大数据管理与分析方法研究北京市重点实验室

一、实验室简介

“大数据管理与分析方法研究北京市重点实验室”是首个以大数据为主要研究内容命名的北京市重点实验室，挂靠于中国人民大学高瓴人工智能学院，联合公管、经济、新闻、社会与人口、统计等五个学院共同建设，是以大数据和人工智能研究为基础，以人文社会科学应用为依托，以实现产学研一体化为目标的多学科交叉研究实体。

1. 历史沿革

2015 年 5 月 21 日，北京市科学技术委员会发布《关于公布 2014 年度北京 市重点实验室和北京市工程技术研究中心认定名单的通知》，中国人民大学申报的“大数据管理与分析方法研究实验室”被正式认定为北京市重点实验室。

2019 年 5 月，重点实验室在由北京市科技委员会主持的第一个三年绩效考评中获得“良好”的成绩，顺利通过考评，目前进入第二个三年建设周期。

2020 年 8 月，经中国人民大学信息学院党政联席会讨论决定并报学校批准，重点实验室的挂靠学院由信息学院变更为高瓴人工智能学院。

2. 实验室主任

重点实验室由文继荣教授担任实验室主任，文继荣教授是国家“千人计划”特聘教授、博士生导师、中国人民大学信息学院院长和高瓴人工智能学院执行院长，2018 年入选首批“北京市卓越青年科学家”，经费达 5000 万元。2019 年担任北京智源人工智能研究院首席科学家，同时担任教育部第八届科学技术委员会委员、国家重点研发计划重点专项“先进计算与新兴软件”专家组成员、中国计算机学会常务理事、北京市第十三届政协委员、中央统战部党外知识分子建言献策专家组专家等。

3. 建设目标

重点实验室以大数据管理和分析方法为主要研究方向，围绕大数据的存储、索引和查询方法，多媒体大数据分析、大规模互联网数据挖掘等内容开展理论研究工

作，在实时大数据查询与分析、社会大数据分析与预测等方面展开系统研发工作，期望为社会治理、舆情分析与预测等应用领域提供理论和技术的支撑。实验室的建设目标包括：

（1）深入研究大数据人工智能理论和系统构建，推动中国人民大学计算机学科的快速发展；

（2）将大数据人工智能与人文社会科学理论模型相结合，促进计算机科学与人文社会科学的学科交叉，探索大数据人工智能驱动的人文社科研究新范式；

（3）推进产学研一体化，为政府和相关企事业单位的战略决策制定提供支持；辅助北京市“智慧城市”建设，为北京地区的社会发展和经济增长贡献力量。

4. 环境及设备

在学校的大力支持下，重点实验室已初步建成社会大数据中心，目前拥有高性能服务器 215 台，可以有效存储和处理 20PB 以上的数据。在软件资源方面，重点实验室在已有的社会大数据原型系统“时事探针”的基础上，进一步拓展为支持人大乃至北京市人文社会学科研究的支撑平台。重点实验室将从硬件环境、研究氛围、业界实践、国际交流等多个方面，为科研人员及立志从事技术和科研的学生提供良好的学习科研环境与成长空间。

5. 成果概述

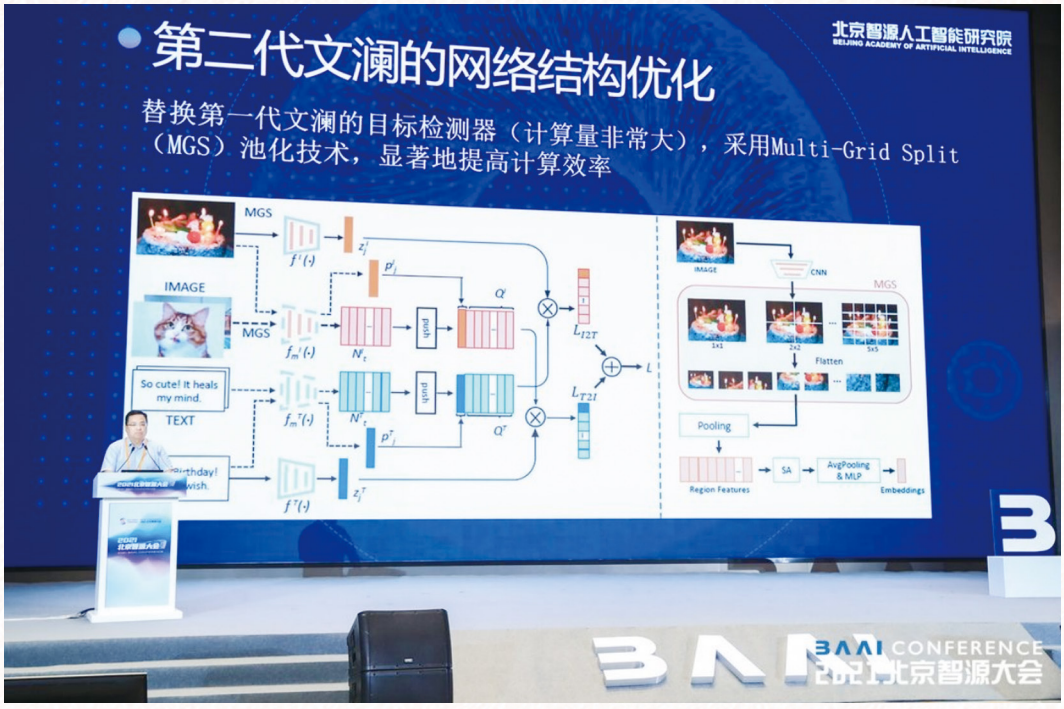
重点实验室先后承担国家级和省部级项目数十项，包括国家 973 计划项目、国家自然科学基金项目、国家社会科学基金项目等。2014 年，中国人民大学（以重点实验室为主要研究力量）与九三学社中央签署合作协议，双方将共同建设公共政策研究中心。重点实验室所开发的“时事探针”互联网大数据分析系统被二百多家国家部委和企事业单位使用，得到了科技部、教育部、中央政法委、中宣部、北京市委等部门领导的高度评价。重点实验室的多项研究成果为中宣部、网信办、九三学社、新华社等部门提供了决策依据与数据支持。在“砥砺奋进的五年”大型成就展中，重点实验室研发的“网络扶贫行动大数据分析平台”得到了中央领导人的高度评价。

二、代表性成果与案例

1. 超大规模多模态预训练模型“文澜”

（1）成果描述：

重点实验室与北京智源研究院合作搭建了“悟道 · 文澜”超大规模多模态预训练模型，该模型是世界上首个基于图 - 文弱相关信息的多模态大规模预训练模型；2020 年 3 月发布悟道·文澜 1.0，基于 3 千万图文对进行双塔模型的训练，在唯一中文多模态公开数据集 AIC-ICC 上获得第一名；2021 年 6 月在智源大会上发布悟道·文澜 2.0 基于 6.5 亿图文对，是目前最大规模的中文通用多模态预训练模型（53 亿参数量，可单卡落地），在三个公开数据集上的实验效果超过 OpenAI 的 CLIP 模型和 Google 的 ALIGN 模型；首次公开了支持 7 种语言的多语言多模态预训练模型，在两个公开数据集上大幅度超越目前最佳结果；设计和实现了小应用《AI 心情电台》和《布灵的想象世界》，具有图像检索歌词和任意句子检索图片的落地能力。



（2）荣誉获奖：

本成果的相关研究成果发表于国内外顶级会议 / 期刊，多次获得国际主流会议的最佳论文或提名，具体获奖情况如下表：

获奖人	论文题目	获得奖项
徐　君	AdaRank: A Boosting Algorithm for Information Retrieval	SIGIR 2019 Test of Time Award Honorable Mention Award
张　静	Arnetminer: Expertise Oriented Search Using Social Networks	SIGKDD2020 Test-of-Time Award
赵　鑫	Comparing Twitter and Traditional Media Using Topic Models	ECIR 2021 Test of Time Award
毛佳昕	Models Versus Satisfaction: Towards a Better Understanding of Evaluation Metrics	SIGIR 2020 最佳论文提名奖
窦志成	Interaction-based Document Matching for Implicit Search Result Diversification	CCIR2021 最佳论文

（3）应用案例：

重点实验室已与华为技术有限公司签署下一代智能信息分发技术研究项目合作协议，双方将在下一代智能信息分发技术领域进行联合创新与长期合作，经费总投入 1650 万元。双方聚焦于“交互式智能信息分发、多模态 / 跨模态信息分发、新型融合计算框架及加速”等前沿方向开展联合攻关，以期实现关键核心技术的自主可控。

此外，2021 年，实验室与中国联合网络通信有限公司签署了战略合作协议，成立联合实验室，利用我校人工智能与计算机学科的专业、人才和技术优势，围绕智能搜索与推荐、人工智能与大数据、智能开放光网络等课题展开长期合作，经费总投入 600 万元。

2. 全国网络扶贫行动大数据分析平台

（1）成果描述：

为贯彻落实习近平总书记关于实施网络扶贫的重要指示精神，践行《网络扶贫行动计划》，发挥互联网助推脱贫攻坚作用，2017 年，在文继荣主任的带领下，实验室与中央网络安全和信息化领导小组开展了全国网络扶贫行动大数据分析平台的建设工作，依托互联网、大数据技术建立“全国网络扶贫行动大数据分析平台”，实时、直观、精准地呈现全国各省、市、县的网络扶贫推进力度。

该平台以互联网舆情信息、网络扶贫行动信息、网络扶贫进展信息、重大扶贫专项数据的实时采集和有效利用为基础，利用自然语言处理、多维分析、可视化分析等大数据技术，对全国网络扶贫行动工作的总体进展及成效从第三方视角实时、精准、客观地进行分析和评估，并以灵活生动的可视化方式加以呈现，为全面掌握互联网对网络扶贫的参与和理解程度，探索网络扶贫工作新模式，实现精准扶贫的科学决策提供了有力支撑。

（2）荣誉获奖：

本成果入选“砥砺奋进的五年”大型成就展，展览由中央宣传部、国家发展改革委、中央军委政治工作部、北京市委主办，累计现场参观人数达 266 万。2017 年 12 月，文继荣教授在第四届世界互联网大会“共享红利：互联网精准扶贫”论坛上发布了“中国网络扶贫大数据分析平台”，引起了与会者的广泛关注，相关工作内容入选中央网信办《网络扶贫动态》，受到中央网信办致信表扬。

（3）应用案例：

重点实验室开发的“时事探针”互联网大数据分析系统被二百多家国家部委和企事业单位使用，得到了科技部、教育部、中央政法委、中宣部、北京市委等部门领导的高度评价，相关研究成果为中宣部、网信办、九三学社、新华社等部门提供了决策依据与数据支持。2014 年，以所在实验室为主要参与研究力量，中国人民大学与九三学社中央签署合作协议，共同建设公共政策研究中心。