体产业链的宏观把控，单个企业仅仅基于自己掌握的独立数据无法探究产业链各个环节数据之间的关系，因此对消费者的刻画和判断会十分有限，甚至片面。

## 3 大数据分析系统创新平台

针对我国数据资源开放共享不足、综合分析能力薄弱、创新应用领域不广等问题，国家发展和改革委员会2017年正式批复由中国科学院计算技术研究所负责组建大数据分析系统国家工程实验室，将重点建设大数据分析技术创新平台，支撑开展大数据分析基础理论(如大数据复杂性和不确定性分析、大数据机器学习理论）研究、异构大数据预处理、挖掘与分析计算系统、大数据挖掘结果与决策支持、基于大数据的智能知识获取算法、可视化交互与开放接口和大数据分析示范应用等技术的研发和工程化，形成面向大数据分析全生命周期的技术与应用环路，如**图1**所示。

具体地，在大数据复杂性与不确定性分析以及大数据机器学习理论支撑下，建设融合计算架构、大数据预处理、深度分析挖掘、智能知识计算与决策辅助、可视化交互与开放语言的大数据分析技术与标准体系，形成大数据分析系统工程化能力。建立标准化、系统化和协同的大数据资源汇聚平台，建设并完善一套分布全国的大数据分析云基础设施与分析系统，形成面向科学发现和政府治理的示范应用，并通过开放的数据资源、技术与应用生态，服务万众创新，辐射带动社会和经济的发展。

构建大数据分析系统创新平台的技术框架主要包含平台支撑、基础理论、大数据分析技术与系统，具体的技术框架关系如**图2**所示[6-10]。

着重突破大数据分析系统面临的技术

瓶颈，针对性地实现大数据分析系统的单点深入、纵向加速和横向打通。

结合机器学习、认知计算和知识工程，研制深度智能的大数据分析系列算法，提高大数据的价值利用率，具体包括以下内容。

- 研究大数据基础分析算法：利用表达学习与增强学习，提升分类、聚类、查询、检索、匹配、关联分析、回归分析等基础算法的精准性和适用性。
- 研究大数据融合分析技术：利用认知计算与深度学习，提升异构表达、跨媒体抽取与内容理解、异构数据融合分析、异常模式识别等高级算法的效能。
- 研究大数据预测决策与可视化：利用机器学习与知识建模，突破大数据预测、知识推演、可视化分析、辅助决策等大数据分析技术。

研制智能分析硬件与弹性流式分析引擎，提高复杂大数据分析的时效性和规模可扩展性，具体包括以下内容。

- 研究智能芯片与垂直化硬件加速：利用智能分析芯片和硬件重构技术，跨越式提升复杂大数据分析计算的速度。

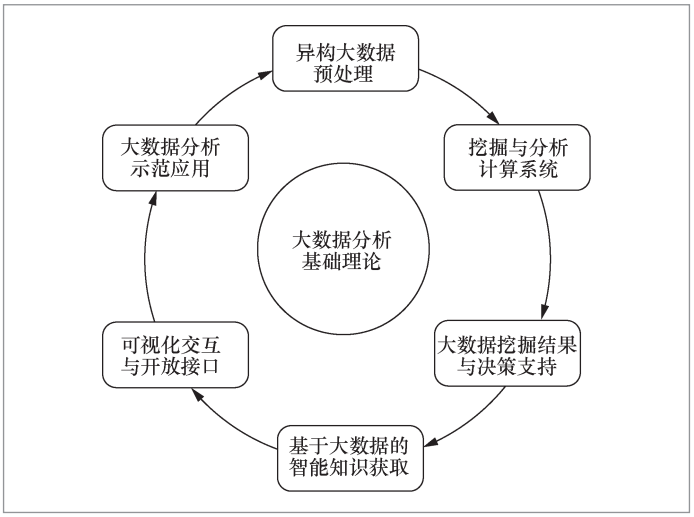


图1 大数据分析技术的价值环路

- 研究泛流式分析架构：研制弹性泛流式分析引擎，支持批式、流式和在线分析。

研制全链路大数据分析软件栈和分层开放编程语言，极大程度提升大数据分析技术的易用性和工程化能力，具体包括以下内容。

- 开发全链条的通用算法工具集：覆盖大数据分析链路的各个环节。
- 研发层次化开放编程语言：实现大数据应用的敏捷开发与应用技术部署。

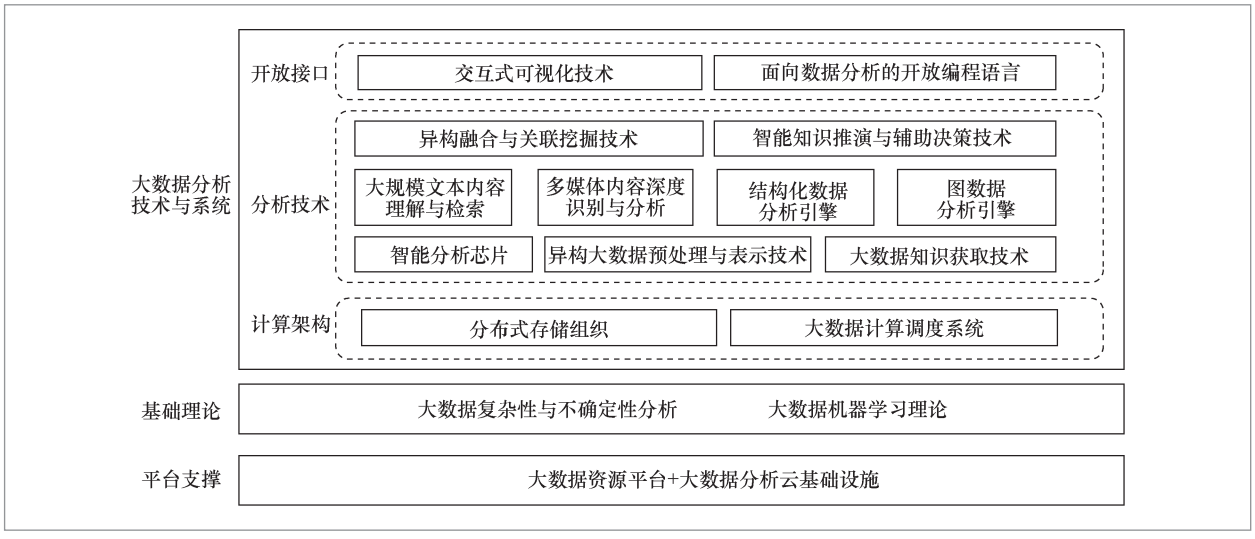


图2 大数据分析系统总体框架

## 4 实现科学决策，提升政府治理能力

大数据分析系统创新平台将为提升政府治理能力提供有力支撑。

首先，大数据特别是政府大数据的交叉汇聚与分析利用将加速政府治理体制从碎片化向网格化的转变。当前，政府治理碎片化是我国政府治理体制存在的主要问题，这种体制给我国的政府治理带来了高成本、低效率的问题。随着大数据时代的到来，所有政府治理主体都可以使用政府大数据进行一系列的处理分析。在这种情况下，将加速实现政府治理大数据的完全共享，不断提升其使用效能、效率和效益，有助于推动各政府治理主体之间的协同与合作，进而加速政府治理体制从碎片化向网格化转变。

其次，大数据分析将促进政府治理方法从以有限个案为基础到“用数据说话”的转变。在大数据时代，政府治理大数据的价值在于提高政府治理决策的科学化与政府治理过程的精细化。为了实现这一目标，各政府治理主体必须对原有的政府治理方法进行大力的改进，重视培养大数据意识，促进相关数据的完全共享，更多地依赖具体数据进行决策，实现从以有限个案为基础向“用数据说话”的转变。

再次，大数据分析将加速政府治理模式从静态治理到动态治理的转变。长期以来，我国政府治理的目标被设定为维持社会稳定，而以这一目标为中心开展的政府治理实践属于静态治理的政府治理模式。与时俱进的大数据技术为全面、及时地掌握政府治理相关数据的变动情况和变动趋势提供了技术保障。对于承担政府治理职责的各个社会主体而言，全面、及时地掌握政府治理相关数据的变动情况和变动趋势，对于实现从静态政府治理模式向动态模式的转变具有非常重要的意义。

最后，大数据分析将有利于推动政府治理方式从简单粗糙向科学决策过渡。目前，我国正处于社会矛盾的凸显期和集中爆发期。在这一阶段，协调利益矛盾，化解社会冲突，维护社会秩序，促进社会稳定，成为大家面临的严峻挑战。为应对这一挑战，亟须对我国原有的自上而下的利益矛盾化解机制进行根本变革，逐步建立上下互动、主体多元的利益矛盾化解机制，促进我国政府治理方式的科学化。大数据分析系统创新平台的建设将为各政府治理主体实现从简单粗糙向精确分析、从关注宏观数据向关注微观数据的转变提供有效支撑，并根据大数据的流程设置相应机构和工作岗位，加速政府治理方式从简单粗放到科学决策的转变。

## 5 提供平台和工具，推动创新创业

当前，大数据已成为一种重要资源，对大数据的分析有利于推动创业创新。大数据分析对于创业的非比寻常的价值主要体现在以下两点。

（1）大数据分析模式可激发创造力

相比先提出假设，然后通过收集和分析数据来验证这种假设的传统研究方法，大数据时代探索世界的方法不再始于假设，而是始于数据，根据数据发现以往不曾发现的联系。

（2）大数据分析技术能预测和满足个性化需求

大数据的核心是预测，预测则是基于对大量结构性和非结构性数据进行的相关性分析。大数据技术可以对人的需求进行

分析预测，以个性化数据为支撑的大数据服务将变得更为精准、有效。

利用大数据分析进行创新创业的主要应用前景如下。

● 订阅式定价模式在大数据分析服务中展现优势，订阅式定价模式是未来大数据分析服务的方向。使用这种模式的创业服务更个性，创业人群更普遍，从而可以有效扩大创业人群。

● 开放数据和开源技术降低了创业门槛，大大增加了创业机会。大数据时代，人们寻找创业机遇，最重要的是通过数据采集和分析，从数据中找到好点子。通过对大数据的收集和分析，可以为创业者提供关于创业的好点子，同时根据市场大数据的分析结果和创业者自身的条件为其量身定制合适的创业项目。通过市场大数据的分析，使项目能够迅速融入市场。

● 大数据技术本身的发展带来全新的创业方向。大数据相关技术的发展，将会创造出一些新的方向，从而引领一些新的市场需求。比如数据采集业，包括定位、支付、社会性网络服务（social networking services，SNS）、邮件等行业；数据技术产业，包括硬件方面的智能管道、数据平台等；数据加工业，包括数据挖掘、数据分析、数据咨询等产业。这些都为创业者们提供了新机遇。

大数据分析系统国家工程实验室基于对大数据复杂性和不确定性等内在特性的深入研究，建立大数据分析的基础理论，在此基础上对大数据分析的关键技术进行深入而系统性的分析与探索，建立相关的技术研发和工程化验证平台，进而实现技术的工程化和产品化，并通过建设大数据开放资源和遍布全国的云分析基础，实现典型的示范应用，最终为社会应用推广提供支撑。通过建设大数据分析系统国家工程实验室，实现大数据分析技术的创新突破，并利用大数据分析系统工程实验室的机制不断更新与完善该系统的核心技术，可以不断保持我国大数据分析技术的先进性。

## 6 建设开放生态，推广技术辐射

大数据分析系统国家工程实验将针对大数据科研、技术与产业创新的需求，产学研用多方共投共建，集产业共性技术和成果转化、企业孵化、技术服务、人才培训交流于一体，以产业化应用为目的，促进国家产业标准和产业政策制订，推动科技成果的产业落地；建设成为国内在学科、学术、技术、产业、机制5个方面均具有引领地位的大型创新平台。构建基于大数据分析系统的开放产学研用生态，实现大数据分析技术的推广辐射，如**图3**所示。

依托中国计算机学会大数据专家委员会，通过组建开放数据共享联盟，组织大数据创新创业大赛与大数据技术、学术、产业大会等形式，大力促进大数据开源社区的建设，包括构建开放共享数据资源池，制订大数据技术与应用标准测试集与测试床等；还将联合中关村大数据产业联盟等行业协会与产业联盟，共同促进大数据产

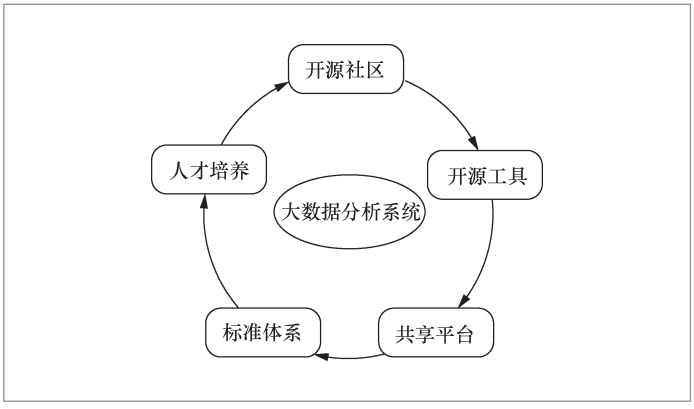


图 3　大数据分析系统开放生态

学研用特别是大数据人才教育培养体系的构建。

我国经济的高速发展急需大数据分析和管理的复合型人才。政府与业务部门更需要能理解数据分析并善于从数据中提炼知识的高层次人才从事现代化公共管理和知识管理活动。大数据分析系统工程实验室的建立将加强对新一代大数据分析和管理复合型人才的培养，营造一个开放、轻松、活泼、创新的科研环境，使实验室成为新学科——“数据科学”的发源地。

## 7 结束语

本文介绍了大数据分析的共性技术与工程化瓶颈，并介绍了大数据分析系统国家工程实验室建设的大数据分析技术创新平台，该平台将有效支撑开展大数据复杂性和不确定性特征描述、异构大数据预处理、挖掘与分析计算系统、大数据挖掘结果与决策支持、基于大数据的智能知识获取算法等技术的研发和工程化，形成面向大数据分析全生命周期的技术与应用环路。依托大数据分析系统国家工程实验室将实现科学决策，提升政府治理能力；为与大数据分析相关的各行各业提供平台和工具，推动创新创业；同时，进一步建设开放生态，推广技术辐射。

## 参考文献:

[1] 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域[J]. 中国科学院院刊, 2012, 27(6): 647-657.
LI G J, CHENG X Q. Big data research: a major strategic area for future science, technology and economic and social development[J]. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657

[2] 李国杰. 对大数据的再认识[J]. 大数据, 2015, 1(1): 2015001.
LI G J. Further understanding of big data[J]. Big Data Research, 2015, 1(1): 2015001.

[3] 王元卓, 靳小龙, 程学旗. 网络大数据：现状与展望[J]. 计算机学报, 2013, 36(6): 1125-1138.
WANG Y Z, JIN X L, CHENG X Q. Network big data: present and future[J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138.

[4] 程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9): 1889-1908.
CHENG X Q, JIN X L, WANG Y Z, et al. Survey on big data system and analytic technology[J]. Journal of Software, 2014, 25(9): 1889-1908.

[5] JIN X L, WAH B W, CHENG X Q, et al. Significance and challenges of big data research[J]. Big Data Research, 2015, 2(2): 59-64.

[6] 靳小龙, 王元卓, 程学旗. 大数据的研究体系与现状[J]. 信息通信技术, 2013, 7(6): 35-43.
JIN X L, WANG Y Z, CHENG X Q. Research system and status quo of big data[J]. Information and Communications Technologies, 2013, 7(6): 35-43.

[7] 孟小峰, 慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169.
MENG X F, CI X. Big data management: concepts, techniques and challenges[J]. Journal of Computer Research and Development, 2013, 50(1): 146-169.

[8] 覃雄派, 王会举, 杜小勇, 等. 大数据分析——RDBMS与MapReduce的竞争与共生[J]. 软件学报, 2012, 23(1): 32-45.
QIN X P, WANG H J, DU X Y, et al. Big data analysis—competition and symbiosis of RDBMS and MapReduce[J]. Journal of Software, 2012, 23(1): 32-45.

[9] 张引，陈敏，廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展，2013，50(suppl)：216-233.
ZHANG Y，CHEN M，LIAO X F. Big data applications：a survey[J]. Journal of Computer Research and Development，2013，50(suppl)：216-233.

[10] 雷军，叶航军，武泽胜，等. 基于开源生态系统的大数据平台研究[J]. 计算机研究与发展，2017，54(1)：80-93.
LEI J，YE H J，WU Z S，et al. Big data platform based on open source ecosystem[J]. Journal of Computer Research and Development，2017，54(1)：80-93.

## 作者简介

**王元卓**（1978-），男，博士，中国科学院计算技术研究所研究员、科研处副处长，大数据分析系统国家工程实验室主任助理，中国计算机学会高级会员，中国计算机学会大数据专家委员会委员，IEEE通信与信息安全技术委员会委员。在大数据分析和数据安全等领域发表论文150余篇，包括重要国际期刊上发表论文30余篇，3次获得国际知名学术会议的最佳论文奖。申请发明专利41项，获授权发明专利15项。曾获得国家科技进步奖二等奖、北京市技术发明奖二等奖、中国电子学会电子信息科学技术奖一等奖等。主要研究方向为开放知识网络计算、社交网络演化分析、网络与信息安全、隐私泄漏行为分析等。

**靳小龙**（1976-），男，博士，中国科学院计算技术研究所副研究员、博士生导师，中国科学院网络数据科学与技术重点实验室网络数据科学研究部负责人，中国计算机学会大数据专家委员会委员。2005年于香港浸会大学获得博士学位。发表国内外学术期刊与会议论文120余篇，获得国际会议CIT-2015、AINA-2007与ICAMT-2003的最佳论文奖。负责或作为骨干参与项目/课题十余项，包括国家重点基础研究发展计划（“973”计划）、国家高技术研究发展计划（“863”计划）、国家自然科学基金重点项目、欧盟FP7项目、英国EPSRC项目等。主要研究方向为社会计算、社会网络、网络数据分析、多智能体系统等。

**程学旗**（1972-），男，大数据分析系统国家工程实验室副主任，中国科学院计算技术研究所研究员、副总工程师、副所长，中国科学院网络数据科学与技术重点实验室主任，中国计算机学会大数据专家委员会秘书长，国家杰出青年科学基金获得者。先后主持并完成了十余项国家自然科学基金、国家重点基础研究发展计划（“973”计划）、国家高技术研究发展计划（“863”计划）、国家信息安全重大专项以及中国科学院知识创新工程等科研任务。两次获得国家科技进步奖二等奖（2012年个人排名第一、2004年个人排名第二），获得第十二届中国青年科技奖、中国计算机学会青年科学家奖、中国科学院青年科学家奖等荣誉。主要研究方向为 Web 信息检索与数据挖掘。

**收稿日期：** 2017-12-15