中国科学院文献情报中心
青年人才领域前沿个人项目任务书

项 目 名 称：　科研人员实体链接和主题标引技术方法研究

资 助 金 额：　10　万元　负 责 人：　于改红

孔贝贝

执 行 年 限：　2019 年 07 月 01 日至 2020 年 06 月 30 日

所 在 部 门：　信息系统部

邮　　箱：　yugh@mail.las.ac.cn

kongbb@mail.las.ac.cn

电　　话：　010-82628382

填 表 日 期：　2019　年　7　月　1　日

中国科学院文献情报中心

2019 年 6 月制

填　表　须　知

一、本表第一至第五项由项目负责人填写。

二、按照表内栏目如实填写，所填栏目不够用时可加附页。

三、本表填写内容表达要明确、严谨，外来语要同时用原文和中文表达。

四、无内容填写的栏目可空白。

五、本表上报原件一份即可。

一、项目基本信息表

<table>
<tr><td colspan="2">项目名称</td><td colspan="8">科研人员实体链接和主题标引技术方法研究</td></tr>
<tr><td colspan="2">成果形式</td><td>A、G</td><td colspan="7">A.论文集 B.研究报告 C.软件平台 D.专利 E.软件著作权 F.专著 G.其他</td></tr>
<tr><td colspan="2">研究期限</td><td colspan="5">2019 年 7 月 1 日至 2020 年 6 月 30 日</td><td>资助金额（万元）</td><td>10</td></tr>
<tr><td rowspan="7">项目负责人情况</td><td>姓 名</td><td>于改红</td><td>性别</td><td>女</td><td>民族</td><td>汉</td><td>出生日期</td><td>1988.08.04</td></tr>
<tr><td>所在部门</td><td colspan="3">信息系统部</td><td>学位</td><td>硕士</td><td>职称</td><td>中级</td></tr>
<tr><td>联系电话</td><td colspan="4">010-82628382</td><td>电子邮件</td><td colspan="2">yugh@mail.las.ac.cn</td></tr>
<tr><td>姓名</td><td>孔贝贝</td><td>性别</td><td>女</td><td>民族</td><td>汉</td><td>出生日期</td><td>1986.09.20</td></tr>
<tr><td>所在部门</td><td colspan="3">信息系统部</td><td>学位</td><td>军事学硕士</td><td>职称</td><td>中级</td></tr>
<tr><td>联系电话</td><td colspan="4">010-82628382</td><td>电子邮件</td><td colspan="2">kongbb@mail.las.ac.cn</td></tr>
<tr><td rowspan="5">主要参加人情况</td><td>姓 名</td><td>出生年月</td><td colspan="2">工作单位和部门</td><td>职 称</td><td>承担任务</td><td>本人签字</td></tr>
<tr><td>钱力</td><td>1981.01.10</td><td colspan="2">中国科学院文献情报中心 信息系统部</td><td>副高</td><td>主题标引框架构建</td><td></td></tr>
<tr><td>谢靖</td><td>1983.10.30</td><td colspan="2">中国科学院文献情报中心 信息系统部</td><td>副高</td><td>实体链接框架构建</td><td></td></tr>
<tr><td>师洪波</td><td>1985.12.20</td><td colspan="2">中国科学院文献情报中心 信息系统部</td><td>中级</td><td>主题标引及实体链接关键技术研究</td><td></td></tr>
<tr><td>胡吉颖</td><td>1988.06.09</td><td colspan="2">中国科学院文献情报中心 信息系统部</td><td>中级</td><td>基于知识图谱的主题标引库构建</td><td></td></tr>
</table>

	余丽	1986.09.08	中国科学院文献情报中心 信息系统部	中级	主题标引、实体链接中机器学习、深度学习方法研究	
	李涵昱	1986.09.27	中国科学院文献情报中心 信息系统部	中级	主题标引、实体链接结果检验	
	常志军	1981.02.28	中国科学院文献情报中心 信息系统部	副高	基于知识图谱的索引搭建	
	张敏	1984 02.26	中国科学院武汉文献情报中心 信息系统部	中级	主题索引体系构建与实现方法研究	
	刘欢	1995.10.21	中国科学院文献情报中心	博士生	深度学习分类算法研究与实现	
	陈小莉	1988.2.22	中国科学院文献情报中心 信息系统部	中级	对科技文献中重要技术进行识别研究	
	刘春江	1984.10.17	中国科学院成都文献情报中心 信息系统部	中级	对基于知识图谱的文献知识库进行分析与构建	
	刘林林	1990.04.02	中国科学院文献情报中心 信息系统部	初级	主题标引及实体链接技术方法实验	
	庞娜	1995.01.20	中国科学院文献情报中心	硕士生	实体链接项目、文献追踪与整理	

二、项目摘要

中文摘要	（对项目的背景、主要研究内容、重要结果、关键数据及其科学意义等做简单概述，500 字以内） 随着网络技术的快速发展，电子化文本数量激增，越来越多的机构将研究成果以电子化文本形式呈现，知识图谱具有强大的知识组织能力，为互联网时代的知识化组织和智能检索应用等奠定了基础，知识图谱中保存的是现实世界中存在的实体及实体之间的关系，由于知识图谱数据来源的多源性，采用实体链接（实体链接是指对于从文本中抽取得到的实体对象，将其链接到知识库中对应的正确实体对象的操作）及主题标引技术保障知识图谱实体及实体关系的精准度、促使知识图谱上层应用服务价值提升很有必要。 本研究以中国科学院文献情报中心的科技大数据知识发现平台的知识图谱作为研究数据，以科研人员实体作为研究对象，就如何高效准确地呈现不同的科研实体类型关注的核心主题，辅助科研人员对科研实体进行集中的重要关联聚焦，引入最新的深度学习分类算法模型，在传统的学术推思路的基础上，结合深度学习分类算法的思想，研究基于深度学习模型的科研实体主题标引方法；同时，基于科研人员的研究主题及其他属性维度，通过建立起适用于中、英文两种科研实体的实体链接算法模型，完成科研人员实体身份的唯一识别，促进实体消歧，优化知识图谱的实体及实体关系数据。
关键词	（不超过 5 个，用分号分开） 知识图谱；主题标引；深度学习；实体链接
英文摘要	With the rapid development of network technology, the number of electronic texts has proliferated. More and more organizations present their research results in electronic texts. Knowledge maps have strong knowledge organization capabilities, and are knowledgeable organizations and intelligent retrieval applications in the Internet age. It lays the foundation. The knowledge map preserves the relationships between entities and entities in the real world. Due to the multi-source of knowledge map data sources, entity links are used. (Entity links refer to entities extracted from text. The object, the operation of linking it to the corresponding correct entity object in the knowledge base) and the topic indexing technology ensure the accuracy of the knowledge map entity and entity relationship, and it is necessary to promote the value of the upper layer application service of the

	knowledge map. This study takes the knowledge map of the science and technology big data knowledge discovery platform of the Chinese Academy of Sciences Document Information Center as the research data, and takes the scientific researcher entity as the research object, and assists the researchers on how to efficiently and accurately present the core themes of different research entity types. The scientific research entity focuses on the important associations and introduces the latest deep learning classification algorithm model. Based on the traditional academic push ideas, combined with the idea of deep learning classification algorithm, the research method of scientific entity subject indexing based on deep learning model is studied. Based on the research topics and other attribute dimensions of researchers, through the establishment of an entity link algorithm model for both Chinese and English research entities, the unique identification of the identity of the scientific researcher is completed, the entity disambiguation is promoted, and the entity of the knowledge map is optimized Entity relationship data.
Keywords	**（limited to 5 keywords, separated by ;）** Knowledge map;Subject indexing;Deep learning;Entity linking

三、研究内容和研究目标

1.研究内容一：基于深度学习的科研实体主题标引技术方法研究

研究目标是探索一种基于深度学习的科研实体主题标引方法，利用深度学习模型对当前海量科研实体大数据进行主题自学习和标引，形成科研实体的主题知识标引库和标引模型。利用该科研实体的主题知识标引库和标引模型可以对科研实体数据，包括期刊、论文、会议、项目、专利、科研人员等进行主题标引等，进而支撑面向文献情报领域各类智能服务应用，提升当前慧科研系列精准服务能力、科技情报监测主题发现服务能力等。

围绕上述研究目标,开展的研究内容包括:

1.1 基于知识图谱的科研实体主题标引知识库构建方法研究

作为项目开展的研究基础和关键，基于主题标引的知识库网络构建对后续深度学习模型的效果具有非常重要的作用。现有的大数据知识图谱网络中，对当前的重要科研实体如（论文、专利、项目、期刊、作者、机构等）从元数据层次进行了组织关联关系构建，但对论文研究主题，作者研究主题以及期刊研究主题之间没有开展的深入的分析研究和对这些科研实体的主题知识单元进行自动标引，因此为了构建下一步主题标引深度学习训练语料库，课题拟开展基于知识图谱的科研实体主题知识库构建模型和方法。

1.2 基于深度学习的主题标引模型研究与实验探索

该部分是本课题的研究重点和关键难点部分，在语料和科研实体主题模型基础上，确定深度学习的训练语料集合和测试集合。该部分拟重点开展在自然语言处理和文本处理上取得突破效果的深度学习模型，如 DNN 模型、LSTM 模型、BERT 模型等，在结合当前主题标引的个性化特征分析，提出适用于主题标引的深度学习模型，并在此基础上开展大量实验研究，实现对不同科研实体的主题标引分类。该部分研究内容包括了基于深度学习的科技文献知识库分类模型构建、分类效果实验和测试、分类模型优化以及对基础数据语料的精炼和优化，每个环节环环相扣、层层关联，构成一个循环的不断优化深度学习优化方案。

1.3 基于深度学习的科研实体主题标引智能服务应用探索

该部分主要围绕当前中心重要战略规划和未来智能知识服务需求，探索课题如何支撑课题服务，提供主题标引引擎工具或标引 API。拟开展三个方向的智能服务探索，（1）支撑慧科研产品服务，主要探索慧科研产品中对科研实体（重点是科研人员以及科研人员的论文成果）进行有效的主题标引，提升精准科研服务能力。（2）支撑领域科技情报监测服务，主要探索在一个垂直领域服务内，对特定的研究机构或研究子领域进行主题标引实验，帮助科研人员把握领域研究重点主题。（3）支撑大数据知识发现服务，发现相关主题的知识组织脉络，提高知识发现效率和检索效率

2.研究内容二：基于知识图谱的科研人员的实体链接方法研究

研究目标是：以科研人员实体作为研究对象，通过对实体链接的关键技术方法进行研究，构建适用于中、英文科研人员实体的链接方法模型，提升当前知识图谱的建设效果，保障知识图谱是科研人员实体及实体关系网络的正确性。

研究内容如下：

1.1 实体链接方法研究

实体链接（Entity Linking，EL）是指把指定文本或数据源的实体链接到目标知识图谱的过程。实体链接的建设过程一般包括：（1）实体查询；（2）生成候选实体集；（3）进行候选实体的排名；（4）NIL（NotIn-Lexicon，未匹配）实体的处理。实体链接的主要挑战是模糊性，实体可能存在有多种形式的表示方式，包括缩写、别名、可变性、多语种表示等，核心是对候选实体集进行排序以挑选正确的映射实体。

1.2 基于科技大数据平台知识图谱的科研人员实体链接模型构建方法研究

当前实体链接的研究多采用 wikipedia 数据，国内外科研人员基于该数据集进行监督学习、无监督学习两种实体链接算法的研究，取得了一定的成果，该数据存在 200 多个语种，当前针对不同语种的实体链接以不同语种的语料为基础或者基于翻译软件进行，但语料的加工需要的人工成本较高时间也较长，已有语料内容不完全适用于本研究；翻译软件的翻译效果是有限的，在科研人员实体方面，无法完成科研人员实体尽可能多的候选实体列表，导致出现 NIL 实体的概率加大。

1.3 基于图的科研人员实体链接模型构建

为了获取更全面的科研人员实体列表，并提升候实体的排序效果，并研究采用如下两种方法：（1）科技文献中的科研人员实体的名称存在多样性，对 SCI、CSSCI、北大核心、南大核心收录的中、英文的出版物的科研人员实体命名方式进行调研，构建中、英文科研人员实体的变体名称，作为科研人员的名称变种，进行科研人员实体检索时，要采用尽可能多的人员名称方式，以保障候选实体的全面性，减少 NIL 实体的出现；（2）在实体排序方面，采用中、英文词向量化完成实体相似度计算，完成跨语言科研人员实体的排序，具有包括三步，第一步，在词向量生成方面，计划采用维基中文、英文文献为基础，采用 STKOS 中的 60 多万英文优先词组及 60 多万的中文词表作为分词基础，为科研人员中英文属性内容（姓名、研究方向等）之间语义相关性的计算打下基础；第二步，采用统计计算的方法，获取科研人员实体对单篇文献中不同关键词的权重值；第三步，通过训练好的词相似度，完成候选科研人员实体的排序。

通过建立起适用于中、英文两种科研实体的实体链接算法模型，借鉴英文科研实体的链接方法，完成科研人员实体身份的唯一识别，促进实体消歧，优化知识图谱的实体及实体关系数据，通过本项目的研究成果，为采用科研文献进行科技知识图谱建设提供算法支撑，保障知识图谱建设的正确性与完整性。

四、研究进度安排及建设成果

（包括月到月的具体进度安排、阶段性成果和最终成果）

1.阶段性成果

1.1 2019 年 6 月-2019 年 8 月

研究当前的知识图谱网络数据组织方法，研究如何基于知识图谱构建基础主题标引库。研究构建基础的主题标引知识库模型和结构，初步完成小规模语料的入库。

实体链接技术相关的文献、项目、会议的跟踪，深入了解实体链接当前采用的有效的链接技术，重点对科研人员实体链接方法、流程研究。

阶段性成果：项目的基础调研

1.2 2019 年 9 月-2019 年 12 月

研究深度学习模型 BERT，并完成模型的构建。设计实验方案，探索模型的有效性和可用性。分析对比实验结果，初步确定一套主题标引模型。

进行有监督实体链接技术的研究，当前有监督实体链接技术的处理流程。进行无监督实体链接技术的研究，无监督实体链接算法的技术特点及方法。进行基于图的实体链接方法及图的向量化表示方法 Graph2vec 等内容的研究，构建实体链接模型。

阶段性成果：项目关键技术深入研究、方法选用、模型初步构建

1.3 2020 年 1 月-2020 年 3 月

不断更新语料库，达到大规模级别。进一步设计不同研究因素变化的实体，探究在不同规模和不同科研实体角度，模型的适用性和扩展性。探索开展智能主题推荐和标引服务。

多语言实体链接技术方法研究与选择，实体链接模型效果（准确度、召回率、F1 评测三方面）评估，科研人员实体链接模型完善，选用不用的链接方法进行测试与效果核验。

阶段性成果：主题标引模型、科研人员实体链接模型优化及评估

1.4 2020 年 4 月-2020 年 6 月

完成智能主题标引引擎 DEMO 工具。

完成科研人员实体链接模型的优化，在科技文献大数据平台上进行应用及进一步的效果提升。

完成项目结题。

阶段性成果：完成主题标引的 demo 示范系统，在科技大数据平台上进行实体链接模型的应用，准备项目结题的相关材料。

2.最终成果

（1）一个示范工具：智能科研实体主题标引引擎工具服务系统；

（2）适用于采用知识图谱技术建设的知识库进行中、英文科研人员实体链接的方法模型；

（3）产出基于知识图谱的科研人员主题标引及实体链接技术相关的论文 1~2 篇；

五、经费预算及详细安排

（包括列出预算支出金额和计算依据）

经费预算表

序号	科目名称	金额（单位：万元）
1	设备费	**0**
1.1	（1）购置设备费	**0**
1.2	（2）研制设备费	**0**
1.3	（3）设备改造与租赁费	**0**
2	材料费	**0**
3	测试化验加工及计算分析费	**0**
4	燃料动力费	**0**
5	差旅/会议/国际合作与交流费	**3**
6	出版/文献/信息传播/知识产权事务费	**2**
7	劳务费	**3.5**
8	专家咨询费	**1.5**
9	其他支出	**0**
10	经费总额	**10**

注：“其他支出”无需列支使用本单位现有仪器设备及房屋、日常水、电、气、暖的消耗补助支出。

预算说明书

（请按照《中国科学院文献情报中心承担院项目（课题）经费管理办法》的有关要求，对各科目支出的主要用途、与项目研究的相关性及测算方法、测算依据进行详细分析说明。）

1．设备费　无

2．材料费　　无

3．测试化验加工及计算分析费　无

4．燃料动力费　　无

5．差旅/会议/国际合作与交流费

课题组成员进行项目科研协作和学术交流（8 人次）等活动发生的费用。按照国家规定的差旅费管理办法住宿标准人均 350 元，伙食补助费人均 100 元，市内交通费人均 80 元，城市间交通费人均 1000 元，出差 3 天计算差旅费：（（100+80+350）*3+1000）*8≈2.1 万.参加学术会议产生的注册费用 ,预计注册费 3000 元,预计 3 人次参加,3000*3=9000 元,合计 3 万.

6．出版/文献/信息传播/知识产权事务费

项目拟发表中文核心期刊 2 篇，每篇按照 3500 元版面费，需要 3500*2=0.7 万，拟投稿英文文章一篇，每篇 8000 元，1.5 万元

购买图书资料,打印印刷费用约 0.5 万元.合计 2 万元

7．劳务费

课题有 1 位博士研究生,1 位硕士研究生参加项目的相关研究工作，按博士 2000 元/月,硕士 1500 元/月，工作 10 个月,合计 3.5 万元。

8．专家咨询费

课题研究过程中拟组织 2 次专家咨询会每次 5 人，每人 1500 元，共需 1.5 万。

9．其他支出

青年人才领域前沿个人项目签批审核表

我接受中国科学院文献情报中心青年人才领域前沿项目的资助，将按照申请书、批准意见和任务书实施本项目。我与项目组成员将严格遵守青年人才领域前沿项目管理规定，切实保证研究时间，按计划认真开展研究工作，按时报送有关材料，及时报告重大变动情况，按规定对资助项目发表的论著和取得的成果进行报告和归档。若填报失实、违反规定，本人将承担全部责任。

项目负责人（签字）：

年　　月　　日

我部门（地区中心）同意　　承担上述青年人才领域前沿项目，并将保证项目负责人及研究队伍的稳定，保证项目实施所需条件，严格遵守中心有关项目管理和财务管理的各项规定，并认真督促实施。

部门（或地区中心）负责人（签字/签章）：

年　　月　　日

业务管理处意见：

按照《中国科学院文献情报中心青年人才领域前沿项目管理实施细则》的相关要求，该项目经申报、论证评审、评审结果公示、中心主任办公会审议等程序，建议准予立项，业务管理处将监督检查该项目的实施过程。

年度拨款计划（单位：万元）

年度	总额	2019 年	2020 年
金额	10	10	0

负责人（签章）：

年　　月　　日

中心意见：

中心主任（签字）：　　（单位公章）

年　　月　　日