新技术论坛

论坛主席及主持人：马帅

论坛主题：大数据智能分析管理

邀请人：吴信东、袁野、童咏昕、崔立真、刘树杰、常雷

论坛组织形式及详细安排（这个根据NDBC的举办方式决定，可以线上线下结合）

拟安排1个上午，第一个报告45分钟，后面5个报告35分钟，共计220分钟。

**报告1：数据中台建设（8:00~8:45，45分钟）**

摘要：数据中台将一个机构（企业，事业，或政府部门）的数据作为战略资产进行管理，是从数据收集到处理应用的一套管理机制，以期提高数据质量，实现广泛的数据共享，最终实现数据价值最大化。数据中台建设覆盖数据的逻辑管理和物理管理，逻辑管理包括数据结构的设计和数据之间相关性的分析，如数据仓库，物理管理包括数据的存储和检索。本报告介绍我们在营销智能和跨姓家谱建设领域的数据中台实践，分析和例证数据中台构建的核心技术和关键问题。

讲者简介：吴信东，明略科技集团首席科学家和明略科学院院长，“大数据知识工程”教育部重点实验室（合肥工业大学）主任，长江学者、IEEE Fellow、AAAS Fellow，营销智能国家新一代人工智能开放创新平台负责人，国家重点研发计划重点专项项目“大数据知识工程基础理论及其应用研究”首席科学家。因为“在数据挖掘及其应用领域的先驱性贡献”，2012年获IEEE计算机学会技术进步奖。 KAIS (Knowledge and Information Systems) 期刊主编和TKDD (ACM Transactions on Knowledge Discovery from Data)的联合主编，数据挖掘国际会议ICDM（IEEE International Conference on Data Mining）的创办人和指导委员会主席 。2005年1月至2008年12月，担任《IEEE Transactions on Knowledge and Data Engineering》(TKDE)两届主编。2014年获IEEE ICDM十年最有影响力论文奖和中国计算机学会优秀博士学位论文奖（导师）。

**报告2：新型大图数据体系结构（8:45~9:20，35分钟）**

摘要：在大数据时代海量多源异构的数据间存在着紧密的关联性，图作为表示数据之间关系的基本结构从传统网络分析领域（如道路交通网络分析、社交网络分析）到新型网络分析领域（如人脑网络分析、知识图谱分析、物联网络分析）有着广泛的应用。新型应用的大图数据除了具有大数据的 4V特点外，还具有局部特征多样性、关联数据复杂性、拓扑结构时变性等新型数据特征。本报告首先将结合实际应用，阐述大图数据的传统特点和新型数据特征，以及由新型数据特点带来的研究挑战和科学问题。其次，介绍我们提出的一套大图数据计算理论框架，及基于该框架的新型大图数据计算模型和管理系统。最后，介绍大图数据分析在医疗健康领域的应用。

讲者简介：袁野博士，现为北京理工计算机学院教授、博士生导师、国家优秀青年基金获得者。主持国家自然基金重点项目，科技部重点研发课题。曾获教育部和辽宁省科技进步一等奖、全国优秀博士论文提名奖、中国计算机学会优秀博士论文奖。袁博士是中国计算机学会数据库专业委员会委员、大数据专家委员会委员、中国计算机学会高级会员，IEEE、ACM高级会员。香港科技大学、香港中文大学、英国爱丁堡大学访问学者。主要研究方向为大数据管理与分析(包括图数据管理、众包数据管理、不确定数据管理、数据隐私保护等)。在SIGMOD、VLDB、ICDE、VLDB Journal、IEEE Trans. TKDE、IEEE Trans. TPDS等重要学术会议和期刊上发表论文70余篇。

**报告3：面向共享出行的大规模群智计算（9:20~9:55，35分钟）**

报告摘要：近年来随着移动互联网技术与共享经济模式的普及，如网约车、送餐与最后一公里派送等共享出行类服务正在快速发展，同时其也为相关的大规模群智计算带来全新的挑战：如何通过线上平台将大规模出行需求智能有效地分配到线下出行服务提供者群体并实现协同优化？时空众包计算是一种基于众包服务提供者的动态时空属性,通过汇聚群体智慧来求解问题的群智计算范式，为上述挑战提供了有效的解决方案。本报告首先简要地回顾了群智计算的发展历史，随后结合本团队近年来与滴滴出行等共享出行类产业界公司的合作案例，特别介绍订单分派与路径规划等技术的最新研究成果。

讲者简介: 童咏昕，博士，北京航空航天大学计算机学院教授，博士生导师，国家自然科学基金优秀青年基金获得者。2014年于香港科技大学获计算机科学与工程学博士学位，随后留校担任研究助理教授，2015年入选北京航空航天大学“卓越百人计划”并加入软件开发环境国家重点实验室工作。长期从事大数据与人工智能领域的研究工作，目前主要研究方向包括：群体智能、众包计算、时空大数据分析与处理、联邦学习与强化学习等。目前共发表学术论文70余篇，其中以第一/通讯作者在CCF A类期刊/会议上持续发表（含接受）论文40余篇，并先后主持了国家重点研发计划课题等10余项科研项目。曾获阿里巴巴集团评选的首届“达摩院青橙奖”、数据库领域国际会议VLDB 2014“杰出演示系统奖”和数据挖掘领域国际竞赛 KDDCup 2020“强化学习赛道分配学习比赛冠军”。担任国际期刊《IEEE Transactions on Big Data》的编委和多个CCF A类会议的程序委员会委员；是CCF高级会员、CCF杰出讲者、CCF会员与分部工作委员会主任助理、“CCF-滴滴大数据联合实验室”学术委员会委员和CCAI智能服务专业委员会常委。

**报告4：全周期数据链与知识驱动的医疗诊断可解释性生成及其应用（9:55~10:30，35分钟）**

摘要：医疗数据具有碎片化、分散广、多模态的特征，如何融合这些多源数据和医疗知识，辅助提高医疗诊断的可解释性具有重要意义。本报告以医保数据为基础，融合来自医疗机构、卫健部门、医保部门的多模态数据，形成完整的数据链，通过对既有全周期全链路海量数据的自主学习，将结构化数据、文本数据和影像数据的特征相结合，自动生成自然语言描述文本且具备可解释性特征的诊断报告。提高诊断报告标准化、规范化程度，大幅度减少医师的诊断报告书写工作量，提高医师工作效率，保证诊断质量，也为管理决策提供科学依据。

讲者简介：崔立真，山东大学教授，软件学院院长、党委副书记，山东大学-南洋理工大学人工智能国际联合研究院联合院长， CCF理事，数据库专委会常委委员、大数据/服务计算/协同计算专委会委员，教育部软件工程专业教指委委员。研究兴趣包括大数据管理与分析，数据智能及其医疗健康应用。先后主持国家自然基金重大研究计划重点项目、面上项目、国家重点研发计划等国家、省部级科研项目30余项。在TPDS、TSC、WWW、KDD、ICDE、AAAI、SIGIR、CIKM、BIBM等发表高水平学术论文60余篇。成果获吴文俊人工智能科技奖、教育部科技进步奖、山东省科技进步奖多项。获AAAI2018\2020两项人工智能创新应用奖。2018年获国家级教学成果奖一等奖。

**报告5：无监督神经机器翻译及其应用（10:30~11:05，35分钟）**

 摘要：不同于通用语种之间的翻译可以使用大规模的双语数据来训练神经机器翻译模型，小语种之间的翻译模型的训练往往缺乏双语数据。本报告将首先介绍无监督神经机器翻译的常用方法，包括无监督双语词典的获取，跨语言词向量的学习，翻译模型的初始化，以及联合训练的无监督神经机器翻译的训练方法。最后将介绍无监督机器翻译在文本风格变换以及代码翻译中的具体应用。

讲者简介：刘树杰，微软亚洲研究院高级研究员。他于2012年从哈尔滨工业大学博士毕业。他的研究兴趣包括自然语言处理，语音处理和深度学习相关技术。他在自然语言处理各顶级期刊和会议（包括CL, ACL, AAAI, EMNLP, NAACL等）上发表论文60余篇，并合著《机器翻译》一书。其研究成果被广泛应用于Microsoft Translator、Skype Translator、Microsoft IME、微软小冰和微软语音服务等微软重要产品中。

**报告6：新一代分布式数据库架构（11:05~11:40，35分钟）**

摘要： 世界上的分布式数据库不下百个，但分布式数据库的架构却只有几种：共享存储，无共享，到较新的计算与存储分离。数据库技术的发展在大数据和云计算的浪潮推动下很快，但一个数据库的架构演进却很慢。这是因为数据库架构的变动需要的开发工作量巨大，牵一发而动全身。本报告将结合讲者设计 Apache HAWQ 和 OushuDB 数据库架构以及与各个行业头部用户（包括国有大行等）真实业务的深度打磨经历，讲述数据库架构的不同带来的对业务场景支持的巨大区别，以及数据库各个核心组件设计的不同，其中包括大规模并行处理，分布式事务，资源管理，查询优化和执行等等。希望能够对数据库的研究提供一些新的视角。

讲者简介：常雷，博士，Apache HAWQ 数据库顶级项目创始人，偶数科技创始人，前 EMC 高级研究员和研发部总监，组建了中国 Greenplum 和 HAWQ 数据库研发团队。长期专注于 AI 和大数据领域，在国内外顶级数据管理期刊和会议（比如 SIGMOD 等）发表数篇论文，并拥有多项专利。中国计算机学会数据库专委会委员，中国大数据产业生态联盟专家和中国人工智能百人专家。 2017 年入选美国著名商业杂志《快公司》“中国商业最具创意人物 100”榜单。他于毕业于北京大学计算机系。