代表性成果登记表

申请人姓名：肖仰华　　所在院系：计算与智能创新学院

研究类别：[　]基础创新类　　[√]应用创新类

成果名称：

大规模知识图谱构建、推理与应用研究

成果支撑材料清单（分点列出组成本项成果的高水平论文、专著、授权专利、政策咨询报告等）：

一、高水平论文

1.Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, Yanghua Xiao（通讯作者）. CN-DBpedia: A never-ending Chinese knowledge extraction system. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2018): 428-438（Google Scholar 引用量达 430 余次）

2.Lu Wang, Yanghua Xiao（通讯作者）, Bin Shao, Haixun Wang. How to Partition a Billion-Node Graph. 2014 IEEE 30th International Conference on Data Engineering (ICDE 2014): 568-579（ICDE 2024 十年影响力论文）

3.Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, C. R. Jankowski, Deqing Yang, Yanghua Xiao. Distilling Script Knowledge from Large Language Models for Constrained Language Planning. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023): 4303-4325（ACL 2023 杰出论文奖）

二、专著

肖仰华等编著，《知识图谱：概念与技术》，电子工业出版社，2020 年

三、授权专利

1.基于 CN-DBpedia 的实体识别与链接系统和方法，专利号：201810173270.0

2.中文通用概念图谱纠错装置，专利号：202010303271.X

3.一种基于知识提示的概念抽取系统，专利号：202210419758.3

成果简介（200-300 字）：

知识图谱是人工智能的重要基础设施，其落地面临构建成本高、推理受限、适配困难三重难题。本项目研究大规模知识图谱的构建、推理与应用：建立面向开放环境的知识持续习得机制，提出基于语言模型蒸馏的知识获取方法，形成大规模图谱分割、搜索与问答方法体系。累计发表 CCF A 类论文 114 篇，授权发明专利 3 项，主编专著 2 部。CN-DBpedia 下载逾 5 万次、API 调用量超 18 亿次；成果荣获 ICDE 2024 十年影响力论文奖、ACL 2023 杰出论文奖，另获教育部自然科学奖二等奖、上海市计算机学会科学技术奖一等奖（自然科学奖），在华为、美团、阿里巴巴等企业规模化落地，并获华为火花奖、华为优秀合作伙伴奖等多项合作奖项。

附说明：

一、科学贡献：

智能的本质是知识的发现与应用。将人类的知识赋予机器是提升机器智能水平的重要方式之一，是实现认知智能的必经之路。知识图谱作为一种大规模语义网络，是表达人类知识、开展知识服务的重要形式，逐渐成为推动人工智能产业发展的重要基础设施。然而，知识图谱的产业落地仍面临三重核心难题：构建上规模巨大导致成本高昂，推理上结构稀疏与语义隐含制约认知深度，应用上场景多元要求高效响应与灵活适配。对此，本项目从知识图谱的应用需求出发，对大规模知识图谱的构建、推理与应用展开深入研究，形成如下三个科学发现：

1. 建立了面向开放环境的知识持续习得机制。针对传统知识获取方法在成本、规模与质量上的瓶颈，提出了以“自动抽取—按需更新—持续优化—众包验证”为核心的知识图谱构建新路线。

2. 提出了基于语言模型蒸馏机制的知识获取方法。针对复杂结构知识获取需求，提出利用预训练语言模型蒸馏机制的知识获取方法，显著降低了知识获取的成本，提升了知识获取规模与质量。

3. 提出了面向大规模图谱高效应用的分割、搜索与问答方法体系。针对超大规模知识图谱在实际部署中面临的存储、检索与交互瓶颈，在国际上首次实现十亿节点级知识图谱的高效分割，荣获中国大陆地区首个 ICDE 2024 十年影响力论文奖。

上述科学发现累计发表中国计算机学会（CCF）认定的 A 类期刊或会议论文 114 篇，授权发明专利 3 项，主编教材与专著 2 部，被多所高校用作教材，并被中国台湾地区引进发行。申请人多次受邀在顶级学术会议进行主旨演讲或学术讲习；成果在大规模百科图谱、领域图谱等构建任务中取得良好应用，在华为多个业务线落地，获华为火花奖、华为优秀合作伙伴奖；在阿里、美团等世界级服务平台中规模化落地应用，荣获首届阿里巴巴研究奖、美团科研合作创新奖等，同时支撑了多个校企联合实验室的建设与运营。相关研究另获教育部自然科学奖二等奖、上海市计算机学会科学技术奖一等奖（自然科学奖）（见图 1、图 2）。

项目整体成果受到国际同行高度关注，申请人受到 ADMA、CIKM、DASFAA 等领域国际会议邀请，以学术讲习的形式向国际同行系统性介绍研究成果，并受 CCF 委托组织了知识图谱构建相关的 CCF 前沿讲习班，引起了强烈反响，连续两年获评 CCF 优秀学术主任，连续两年入选 CCF 杰出演讲者。申请人还受邀在国内外知名大学、世界五百强公司等作报告，并在 WAIC、外滩大会等国际重要会议上累计作主旨演讲或特邀报告三百余次。

二、代表性论著评价

申请人构建的中文百科知识图谱 CN-DBpedia（论文 1）被视为一种高质量的知识源及知识获取方法，广泛应用在领域图谱构建等任务中。中国科学院院士王金龙、中国工程院外籍院士郭毅可在内的多位知名学者对此高度评价并进行了多次引用，Google Scholar 引用量达 430 余次。王金龙院士评价“CN-DBpedia 是国内较早推出的也算目前规模较大的开放百科中文知识图谱”。IEEE/ACM Fellow Philip S. Yu 在 2019 年发表于人工智能顶级会议 IJCAI 的研究工作中将 CN-DBpedia 用作实体的核心知识图谱来源来构建其异构社交事件信息网络。IEEE Fellow Mohsen Guizani 在 IEEE Access 2019 的论文中认为项目构建的百科图谱和概念图谱都是大规模知识图谱的典型代表（“there are several large-scale knowledge bases … such as CN-Probase and CN-DBpedia”）。

申请人提出的基于多级标签传播的图分割方法（MLP）（论文 2）被国际同行认可为一种经典的图分割方法，多位 IEEE/ACM/AAAS Fellow 等知名学者将其与 AAAI/IEEE/ACM/AAAS/SIAM Fellow Vipin Kumar 提出的 METIS 方法共同作为典型工作引用或比对。加拿大皇家科学院院士、IEEE/ACM/AAAS Fellow M. Tamer Özsu 教授认为该工作提出的方法 MLP 是与 METIS 并列的两大划分方法。MLP 受到数据库三大权威国际学术会议之一的 ICDE（CCF A 类会议）认可，荣获 ICDE 2024 十年影响力论文奖，系中国大陆地区首获该殊荣，授奖委员会认为该工作极大地降低了图系统的响应延迟和图数据访问成本，进而使得未来需要访问世界知识的大模型系统能够有效落地。社团搜索相关工作得到了众多学者的认可与跟进研究，被不同行业的多位学者应用于社交网络分析、生物信息研究、电商推荐系统、异常行为分析等众多场景，提供了结果精准、响应迅速的高质量检索方案。

申请人提出的基于语言模型蒸馏机制的知识获取方法（论文 3）在自然语言处理顶级会议 ACL 2023 的 4864 篇投稿中脱颖而出，荣获 ACL 2023 杰出论文奖（Outstanding Paper Award）。该工作首次定义了约束语言规划任务，提出了超生成再过滤（over-generate-then-filter）的知识蒸馏方法，从大语言模型中蒸馏脚本知识，构建了包含 5.5 万条脚本的 CoScript 数据集，使约束语言规划的准确性较此前方法提升了 26%，并有效赋予了小型语言模型约束语言规划能力，Google Scholar 引用量达 67 次。组委会评价该工作“通过从大型模型中提炼知识，来教会更小、更高效的语言模型如何执行约束语言规划”；该工作还被 IEEE/AAAI/AAAS/ACM Fellow Amit Sheth 等知名学者作为典型方法进行介绍。

申请人主持编写了《知识图谱：概念与技术》教材，自 2020 年首版以来已发行超过 4 万册，被 17 所高校用作教材，多次进入京东等各大平台计算机书籍销售榜前三，被读者评价为学习知识图谱的“必备读物”，并被中国台湾地区引进发行（2021 年正式出版）（见图 3）。构建了《知识图谱》数智教材平台，获得复旦大学教材建设特等奖。该书被电子工业出版社列

为重点推广书目，申请人荣获2019年度电子工业出版社优秀作者称号。依托知识图谱研究成果，主编《人工智能（高中）》、参编《大数据技术》等多本教材和教参，并构建《知识图谱》在线教材平台积极推广数智教材技术，受聘为首批教育部高教司重点领域教育教学资源及新型教材建设项目技术专家组成员，受教育部及辽宁、广西、陕西等省（自治区）教育厅邀请作报告5次，累计向全国12万余名高校教师及专家介绍经验。

三、研究成果应用与落地

相关成果在2019年中国计算机学会和中国中文信息学会联合主办的语言与智能技术竞赛中荣获全球第一名。在此基础上，该项目进一步构建并公开发布了大规模中文百科知识图谱CN-DBpedia等多个大型图谱，支撑了众多中文知识图谱的相关研究工作。截至2026年7月，CN-DBpedia在多个平台上累计下载次数达到53852次，API调用量超18亿次，不仅推动了中文知识图谱领域的发展，更以更低的更新成本、更高的时效性达到国际领先水平，展现了我国在图谱建设与应用中的技术优势。知识图谱构建相关成果落地应用于华为十多个核心业务线，有力支撑了华为各业务线智能化发展。为华为构建了业界首个企业级知识图谱（包含数千万实体），实现了领域知识的自主更新及知识闭环；其网络运维智能分析模型在华为数字金融、数字政府等场景中应用，带来了数千万元经济效益；在华为整个集团的文档搜索与推荐业务中取得了显著效果，搜索准确率提高10%，知识搜索满意度提升30%以上，知识推荐满意率达到85%。申请人因此荣获华为优秀合作伙伴奖、华为火花奖（见图4）等荣誉，并入选“2019赋能中国人工智能知识图谱20位标杆人物”。

相关成果落地美团生活服务平台，构建了千万级生活服务领域的概念图谱（准确率达到88%，是当时质量最高的产品级领域概念图谱），利用概念性常识形成生活服务领域的认知能力，提升了美团在线业务中的用户搜索理解与商户标签理解水平，显著提升了商品自动化标签匹配的准确率，申请人获得美团科研合作创新奖（见图5）。成果同样落地阿里巴巴飞猪电商平台，利用概念图谱指引平台用户的行为理解分析，大幅提高了用户点击率；面向旅行场景的缩略词生成模型与情感三元组抽取模型上线飞猪APP搜索系统，商品转化率相对提升2.03%。申请人荣获首届阿里巴巴研究奖、阿里巴巴学术合作最佳合作伙伴奖（见图6）。

巨图分割相关工作应用于微软的图数据库系统 Trinity（微软分布式大规模图数据处理引擎Graph Engine的前身）中，为微软知识图谱ProBase提供了基础设施支撑，受到了学术界和工业界的广泛关注与认可。项目团队将相关技术应用在与中航工业、国家电网等多个传统行业龙头企业的合作项目中，支撑了相关行业及企业的客服问答、知识推荐等智能化服务。相关技术吸引了多个企业与投资人对项目成果转化的关注，项目团队与上海爱数信息技术股份有限公司合作成立了国内首个面向内容管理智能化的认知智能校企联合实验室。



图 1 大规模图结构数据管理获教育部自然科学奖二等奖



图 2 大规模知识图谱构建、推理与应用研究获 2025 年度上海市计算机学会科学技术奖一等奖（自然科学奖）



图 3 《知识图谱：概念与技术》教材及中国台湾地区引进发行版



图 4 网络运维智能分析模型在华为数字金融、数字政府等场景中应用，荣获华为优秀合作伙伴奖、华为火花奖



图 5 基于常识知识图谱的大规模细粒度概念提取、缩略词预测、产品上下位关系补全、复合名词解释等工作应用于美团生活服务领域，荣获美团科研合作创新奖



图 6 面向旅行场景的缩略词生成模型与情感三元组抽取模型上线飞猪 APP 搜索系统，荣获阿里巴巴学术合作最佳合作伙伴奖、Alibaba Research Fellow