大数据分析与应用技术创新平台

张平文[1]，鄂维南[2]，袁晓如[3]，傅毅明[4]
1. 北京大学数学科学学院，北京 100871；2. 北京大学大数据科学研究中心，北京 100871；
3. 北京大学信息科学技术学院，北京 100871；4. 北京大数据研究院，北京 100871

摘要
针对我国大数据挖掘与分析能力弱、大数据算法应用和综合能力不高等问题，系统地介绍了大数据分析与应用技术创新平台的总体技术框架，详细分析了我国大数据分析与应用五大共性技术存在的不足和解决思路，并阐述了创新平台中四大支撑平台的设计思路与应用方向，最后对大数据分析与应用技术国家工程实验室未来的发展方向和重点工作进行了介绍。

关键词
大数据；数据处理；质量控制；数据挖掘；可视分析；智能数据管理；国家工程实验室

中图分类号：TP391　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2018043

Big data analysis and application technology innovation platform

ZHANG Pingwen[1], E Weinan[2], YUAN Xiaoru[3], FU Yiming[4]
1. School of Mathematical Science, Peking University, Beijing 100871, China
2. Center for Data Science of Peking University, Beijing 100871, China
3. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
4. Beijing Institute of Big Data Research, Beijing 100871, China

Abstract
In view of bottlenecks faced by big data analysis and application in China, such as weak big data mining and analysis ability, low big data algorithm application and comprehensive ability, the overall technical framework of big data analysis and application technology innovation platform was systematically introduced. The shortcomings and solutions of the five major common technologies in China's big data analysis and application were analyzed in detail. The design ideas and application directions of the four supporting platforms in the innovation platform were introduced. Finally, the future development direction and key work of the National Engineering Laboratory for Big Data Analysis and Application were introduced.

Key words
big data, data processing, quality control, data mining, visual analytics, intelligent data management, National Engineering Laboratory

1 引言

目前，我国在大数据发展和应用方面已具备一定基础，拥有一定的市场优势和发展潜力，但也存在政府数据开放共享不足、产业基础薄弱、缺乏顶层设计和统筹规划、法律法规建设滞后、创新应用领域不广等问题，这些问题亟待解决。《国务院关于印发促进大数据发展行动纲要的通知》提出“加大大数据关键技术研发、产业发展和人才培养力度，着力推进数据汇集和发掘，深化大数据在各行业创新应用，促进大数据产业健康发展”，强调推进基础研究和核心技术攻关。

针对我国大数据挖掘与分析能力弱、大数据算法应用和综合能力不高等问题，急需开展关键技术的攻关研发与突破工作。为践行国家大数据战略，北京大学牵头，联合中国科学院数学与系统科学研究院、中山大学、北京奇虎科技有限公司、北京嘀嘀无限科技发展有限公司、中国信息安全研究院有限公司等多家单位，共同申请建设大数据分析与应用技术国家工程实验室，并于2017年初获得了国家发展和改革委员会的正式批复。其目标是围绕数据科学理论体系、大数据计算系统与分析理论等重大基础研究进行前瞻布局，开展数据科学研究，引导和鼓励在大数据分析的理论、方法及关键应用技术等方面展开探索；加强数据清洗、大数据分析发掘、大数据可视化等领域关键技术攻关；围绕多源异构数据分析，针对视频、文本等多种数据类型，研究大规模机器学习的基础和创新方法，依托合作支撑单位开展面向大数据分析的全流程框架和工业标准的设定、大规模数据分析系统平台的搭建；针对多项代表性应用的大数据分析应用平台进行设计和优化，注重大数据分析理论和系统的源头创新，提升数据分析处理能力、知识发现能力和辅助决策能力；切实和有效地突破目前我国在大数据分析和应用方面的技术瓶颈，形成一整套具有国际影响力的大数据分析系统软件，推动我国大数据产业的发展。本文将系统地分析我国大数据分析与应用的五大共性技术的瓶颈和解决思路，介绍大数据分析系统开发平台、大数据系统测试评估平台、大数据分析可视化展示平台、重大应用示范与系统集成平台四大支撑平台的设计与应用，并对大数据分析与应用技术国家工程实验室未来的发展方向和重点工作进行展望。

2 五大共性技术

如**图1**所示，在一个典型的大数据分析与应用流程中，数据经过预处理后，采用以大数据统计为代表的共性模型和算法与大数据挖掘技术进行计算分析，再结合高度智能灵活的可视分析，最后支持复杂场景下的智能决策。分析的结果将进一步反馈到数据处理，并再次进入分析过程，进一步提供基于数据的洞见能力和决策支持。可见，在大数据分析与应用中，大数据预处理与质量控制技术、大数据分析支撑理论与算法、大数据挖掘技术、大数据可视分析技术、大数据智能知识管理与决策支持技术构成了大数据分析与应用的五大共性技术。这五大共性技术也是构建大数据分析与应用技术创新平台的关键环节。

2.1 大数据预处理与质量控制技术

大数据环境下，数据来源的多样性决定了不同数据源提供的数据可信度是不同的。数据的错误可能来自于数据生成和融合的

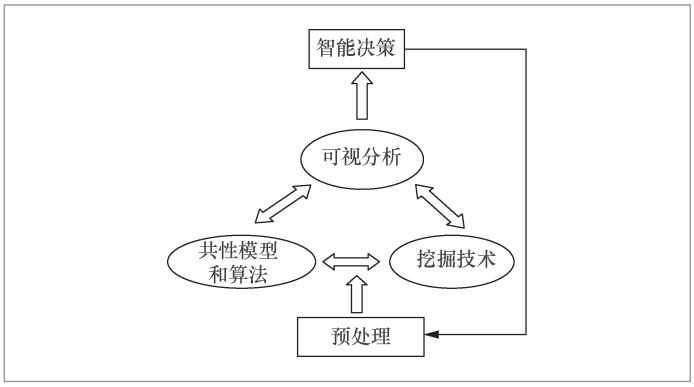

图1 大数据分析与应用技术创新平台的总体框架

各个阶段，包括实体和关系抽取、融合等。

（1）数据中的可信度建模问题

在对复杂结构数据的可信度进行建模时，需要采用“不确定性”和“数据溯源”两个维度。对于不确定性而言，可以采用概率图模型对所研究的复杂结构数据进行数据建模，其中根据应用的需要考虑图模式中的点、边的不确定性以及它们之间的互相依赖关系。除此以外，还考虑到复杂结构数据的多源特性，不同的数据源提供的数据的可信度是不同的。在对数据的不确定性建模的基础上，还要考虑数据的“溯源”问题。大数据分析与应用技术创新平台将数据的不确定性和溯源作为一种统一的元数据，为描述关联数据的可信度提供建模。

（2）数据中的数据质量检测和评估

数据质量决定了数据的可用性。传统的数据质量检测方法中，最经典的方法是挖掘数据之间的函数依赖关系以及带有约束的函数依赖等。然而这些方法只考虑了关系数据库表中各个列之间的函数依赖关系。数据质量的检测需要研究复杂数据的数据质量问题。例如知识图谱数据的数据质量问题需要考虑两个方面的特征：图结构特点以及知识图谱数据中的本体语义。

（3）大数据抽样

大数据分析中有很多问题，其中一个很重要的问题是计算精度和计算复杂度的平衡。一个好的计算方法一方面要误差小，另一方面要计算速度快。大数据分析中这两方面目标几乎不能同时达到。当前应发展更普适的、更稳健的再抽样方法，其普适性将表现在适合于广义线性模型以及一般凸优化问题，其稳健性将表现在统计推断结果的精确性，即均方误差的减小。

2.2 大数据分析支撑理论与算法

数据分析算法大致可以分为计算数学算法、统计算法、计算机算法、优化算法等。由于这些算法来源于不同领域，其侧重点和使用的主要工具都有所不同，然而它们之间又有着千丝万缕的关系，相互关联。

（1）大数据分析的计算数学理论与算法

计算数学包括数值分析、数值代数和最优化等，为大数据分析提供了坚实的理论和算法基础，也拓展了大数据分析的应用领域与范围。在大数据的研究中，要推动变分法、偏微分方程、低秩和稀疏优化的结合，为图像科学和医疗影像提供新的数学模型和数值方法；推动计算数学与机器学习的结合；推动偏微分方程与机器学习的结合，发展高维数据分类、聚类和降维的新方法和理论。针对数据模型，可以推广基于多尺度的降维方法在物理模型中的应用，例如图像表示的小波方法就可以看成一个多尺度方法。深度学习网络的训练也和多重网格算法有很多相似之处，可以系统地研究多尺度方法在非结构数据分析中的应用。基于多层次神经网络的深度学习方法可以看成一种自适应、多尺度、多层次的表示方法，加深了对深度学习的理解，有助于找出其基本原理并推广。神经网

络也可以看成一种特殊的图模型，从图模型的角度来看，神经网络有希望发展出更一般的多尺度、多层次模型。多年来，在计算数学领域发展出了一系列行之有效的并行算法、随机分析方法和张量分析方法，这为大数据分析提供了有效的途径。优化算法是计算数学和大数据分析的主要课题之一，实验室将研究在大数据中如何发展有效的优化算法。

（2）大数据统计分析方法

在分析大数据时，比较大的一个挑战是如何处理大数据的复杂性和不确定性，需要运用统计学原理及方法、机器学习方法来处理这两方面分析中的技术问题。数据的不确定性会带来预测结果的不确定，在统计中可以用随机变量的分布来描述数据的不确定性，然后用置信区间或置信带来量化预测结果的不确定性。

随着计算机硬件的发展与大数据时代的来临，深度学习作为一种通用的人工智能算法，在近些年取得了巨大的成功。对比传统的模式识别方法，深度学习不再需要人工进行特征设计，而是采用端到端的方式进行训练，并在计算机视觉、语音识别和自然语言处理等很多领域都取得了突破性的进展。目前的深度学习算法主要面临模型复杂度较高、速度慢以及需要大量训练数据等问题。为了进一步推动深度学习的发展和广泛应用，仍需要突破以下关键技术：深度神经网络模型的预测加速和模型压缩、类人的深度学习过程以及新型的深度学习优化方法等。随着深度学习的不断发展，各种复杂的网络结构，特别是深层神经网络，由于灵活性较高，具有良好的表达能力，引起了越来越多的关注。但由于缺乏理论上的指导和支撑，目前深层神经网络的大量自由参数大多数情况下仍需依赖经验来调节与优化，离理论上的最优性能仍有较大差距。此外，如此复杂的模型很容易在特定的数据集上得到近乎理想的拟合效果，然而在推广泛化性能上却往往很难得到保障。为了推动深度学习的发展，继续提高模型的性能，需要对复杂神经网络的优化算法进行进一步的研究。

深度学习在许多领域都有广泛的应用。在实际应用中，根据计算平台资源等方面的限制，往往需要设计出不同计算复杂度的模型。由于缺乏系统性方法的指导，很多神经网络的超参数（如卷积神经网络的层数、每层卷积神经网络卷积核的类型与数量等）都需要依赖经验来设定，然后通过大量的实验来验证模型的性能。因此，将深度学习算法在特定问题上进行应用的关键是如何在学习的过程中简单有效地完成超参数的自动选择。

深度学习在监督学习领域已经取得了巨大的成功。基于监督学习的深度学习方法通常需要海量的标注数据进行训练，从而保证模型的推广性能。但在很多应用中，标注数据往往很难获得或者获得成本过高，而没有标注的原始数据往往很容易获得。相对地，人脑可以在没有监督信息时主动地从周围环境中学习。因此，要实现更加高级的智能行为，现有深度学习算法需要突破传统的监督学习模式，从监督学习向半监督学习乃至无监督学习转变，借鉴人与环境之间的交互过程，主动且自动地完成增强学习，以摆脱对监督信息的依赖，在更严苛的环境下完成学习任务。

2.3 大数据挖掘技术

在大数据挖掘技术方面，该实验室主要关注文本、视频、时空等典型数据类型，同时大力开展大规模并行数据挖掘技术的研发。

（1）文本大数据分析技术

深度语义分析、跨领域跨语言情感分析、大规模高精度文本知识挖掘、智能文本

对话、语义搜索引擎等是实现大规模文本数据挖掘和自然语言理解的基础技术。设计制造自动人机对话平台（即智能问答机器助理）被认为是现代人工智能的一项充满挑战也极具困难性的任务。能主动引领新话题的对话模式（即主动式智能问答机器助理）对于现有的被动式人机对话模式而言是全新的突破。语义搜索引擎的应用可以大规模定制化采集任意多个网址、网页和文档信息，利用语义搜索引擎分析相关内容，筛选有价值的信息，构建语义索引，自动汇聚。该平台用于大规模文本的信息筛查和收集，可有效实现基于语义的信息与知识汇聚，有利于基于文本大数据技术的应用，实现科学决策。基于该平台，可以开发行业知识服务与决策支持系统（如智能辅助诊疗、智能新闻报道）、智库信息服务（智能产品服务、智库系统）、通用个人知识助理等行业、专业及通用的知识服务应用。

（2）面向海量视频图像数据的语义分析技术

基于语义的图像视频分析是计算机视觉和多媒体领域的一个基本问题，与人工智能、机器学习、人机交互等领域紧密相关。其中，图像数据相关的主要任务有图像分类、物体检测等，视频相关的任务包括多媒体语义事件检测、视频语义标注等。由于视觉数据的特征表达与真实的语义之间存在“语义鸿沟”的问题，尽管现有的研究工作提出了大量理论与算法，上述问题仍然面临着巨大的挑战。特别地，在图像、视频数量爆炸式增长的今天，如何充分挖掘海量视觉数据及其语义标注信息成为新的技术难题。传统的机器学习模型往往无法充分表达海量视频图像数据的语义复杂度，并进而构建高效的用于预测的模型。因此，需要针对海量视频图像数据的特性进一步扩展与增强现有的模型、理论和算法，并在公开的数据集上评测其性能。面向海量视频图像数据的语义分析技术研究需要突破的关键问题有图像快速索引、多媒体事件检测、视频语义标注等。

（3）时空数据分析技术

时空数据挖掘主要分为以下几类：时空模式挖掘、时空聚类、时空分类、时空异常检测等。在应用层面，可以以社会经济现象感知为目标，以时空大数据为基础，并以时空数据挖掘算法为支撑，发展一系列分析方法，包括个体移动行为模式分析、活动时间变化特征分析、场所情感语义分析等。移动是个体层次空间行为最直接的外在表现，个体移动行为模式分析是利用时空大数据中的移动轨迹信息研究个体移动模式或模型的方法；活动时间变化特征分析是指基于城市不同区域对应的活动日变化曲线，研究其用地特征和在城市运行中所承载的功能；同时，社交媒体（如推特、微博等）中包含了大量文本数据，这些数据成为语义信息获取的重要来源，其中带有位置的社交媒体数据通常占3%，可以利用这部分数据揭示与地理位置有关的语义信息，即进行面向场所的情感语义分析。

（4）大规模并行数据挖掘技术

考虑到当前开放互联环境中海量数据挖掘和应用的需求，该实验室将充分考虑分布式数据的特点，结合不断发展的、多样的存储和计算环境，研究海量分布式数据挖掘的理论、方法和技术、规范体系，力争使我国在相关领域的研究成果进入国际领先行列，重点突破海量数据挖掘系统体系结构、基于云计算的数据模型、分布式数据挖掘算法（包括分类算法、关联规则算法、主成分分析算法）等关键技术。

2.4 大数据可视分析技术

可视分析是对大数据进行分析的一种有效手段，日益受到重视。在中国计算机学会大数据专家委员会发布的2014—2016年大数据发展十大趋势中，可视分析技术连

续3年被列入其中。可视分析技术的目标是使数据分析过程透明化。它结合了可视化、人机交互和自动分析技术。在一个典型的可视分析流程中，自动分析的结果通过可视化展示给用户，用户通过人机交互技术评价、修改和改进自动分析模型，从而得到新的自动分析结果。通过这种方式，可视分析技术将人的经验智慧与机器的运算能力紧密地结合在一起。这其中，由人来定义分析任务的模式，由机器来存储和分析大量的数据。分析结果的可视化则成为人与机器合作的桥梁。

● 大数据原位可视分析技术。最初的面向科学的原位可视化技术将数据处理与可视化代码、数值模拟代码进行耦合，在数值模拟程序运行的过程中进行若干处理与可视化工作，即时对相关特征进行提取并存储，从而能在后续处理中得到完整、精确和可靠的数据分析和可视化结果，用以支持用户的探索与分析。原位处理能在保证数据精度的同时，大幅减少所需数据存储、数据传输以及后续处理的代价。同时，原位处理的过程中，也能即时产生相关的可视化内容，用户可以基于内容对模型模拟过程进行监测、调整和操控。同样的思路可以被推广到一般意义上的复杂大数据分析中，在大数据产生和运输过程中，同步开展有效的可视化工作，可解决大数据可视化的瓶颈问题。

● 多尺度时空可视分析技术。面对多种不同来源、不同粒度的数据，如何在时空内进行融合是分析时空数据的巨大挑战。在统一空间内，将多源数据进行可视化处理并提供有效的可视分析手段是目前的研究热点之一。多尺度时空可视分析技术是探索应对大规模、异构的时空数据时较为通用的交互式分析和探索方法。

● 支持态势感知的大数据可视分析技术。传统的数据分析任务主要是针对单个或者一类数据进行的分析，随着大数据的出现，各式各样具有隐性相关关系的数据涌现出来。在一些复杂任务中，往往需要同时对不同类别的数据进行采集、监控、分析以及响应。支持用户对这类分析过程进行实时感知把握是今后可视分析需要具备的重要功能。

2.5 大数据智能知识管理与决策支持技术

面向大数据的知识发现、管理和决策支持服务平台是大数据分析流程中关键的一环。其中包括4个方面的主要内容：一是面向多源异构数据源的知识图谱构建和融合，将不同的数据源的数据进行有效的整合，形成面向一个领域或者开放领域的完备的知识图谱；二是海量知识数据的存储和查询等数据管理问题，一个面向海量知识数据的高效的知识图谱数据管理系统是提供知识服务和支撑上层决策的基础；三是智能知识检索和分析，笔者希望为用户提供一种基于可视化技术的、交互式的知识数据访问和分析的平台接口，方便普通用户使用知识图谱；四是在上述知识图谱的构建和分析的基础上，针对复杂的问题提出定性定量综合集成建模体系，从而有效支持相关决策。

3 大数据支撑平台

针对上述技术，大数据分析与应用技术创新平台将搭建相应的支撑平台，支持方法和系统的研发与落地。

（1）大数据分析系统开发平台

大数据分析系统开发平台建设专门面向大数据分析技术的支撑平台与系统，实现大数据机器学习算法的高效并行化，为各种机器学习算法提供统一的、用户透明的调度过程。构建两个面向大数据分析的算法库：面向大数据统计与机器学习分析

处理基础算法的并行算法库、面向大数据挖掘技术核心算法的并行算法库。依托天河二号超级计算机，设计面向大数据分析与处理的软件支撑平台，包括基于天河二号的大数据可视化分析系统、支撑大数据可视化的基础并行算法库、实现知识图谱的基础并行算法库与软件系统支撑平台。重点突破大数据背景下深度学习、数据挖掘等各类应用的共性问题，研究解决这些问题的基础算法的并行性，研究并行编程框架及其支撑系统，创建基于天河二号的大数据分析软件支撑平台与系统。

（2）大数据分析系统测试评估平台

大数据分析系统测试评估平台是一个综合性试验测试平台，能够实现各种大数据分析方法和系统在多种数据对象、应用场景下的主客观质量测试和应用验证。大数据分析系统测试评估平台将建立标准测试方法和手段，并建立综合性的测试数据集合，能够提供多种情况下的测试评估能力，对其他平台的结果具有良好的支撑作用。

（3）大数据分析可视化展示平台

大数据分析可视化展示平台将提供先进的可视化显示环境，提供对目标数据分析效果的高效可视化，提供高度沉浸式、高像素分辨率的大规模可视化设施。大数据分析的可视化通常需要同时展示大量、高精度、多角度的复杂图形信息，以便使用者进行观察与分析。

（4）重大应用示范与系统集成平台

重大应用示范与系统集成平台主要通过与共建单位合作，在安全、交通等方面建设包括多种应用在内的大数据重大应用示范与系统集成。

4 未来发展方向

大数据分析与应用技术国家工程实验室瞄准国家的重大需求进行发力和突破，同时，还将以企业和市场的需求为导向，以提高产业自主创新能力为目标，力争在我国大数据分析与应用领域形成一批重大的标志性成果。

在科研与人才的体制机制上，积极进行探索和创新，与北京大学大数据科学研究中心、北京大数据研究院等院校机构和科研单位进行深度协作，整合各方资源，发挥各方优势，努力实现机构共建、人才联聘、成果共享的“三位一体”运行新机制。

在大数据人才培养上，将以往的分学科、分阶段的“串联式”培养方式升级为多学科交叉、理论与实践并行的“并联式”培养方式，联合北京大学数学科学学院、信息科学学院等院系，聚合数学、计算机、工程应用等多方面的资深专家，为大数据领域的优秀人才营造良好的成长环境。在提高人才培养效率的同时，使其具备利用大数据相关理论与实践经验解决具体问题的实战能力。

大数据分析与应用技术国家工程实验室依托北京大学建立研究环境与核心团队，构建创新生态。针对大数据分析技术，建立研究开发和试验平台，开展针对大数据的预处理和质量控制、大数据的计算数学理论与算法、大数据的统计方法及理论、大规模机器学习方法、面向异构多源大数据的挖掘和分析、大数据的可视化以及知识计算和决策支持等方面的研究；针对重点行业应用和数据产业的发展需要，建立大数据分析系统和数据产业工业化应用的全流程开发框架和系统标准；并以支撑平台为基础，加强与支持单位以及优势单位和地区开展的合作研究，突破现有大数据分析的技术瓶颈，引领国内外大数据分析产业的发展，涌现一批具有自主知识产权、形成国际影响力的大数据分析系

统平台，为我国大数据分析产业的发展以及基于大数据的政府决策、企业创新等提供基础支撑。

5 结束语

本文从我国大数据分析与应用的共性技术、大数据支撑平台和应用的现状和存在的问题出发，介绍了大数据分析与应用技术国家工程实验室建设的大数据分析与应用技术创新平台技术架构。创新平台围绕数据科学理论体系、大数据计算系统与分析理论等重大基础研究进行前瞻布局，开展数据科学研究，引导和鼓励在大数据分析的理论、方法及关键应用技术等方面展开探索。创新平台将引领大数据分析技术和国际标准发展，为我国大数据分析产业健康发展提供技术支撑、实验中心和人才基地。在产学研结合方面，该平台将成为支撑产业健康发展的核心试验支撑平台和充满活力的开放式研究开发环境。

作者简介

张平文（1966-），男，中国科学院院士，大数据分析与应用技术国家工程实验室主任，北京大学数学科学学院教授，北京大学学科建设办公室主任、科学与工程计算中心常务副主任，发展中国家科学院院士，长江学者，中国工业与应用数学学会（CSIAM）理事长。曾在第七届国家工业和应用数学大会（2011年）、美国工业和应用数学学会年会（2014年）、2018年国际数学家大会（ICM2018）做大会邀请报告。2014年获得国家自然科学奖二等奖。

鄂维南（1963-），男，中国科学院院士，大数据分析与应用技术国家工程实验室技术委员会主任，普林斯顿大学数学系和应用数学研究所教授、运筹和金融工程系兼职教授，北京大学大数据科学研究中心主任，北京大数据研究院院长，CSIAM大数据与人工智能专业委员会联合主任。1996年获首届美国青年科学家与工程师总统奖，2003年获国际工业与应用数学协会科拉兹（Collatz）奖，2009年获美国工业与应用数学学会克来曼（Kleinman）奖，2014年获美国工业与应用数学学会卡门（Theodore von Kármán）奖。首届美国数学学会、美国工业与应用数学学会Fellow。

袁晓如（1975-），男，北京大学信息科学技术学院研究员，博士生导师，大数据分析与应用技术国家工程实验室常务副主任。中国图象图形学学会理事、可视化与可视分析专业委员会主任。中国计算机学会理事，杰出会员。2008年初在北京大学建立可视化与可视分析实验室，担任信息科学中心/智能科学系副主任，机器感知与智能教育部重点实验室副主任，主要研究方向为复杂流场数据可视化，高维/时空数据，交通、社会媒体数据的分析，可视化的快速构建方法。数十次担任IEEE VIS、EuroVis、IEEE PacificVis等国际可视化会议程序委员会委员。2017年担任IEEE VIS大会论文共同主席。

傅毅明（1984-），男，博士，大数据分析与应用技术国家工程实验室主任助理兼办公室主任，北京大数据研究院大数据分析技术创新中心高级工程师、主任，中国工业与应用数学学会大数据与人工智能专业委员会委员、副秘书长。主要从事大数据与政府治理研究、大数据应用产学研协同创新平台的组织建设等工作。

收稿日期： 2018-05-09