第一届国际数据科学大会在国科大召开

文章来源：中国科学院大学 发布时间：2014-06-04 【字号： 小 中 大 】

5月27日至28日，由中国科学院虚拟经济与数据科学研究中心主办（以下简称“中心”），上海市数据科学重点实验室（复旦大学）、悉尼科技大学数量计算与智能系统(UTS: QCIS)以及西安交通大学管理学院联合协办的第一届国际数据科学大会(ICDS2014)在中国科学院大学雁栖湖校区国际会议中心召开。本次会议以“大数据时代背景探索数据科学新领域”为中心议题。来自美国、英国、日本、澳大利亚、香港和中国大陆地区的80余位专家学者就大数据现有的研究水平、大数据带来的机遇和挑战、数据科学研究未来的发展方向等热点问题展开了颇有成效的探讨，在诸多数据科学相关问题上达成了一定的共识，取得了十分重要和深刻的成果。

27日上午，中国科学院院士、中国科学院大学副校长吴岳良教授致辞，他代表中国科学院大学对各位海内外专家的到来表示欢迎，并对大数据环境下的科研提出了殷切期望，肯定了中心在数据研究方面的科研工作。随后，第九届、第十届全国人大常委会副委员长、中国科学院虚拟经济与数据科学研究中心主任成思危先生致辞，并作题为“静思大数据”的大会报告，他指出目前大数据的研究正处在初步发展阶段，数据量的巨大和繁杂使得数据处理十分困难，如何快速准确地从数据中“沙里淘金”成为研究的热点。成思危进一步指出，对数据的运用已经到了一个前所未有的阶段，但数据分析的主要任务仍旧是预测和评价两种：预测是对未来发展趋势的估计，能提供各种决策支持；评价是对质量的判定，它建立起一定的标准体系、测度方法、表达方法和综合评价方法。成思危认为，通过对大数据的分析研究，我们可以判断事物之间的相关关系，进而找寻它们之间的因果关系。最后，成思危谈到了大数据的三个局限性：第一，大数据虽然为定量研究提供了很好的基础，但对定性分析提出了艰难的挑战；第二，数据分析的结果只能提供决策支持而不能直接代替决策，这就是所谓的“电脑不能代替人脑”；第三，在大数据时代的背景下，如何更好地保护好人的隐私逐渐成为了热点问题。

随后，中国科学院院士、西安交通大学副校长徐宗本教授谈到了大数据研究中的几个关键科学问题。他指出，大数据的显著特点是不能集中存储，这使得很多常规的数据分析方法不能奏效，另外，虽然大数据的价值密度相对较低，但它的出现对社会科学和管理科学有着重要的意义，即提供了量化分析和评价公共政策的方法。徐宗本指出，研究大数据需要学科交叉，更进一步是数据资源管理与公共政策、大数据信息技术、大数据分析的统计和计算以及大数据工程等多方面问题的叠加，从而最终形成大数据产业。具体地，徐宗本提出了大数据研究中的六大问题：第一，高维问题；第二，抽样问题；第三，计算复杂性问题；第四，拼接编码问题；第五，非结构化处理问题；第六，可视化问题。针对这六大问题，他用自己团队所做的研究实例，深入浅出地阐述了解决这些问题的取得的成果和遇到的问题。

接下来，来自美国伊利诺伊大学芝加哥分校的数据挖掘专家Phillip S. Yu谈到了大数据背景下社交网络分析和信息融合等问题。进一步，Yu教授还介绍了其最近的研究成果，即如何通过将用户在社交网络中的活动特征和社会结构特征叠加和综合到用户的基本特征中去，挖掘出更为准确和深层的模式，从而在不同的网络识别相同的用户。紧接着，来自辽宁工程技术大学的著名学者汪培庄教授做了以因素空间为主题的报告，他介绍了因素空间的基础理论，并就如何运用因素空间理论对大数据进行结构化处理谈了自己的看法和观点，同时给大家介绍了近期的研究成果，并指出了运用因素空间处理大数据问题的研究核心和发展方向是寻找合适的基向量，这些新颖的观点引起了与会专家的广泛关注。同时，中国科学院预测科学研究中心主任、中国科学院大学管理学院院长汪寿阳教授向大家介绍了大数据在经济预测与分析方面的进展，尤其是预测中心成功运用独创的TEI@I方法，预测出世界原油和期货价格的变化趋势，得到了国内外同行的高度评价。来自中国科学院计算技术研究所的程学旗教授就大数据的复杂性问题介绍了自己的研究工作。他指出，大数据除了可以分析预测经济走势，还可以发现大规模中的小细节，他举了自己为公安部做的追查逃犯的工作，通过逃犯上网记录，追查其踪迹，最终将其成功抓获。程教授指出，这些研究就涉及到如何在网络大数据环境下寻找罪犯留下的蛛丝马迹以及社交网络中的身份识别问题，更进一步的技术是在网络

中对群体的识别和检测，处理好这些研究中的复杂性问题，能够为反恐和维护国家安全稳定做出重要贡献。

27日下午的主题演讲环节，来自海内外的专家从管理、理论、技术和应用多个方面，对大数据的研究提出了自己的观点。在大数据管理方面，中国科学院大学管理学院的吕本富教授提出了大数据分析师的“十条军规”，这是从管理学角度提出的做数据分析必须遵循的十个重要的规律和规则。同时，他指出大数据带来管理科学研究范式的转变，从以前的“大模型、小样本”发展到现在的“简模型、大数据”，进一步带来了管理实践的转变，这些赋予数据分析新增的价值。在研究理论方面，来自广州大学的黄文学教授探讨了混合类型数据的测度问题，来自香港理工大学的陈小君教授就如何利用p模规划探求稀疏解问题展开了深入浅出的讨论，来自内布拉斯加大学奥马哈分校的王震源教授介绍了在模糊数排序和排秩方面的研究成果，来自计算所的何清研究员汇报了独创的基于超曲面机器学习算法研究工作，来自佐治亚大学统计系的马平教授从统计角度分析了杠杆算法的理论性质和应用价值。在技术和应用方面，来自新泽西州立大学罗格斯特分校的熊辉教授介绍了运用大数据分析技术对移动设备应用中排名欺诈进行研究，来自加州大学洛杉矶分校的王伟教授介绍了图聚类的新方法，来自香港中文大学的Jeffrey Xu Yu教授介绍了在大图像的Random-walk Domination方面的研究成果，来自香港浸会大学的刘际明教授和来自澳大利亚维多利亚大学的张彦春教授分别介绍了自己对健康领域医学大数据数据分析的研究和建模，取得了卓有成效的结果。另外，来自英国布鲁内尔大学的刘小慧教授就智能的数据分析谈了自己的看法，他指出对大数据的研究已经从基本的挖掘分析应用上升到科学问题的高度，使得研究者不需要太多专业知识就可以通过分析大数据得到需要的结果，而且这种分析过程必须是可重复的，能够形成一定的理论方法体系。

28日上午的报告中，来自日本前桥技术研究所的钟宁教授介绍了脑信息科学的研究成果，对脑大数据的计算和健康分析进行了细致的阐述；来自美国佛罗里达大学的朱兴全教授利用哈希方法处理大数据挖掘问题，取得了一定的研究成果。另外，来自德州大学圣安东尼奥分校的孙明和教授介绍了在数据挖掘抽样技术在商品营销方面的应用，利用用户的行为数据，借助张量核函数，构造了支持张量学习机，取得了不错的预测效果。中国科学院大学计算机与控制学院院长黄庆明教授介绍了利用深度学习和多层学习的方法，对多媒体数据进行知识发现，并通过在线聚类方法来减弱大规模噪声的影响。中国科学院大学数学科学学院院长郭田德教授介绍了自己与公安部合作开发的新一代指纹身份证系统，解释了自己在指纹识别问题中采用的新方法和手段，包括基于稀疏表示的指纹图像压缩技术，都取得了显著的成果。

在这些技术层面之外，来自复旦大学的朱扬勇教授就如何训练数据科学家谈了自己的看法，他指出要培养一批真正研究数据本身的科学家，这些科学家是层次性的，掌握从基础到应用的各种知识的同时，要对数据本身做科学研究，这就需要构建数据科学家团队。朱扬勇罗列了国内外在培养数据科学家方面已经开展或正在开展的工作，并进一步介绍了自己领导的上海市数据科学重点实验室的发展情况。西安交通大学管理学院院长黄伟教授做了有关大数据管理研究的报告，他指出，大数据已经发展成一个产业，具有巨大的潜在价值，其技术研究已经取得了很多突破，但与此同时，大数据的隐私保护和管理没有相应地发展起来，如何正确使用大数据来改变世界成为了目前面临的重大决策问题。进一步，黄伟提出了大数据产业链巨大价值背后所遇到的诸多问题和挑战，指出培养专门的数据科学人才——首席数据官(CDOs)迫在眉睫。最后，黄伟提出了发展大数据产业的七点建议：第一，在国家层面制定大数据产业发展战略；第二，制定政策与法规，培养相关人才，引导大数据产业的发展；第三，在全国建立一批研究中心和基地，进行深入的大数据产业研究和发展；第四，建立大数据产业研究中心和基地的示范，供其他省市参考，助其快速发展大数据产业；第五，更改传统的管理和决策模式（从经验和直觉驱动，到数据驱动）；第六，我国需要迅速开展数据质量领域的相关研究，争取在国际数据质量研究和标准ISO8000的制定领域占据一席之地，有话语权；第七，可靠而精准的数据质量和稳定而高效的数据传输是有效处置公共突发事件的基础。

中国科学院虚拟经济与数据科学研究中心常务副主任石勇教授在他的报告中谈到了数据科学和大数据的挑战和发展趋势，指出大数据的到来将深刻改变人们的生活。他提出了目前大数据研究所面临的两个核心问题：第一，研究异构数据(非结构化和半结构化数据)的不同表现形式之间的逻辑关系以寻求基于异构数据的 “多维数据表” 的一般规律；第二，探索大数据复杂性、不确定性特征描述的刻画方法及大数据的系统建模。进一步，对于数据分析方法的发展，石勇提出大数据方法的核心是数据科学。另外，他还介绍了目前大数据的发展趋势，包括大数据的国际会议和国际期刊，并指出中国应该早日公开自己掌握的大量数据，以促进大数据研究的发展。最后，石勇还向与会专家学者介绍了中心在大数据挖掘研究方面已经取得的成绩和几个已经结题的重要应用项目。在与企业界的交流过程中，来自新华社的尹小愚处长和秒针系统的冯是聪副总裁分别介绍了新华08项目的发展情况和秒针系统为多家跨国公司所做的数据分析项目，引起了大家的讨论。

28日下午是会议的分组讨论时间，共有包括数据科学与教育、数据科学和金融应用（新华08）、科学大数据和应用等主题的讨论。另外，本次会议共收到11篇科研文章，这些文章的作者也在分组讨论中汇报了自己的研究成果。

本次数据科学国际会议聚集了一大批国内外数据科学界顶级的专家学者，它为科学界对大数据研究的新思路提供了交流平台，同时，一些企业界人士的参加也加强了学界和商界之间的联系和交流，为促进大数据科研和应用的发展起到了推动作用和有力支持。

打印本页　　关闭本页