网络首发时间：2024-05-28 15:48:26
网络首发地址：https://link.cnki.net/urlid/11.5359.G2.20240527.0931.002

基于社会资本理论的华侨华人知识图谱构建与应用

洪宝惜　林泽斐
（福建师范大学社会历史学院，福州 350117）

摘要：华侨华人群体是推动我国现代化建设和民族复兴的重要力量。以社会资本理论为基础，提出华侨华人知识图谱的构建与应用方案，为面向华侨华人的知识服务提供数据支撑。首先，根据社会资本的来源，分析华侨华人领域的相关概念，进行华侨华人领域本体建模；然后，从华侨华人领域文献中抽取知识单元，使用Neo4j图数据库存储华侨华人领域知识；最后，从社会资本理论的3个维度出发，设计面向华侨华人知识图谱的语义检索和问答策略。基于社会资本来源所设计的华侨华人知识图谱，可揭示华侨华人知识间复杂多样的关系；基于社会资本理论3个维度设计的语义检索和智能问答系统，可以实现华侨华人领域知识的细粒度、多维度呈现。

关键词：社会资本理论；华侨华人；知识图谱
中图分类号：G350　DOI：
引文格式：洪宝惜，林泽斐. 基于社会资本理论的华侨华人知识图谱构建与应用[J]. 数字图书馆论坛，2024，XX（XX）：1-XX.

“华侨”是定居国外的中国公民；“华人”是已入外籍的原中国公民及其后裔[1]。20世纪以来，华侨华人在全球范围内开展了广泛而深刻的政治、商业、文化活动，为中外经贸往来和人文交流作出了巨大贡献，是中国与世界联系的重要纽带。对华侨华人领域知识的数字化开发利用，既有助于华侨华人历史文化的教育与传播，也有助于海外侨胞增强民族认同，保障我国的国际话语空间。然而，华侨华人领域知识广泛多样，知识结构错综复杂，这为华侨华人知识的数字化开发带来了不小的挑战。

目前，对华侨华人知识的数字化开发工作主要基于关系数据库和全文数据库，但这两类数据库在对知识的细粒度语义关联检索方面均存在着一定的局限性。知识图谱（Knowledge Graph）是一种图结构的语义知识库，以结构化的方式描述客观世界中的概念、实体及其相互关系。本研究以社会资本理论为基础，提出了华侨华人知识图谱的构建方法。基于该方法构建的知识图谱能够从知识单元层面细粒度地揭示华侨华人与社会组织、历史事件等实体的深层次关联，从而为华侨华人知识发现、问答系统等下游知识服务提供数据支撑。

1　研究背景

改革开放以来，华侨华人领域知识开发取得了一批实践与研究成果。一系列华侨华人工具书及文献汇编出版，如：北京大学出版社出版的《世界华侨华人词典》，收录了华侨华人人物、华侨华人社团、相关历史事件与重大活动等7 000余个词条[2]；中国华侨出版社出版的

收稿日期：2024-03-05

《华侨华人百科全书》共分为人物卷、社团政党卷、侨乡卷、历史卷、教育科技卷及新闻出版卷等12个分卷，全面系统地介绍了华侨华人的历史、人物、社团政党、文化教育、政策条例法规等各个方面的内容，是改革开放以来侨史研究的重要成果[3]。近年来，数字技术也被应用于华侨华人知识的开发，如：暨南大学图书馆研发了海外侨情数据库等涉侨文献数据库；泉州市图书馆建设有两岸关系谱牒库、泉州馆藏谱牒库等。除上述机构开展的实践工作外，一些学者也对该领域的理论和方法进行了探索，相关研究主要涉及：①华侨华人文献的传承与保护策略[4-5]；②华侨华人文献开发利用对策[6-7]；③华侨华人社会网络指标体系构建[8]。

从上述实践与研究成果来看，虽然华侨华人领域知识的数字化开发已受到关注，但当前的研究主要依赖相对传统的数字化技术，包括将华侨华人文献转化为书目数据库、全文数据库或构建面向基本事实的关系数据库，尚未见有研究者利用知识图谱这一新兴技术手段对华侨华人文献进行开发和利用。

随着关联数据与图数据库技术的发展，一些学者开始对文献中分散的人物数据进行关联集成，构建人物知识图谱，以支持面向人物的语义检索、关联分析和可视化展示，如：Leskinen等[9]基于28 000名芬兰和瑞典学术人员的传记自动抽取与构建家谱知识图谱；徐彤阳等[10]以徐朔方《晚明曲家年谱》为研究素材，构建了晚明曲家年谱本体及知识图谱；沈雪莹等[11]基于李白和杜甫传记资料构建了“李杜”生平知识图谱；张强等[12]构建了皖籍开国将军的知识图谱；程结晶等[13]以《汉书·艺文志》记载的西汉经学家群体为例，探讨了古籍史料中的人物关联组织。上述研究表明，人物知识图谱在挖掘人物关系、揭示人物特征等方面具有独特优势，但当前的人物知识图谱研究主要涉及学术人物、艺术人物和军事人物等，尚未涉及华侨华人这一重要群体。

此外，社会资本理论（Social Capital Theory）是一种分析社会网络及其资源优势的理论框架，在社会学、经济学、信息资源管理等领域均有广泛应用[14-16]。华侨华人和企业在迈向国际舞台的过程中，利用社会网络来获取资源，创造社会价值，这一过程本质上也是对社会资本的利用过程。社会资本理论在华侨华人领域具有良好的适用潜力，但目前还未见有研究将其应用于华侨华人知识的组织和开发。

基于此，本研究以社会资本理论为指导，设计华侨华人领域本体，并从华侨华人文献中抽取知识单元，构建华侨华人知识图谱，并探讨如何从社会资本的3个维度对华侨华人知识图谱进行利用。希冀通过本研究，对华侨华人领域知识的数字化开发起到推动作用。

2　研究框架与数据来源

本研究以社会资本理论为基础设计研究框架，以指导华侨华人知识图谱的构建与应用。社会资本理论由法国社会学家皮埃尔·布迪厄（Pierre Bourdieu）在1980年首次提出，他将经济学的“资本”术语差异化为经济资本、文化资本和社会资本，在这3种资本类型中，社会资本被视为蕴含于社会关系（如群体关系、组织关系以及工作关系）的资源，其强度通常通过某种制度性的关系得以增强[17]。个体或集体可以通过利用社会资本，创造利益和社会价值。

社会资本所依托的关系可以是个体之间的，也可以是群体之间的，甚至可以是跨越整个社会的。按照社会资本来源的不同，可将社会资本细分为微观社会资本、中观社会资本和宏观社会资本[18-19]。微观社会资本是指个体所具有的社会关系网络以及嵌入其中的情感、信任、规则等，其主要来源包括个人的人力资本、职业技能和个人流动性等；中观社会资本是指企业、社团、社区等组织或小型群体所拥有的社会关系网络中嵌入的社会资本，其来源主要为网络、志愿组织和家庭；宏观社会资本则代表了一个国家或地区的社会资本，其主要来源包括社会或国家层面的信任、互惠和规范。

社会资本的理论框架还涉及3个重要的维度：关系资本、结构资本和认知资本[20-21]。关系资本主要是指个体或组织通过其社会网络获取的可利用资源；结构资本则关注社会网络的结构形式、网络密度等结构特征，以及个体或组织在社会网络中的位置；认知资本则包括主体之间的共享语言、愿景、经验、习惯等，这些因素在沟通、理解过程中起到关键作用。

社会资本理论提供了一个理解华侨华人社会关系、信任与合作的宝贵视角。通过这一理论，可以更深入地剖析华侨华人的互动与关联，更好地理解华侨华人群体的特殊性和优势，同时也可以帮助华侨华人群体更好地积累和利用社会资本，推动自身与社会的发展。

基于社会资本理论，本研究构建了由数据来源层、知识图谱层和知识服务层组成的研究框架，如图1所示。

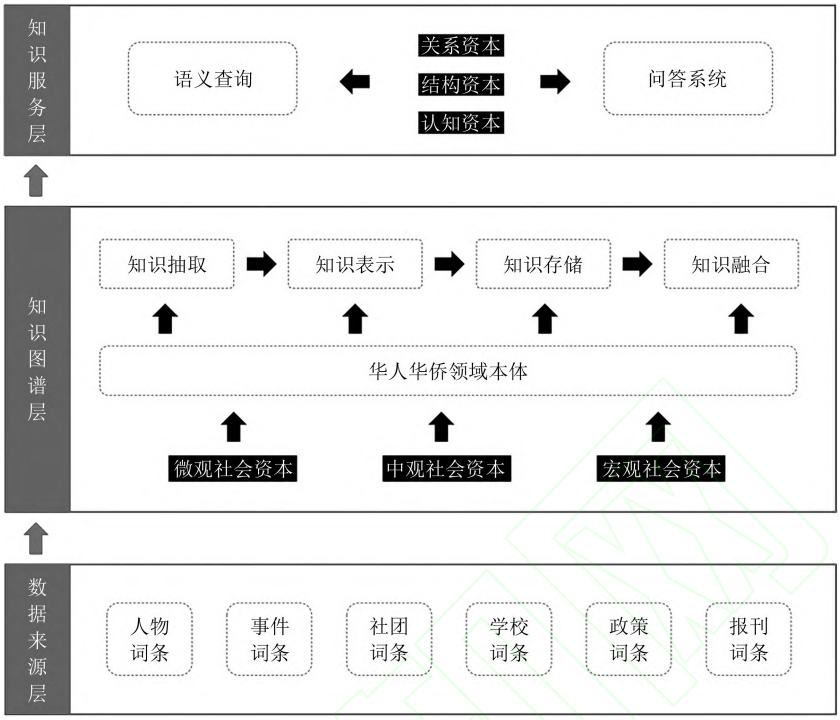

图1　研究框架

数据来源层旨在构建华侨华人知识图谱所需的底层语料库。本研究数据来源为《世界华侨华人词典》和《华侨华人百科全书》两部工具书中的词条。在时间范围上，《世界华侨华人词典》收录了公元2世纪至20世纪90年代重要华侨华人的相关词条，《华侨华人百科全书》则涵盖了公元2世纪至21世纪初各领域代表性华侨华人信息。在地理范围上，两部工具书主要收录了东南亚、北美洲、非洲、欧洲、大洋洲等地区的华侨华人相关词条。在两部工具书中，东南亚、北美洲、欧洲相关词条数量位列前3，反映了这些地区是近代以来华侨华人的主要聚集区。总体而言，两部工具书所收录词条的时间和空间跨度较大，涵盖了华侨华人发展的主要历史时期和主要地域范围。从两部工具书中人工筛选出与华侨华人相关的6类共19 767个词条，包括人物词条5 426个、重要历史事件词条2 610个、社团词条4 763个、报刊词条3 087个、学校词条1 294个以及法律条例政策词条2 587个。通过光学字符识别方式将上述词条转换为数字文本，并通过正则表达式完成数据清洗。

知识图谱层旨在构建华侨华人知识图谱。首先以社会资本的3个来源（微观社会资本、中观社会资本、宏观社会资本）出发，设计华侨华人领域本体，明确华侨华人知识图谱的核心概念及模式框架。在本体模式层所定义的概念体系约束下，从华侨华人相关词条中抽取三元组，并通过知识融合、知识表示和存储，加工形成完善的华侨华人知识图谱。

知识服务层旨在构建华侨华人知识服务系统。该系统以社会资本的3个维度（关系资本、结构资本和认知资本）为思路，以研究构建的华侨华人知识图谱为数据支撑，提出面向华侨华人领域知识的语义查询及问答策略，使华侨华人知识图谱蕴含的潜在知识得以彰显。

3　华侨华人知识图谱的构建

3.1　华侨华人领域本体的构建

在知识图谱领域，本体是在一定的知识范围内对所涉及的概念及其相互关系的形式化表达。通过本体，可定义出一个数据模式，以约束知识图谱中知识的组织结构。与其他人文领域相比，华侨华人知识领域更加强调实体间的跨国籍联系，且华侨华人知识涉及大量的社团、学校、作品等，具有较丰富的内涵。结合华侨华人知识领域的特殊性，本研究基于社会资本理论构建了一个关注华侨华人社会关系、信任与合作的领域本体。

3.1.1 定义核心概念及其等级体系

本体中的概念也称为类。在社会资本理论中，社会资本的来源分为微观、中观和宏观3个层面。本研究基于这3个层面，定义华侨华人的核心概念及其等级体系。微观层面的社会资本主要关注个体，涉及人物的职业技能、社会地位和个人情感等，由此构建人物（Person）、作品（Works）两个一级类；中观层面的社会资本注重群体和组织，由此构建组织（Organization）1个一级类，以及华侨华人社团（OverseasChineseAssociation）、华侨华人学校（OverseasChineseSchool）、华侨华人出版机构（OverseasChinesePress）、其他组织（OtherOrganization）4个二级类；宏观层面的社会资本关注地区或国家层次的社会因素，涉及更广泛的时空环境，由此构建时间（TemporalEntity）、地点（Place）、事件（Event）和法律法规（LawAndRegulation）4个一级类。

3.1.2 定义属性及联系

本体中的属性与联系可分为数据属性和对象属性，前者的作用是丰富实例的特征描述，后者用于建立实例与实例之间的关联，以便实现关联检索。本研究一共构建了28个对象属性和7个数据属性，图2展示了华侨华人领域本体的一级类以及关键的属性与联系。数据属性和对象属性的含义如表1~2所示。

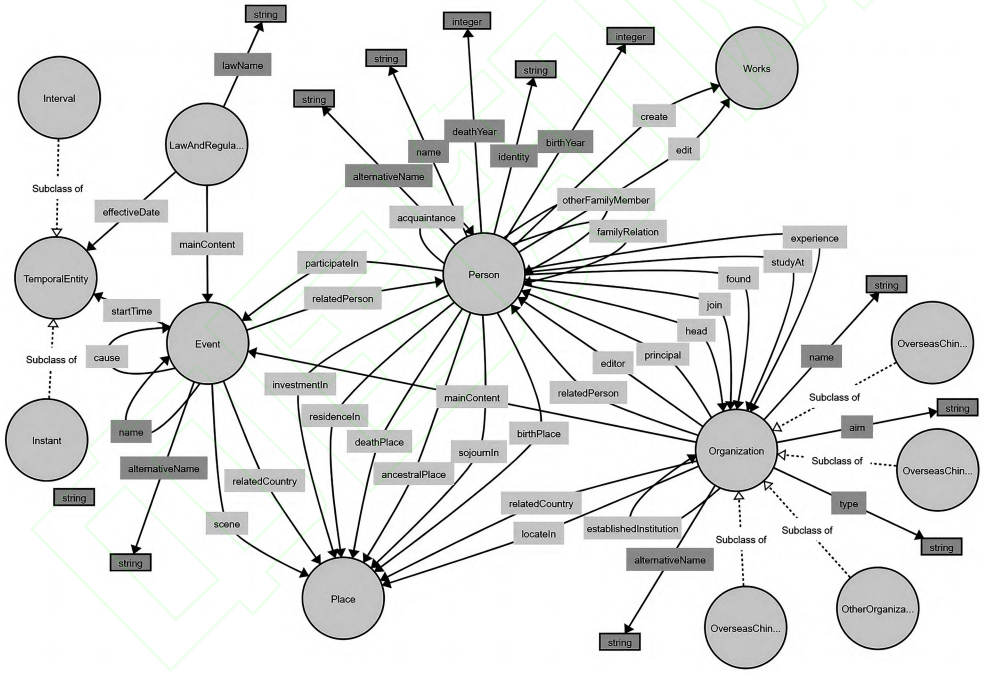

图2 华侨华人领域本体的核心类和属性

3.2 知识抽取

本研究的数据来源词条蕴含着涉及人物、地点、组织等多方面信息的丰富表述，如“陈嘉庚”词条的内容为“陈嘉庚（Tan Kah Kee，1874—1961）爱国华侨领袖。福建同安人。1890年赴新加坡随父经商。两年后任其父开设的顺安米号经理。1904年开始自立门户，先后在新加坡开设新利川和日新菠萝罐头厂，经营福山菠萝园和谦益米店。1906年后又开始经营橡胶种植业。1910年在新加坡加入中国同盟会，曾募捐资助孙中山从事革命活动……”。值得注意的是，词条的主体内容以非结构化文本形式体现，此类文本并不具有固定的格式，不同词条间存在较大差异。华侨华人知识图谱构建的一大关键挑战是从此类非结构化词条文本中提取出本体所限定的结构化信息。

知识抽取是在本体所定义的知识体系指导下，从非结构化文本中自动抽取结构化知识的过程。本研究采用UIE（Universal Information Extraction）框架实现结构化知识的抽取。UIE是于2021年提出的一种通用信息抽取框架，该框架基于百度ERNIE 3.0预训练语言模

表1　华侨华人本体中的数据属性及约束条件

属性名	含义	定义域	值域
name	名称	Event	xsd：string
		Person	
		Organization	
		LawAndRegulation	
alternativeName	替代名	Event	xsd：string
		Person	
		Organization	
birthYear	出生时间	Person	xsd：integer
deathYear	死亡时间	Person	xsd：integer
identity	主要身份	Person	xsd：string
type	类型	Organization	xsd：string
aim	宗旨	Organization	xsd：string

型[22]构建，在实体抽取、关系抽取、事件抽取等多项抽取任务中具有优秀的性能表现。尽管以ERNIE为代表的大型语言模型具有很强的任务迁移和泛化能力，可在零样本的场景中根据提示的指导抽取实体及关系，但为了提升抽取性能，通常还需要针对具体下游任务的需要，对预训练语言模型进行微调。

微调即利用人工标注的监督学习样本，对预训练的基础模型进行进一步训练，以使模型更好地适配具体任务需求。本研究从数据来源中随机抽取词条样本，使用语义标注工具对样本词条涉及的实体名称、关系类型进行人工标注，利用人工标注的样本数据，对UIE默认的预训练模型进行微调，以提高实体和关系抽取性能。具体过程为：使用程序随机抽取550个词条，将词

表2　华侨华人本体中的关键对象属性及约束条件

属性名	含义	定义域	值域
birthPlace	出生地	Person	Place
ancestralPlace	祖籍地	Person	Place
create	创作作品	Person	Works
studyAt	就读学校	Person	Organization
found	创立组织	Person	Organization
join	加入组织	Person	Organization
lead	领导组织	Person	Organization
experience	工作经历	Person	Organization
participateIn	参与事件	Person	Event
investmentIn	投资地	Person	Place
sojournIn	旅居地	Person	Place
residenceIn	侨居地	Person	Place
familyRelation	亲属关系	Person	Person
acquaintance	相识关系	Person	Person
otherFamilyMember	其他家人	Person	Person
locateIn	所在地	Organization	Place
principal	负责人	Organization	Person
relatedPerson	相关人物	Event	Person
		Organization	
relatedCountry	相关国家	Event	Place
		Organization	
		LawAndRegulation	
effectiveDate	实施时间	LawAndRegulation	TemporalEntity

条内容导入标注工具doccano，人工标注词条中的实体名、实体间关系和实体属性，将人工标注后的数据集按照0.70∶0.15∶0.15比例随机切割为训练集、开发集和测试集，使用训练集和开发集进行模型微调和阶段性评估（迭代30轮），使用测试集进行微调后的最终评估。表3~4列出了微调前后对所有实体与关系进行抽取的宏平均精确率、召回率和F1得分，可以看出，经过人工标注样本微调后，模型的抽取性能得到了较大提升。

完成微调后，使用经过微调的知识抽取模型，对所有19 767个词条进行知识抽取，共抽取得到12 927个人

表3　微调前后实体识别性能对比（宏平均）

类别	精确率	召回率	F1得分
微调前	0.148	0.091	0.093
微调后	0.689	0.557	0.606

表4　微调前后关系抽取性能对比（宏平均）

类别	精确率	召回率	F1得分
微调前	0.416	0.256	0.295
微调后	0.730	0.589	0.637

物实体、3 034个作品实体、23 339个组织实体、5 413个时间实体、2 947个地点实体、5 750个事件实体和1 690个法律法规实体，这些实体间共存在着41 108对关系。在此基础上，本研究对抽取结果进一步进行人工检验与修正，以提高知识抽取的准确性。

3.3　知识表示与存储

本研究采用属性图模型来实现知识的表示，使用开源的图数据库管理系统Neo4j作为知识存储引擎。Neo4j是被广泛使用的图数据库，其支持高效的图结构数据存储和查询，并且具有良好的可扩展性和灵活性；属性图模型是Neo4j所采用的数据模型，在属性图模型中，每个节点代表一个实体，边则表示这些实体之间的关系。每个节点和边都可以有多个属性，这些属性提供了关于实体或关系的附加信息；每个节点和边通常带有标签，以区分不同类型的实体和关系。本研究首先对从文本中抽取的实体、属性和关系进行整合处理，转化为CSV格式文件。这些文件被保存在Neo4j根目录下的import文件夹中。然后，使用Cypher查询语言中的load csv命令将所有节点、属性和关系批量导入Neo4j。完成知识表示和存储后，生成的Neo4j数据库共包含55 100个节点和41 108条边。最后为节点的常用属性创建模式索引，以便提高查询速度。

3.4　知识融合

知识融合是指对抽取得到的知识进行规范化处理，消除数据中的冗余和歧义，以便将多源数据集成在一起，形成更全面、更准确的知识图谱。本研究主要针对华侨华人知识图谱中的地理信息进行知识融合。

华侨华人知识图谱中的地理信息至关重要，它揭示了华侨华人的地理分布特征和迁徙模式，对于华侨华人知识的语义查询、知识推理和可视化等应用都具有重要作用。然而，从文献中直接抽取得到的地名存在着两个方面的问题：一是地名表述形式多样，如旧金山又可表述为三藩市；二是文本内容缺乏足够的背景地理信息，例如，仅凭词条描述，往往难以得知某地所在的城市、省份、国家和地理坐标。上述问题可能导致地理信息语义关联困难，从而影响华侨华人知识图谱的准确性和完整性。

本研究采用以下方法对抽取得到的地名进行规范化处理：首先，通过地理信息应用程序编程接口[23]识别地理节点中地名对应的国家、省、市行政区；然后，在上一阶段创建的图数据库中，为识别到的行政区创建三级行政区节点（节点标签分别为country、province和city），每个行政区节点均以属性方式写入应用程序编程接口提供的中心点经纬度坐标；最后，建立三级行政区节点与原知识图谱中地点节点之间的关联。图3所示为地名规范化的典型实例。根据知识抽取结果，人物林同春和陈孝奇的出生地分别为福建福清和福建福州，经过地名规范化后，知识图谱自动创建城市节点福州市、省份节点福建省和国家节点中华人民共和国3个标准行政区节点，分别与地点福清、福州相连接，由此可知两人物出生地有明显关联。通过上述方法，通过应用程序编程接口识别到634个城市级别行政区、309个省级行政区和95个国家。在知识图谱中创建三级行政区节点并将其与地点节点关联后，最终形成包含56 138个节点和46 213条边的华侨华人知识图谱。

4　华侨华人知识的语义检索与可视化

社会资本有助于华侨华人建立良好的社会关系，促进自我实现，最终实现个人和社会的全面发展。社会资本涉及3个维度，分别为关系资本、结构资本、认知资本。本节探讨如何从3个维度的社会资本出发，实现对华侨华人知识图谱的语义检索和可视化。

4.1　基于关系资本的语义检索与可视化

华侨华人是一个具有深厚社会关系的群体，他们通过血缘、友情、地缘和业缘等关系形成了复杂而紧密的网络。关系资本指个体或组织通过其社会网络获取的可利用资源。这种资源可以包括信息、支持、合作机

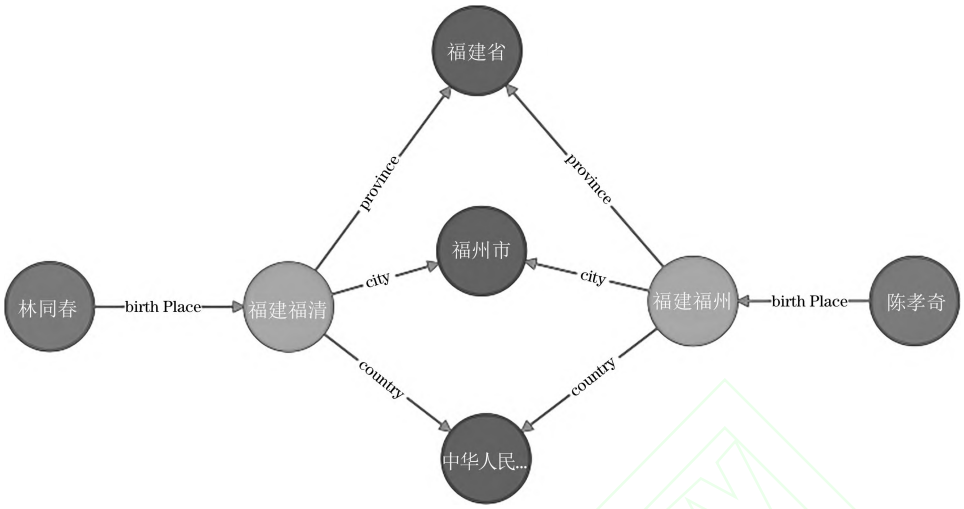

图3　经地名规范化后的地理信息关联实例

会等，来源于个体或组织的社会关系。关系资本往往被视为社会资本积累的逻辑起点。

最基本的关系资本为个体与个体之间的二元关系，可通过血缘、婚姻、友情、合作伙伴等形成。在华侨华人知识图谱中，可通过查询某一中心人物的相关人物节点，或查询两人物间的最短关系路径获取二元关系。如图4~5所示，分别以Cypher查询语句“Match（p1：Person）-[r]-（p2：Person{name：‘林文庆’}）Return p1，p2；”和“Match（p1：Person{name：‘杨纯美’}），（p2：Person{name：‘李清泉’}），p=shortestpath（（p1）-[*..10]-（p2））Return p；”在Neo4j中检索得到新加坡华人林文庆重要人物关系，以及杨纯美、李清泉两位爱国华侨的最短关系路径。

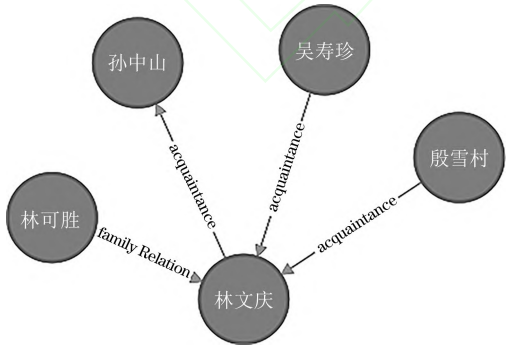

图4　林文庆的重要人物关系

多元关系则是更为复杂的社会结构，涉及超过两个个体之间的互动。这些关系往往是基于共同的背景如同校、同乡、同事等所形成的社会网络，个体可以通过这个网络获取各种可利用的资源，得到职业机会、行

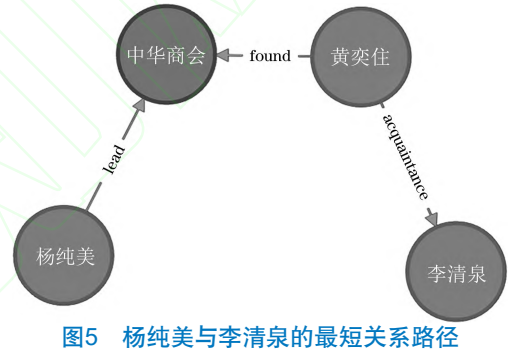

图5　杨纯美与李清泉的最短关系路径

业信息等，校友关系就是一种典型的多元关系。如图6所示，以Cypher语句“Match（p1：Person）-[：studyAt]->（o：Organization）<-[：studyAt]-（p2：Person）Where o.name=‘南洋大学’ Return p1，p2，o；”检索得到南洋大学的部分华侨华人校友关系。对图6中的人物关系作进一步分析，可发现黄孟文、何万成、谢诗坚、林通光的祖籍地均为广东，杜红与杨松年的祖籍地均为福建，林任君与林通光均在《星洲日报》工作过。籍由此类多元关系，个体被置于各类网络结构中，这些网络结构交叉、重叠并发展扩大，促进了华侨华人之间的广泛沟通和信任。

4.2　基于结构资本的语义检索与可视化

结构资本将网络分析提升至全局系统层面，强调网络各参与者间相互作用所形成的结构性关系。

某些参与者可能位于连接不同群体的优势位置，

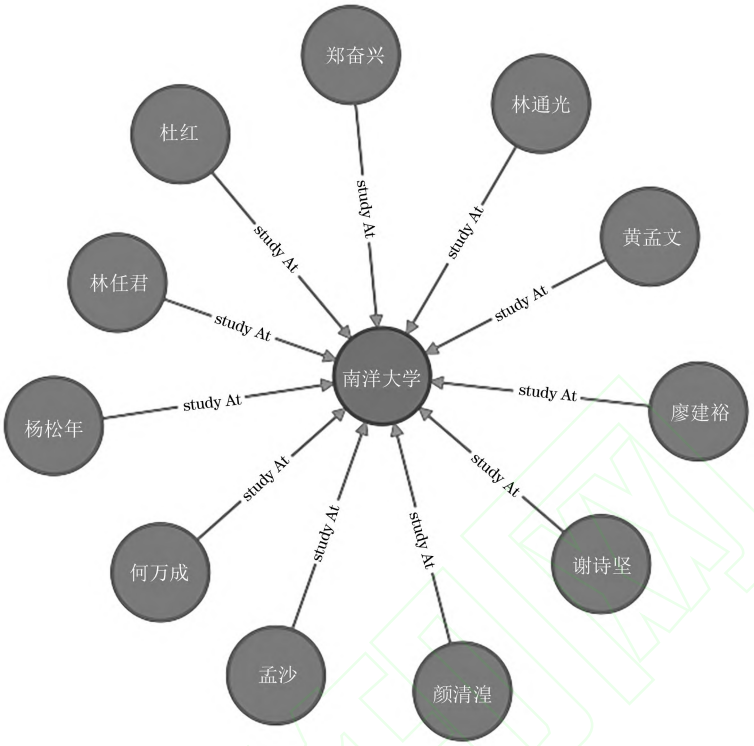

图6　南洋大学华侨华人校友关系（代表性人物）

因此具有获取信息的优越性。在社会网络分析中，通常会采用中心性来衡量个体在网络中的核心作用。度中心性即一个节点直接连接的其他节点的数量。通过计算华侨华人中人物节点的关联人物节点数量，可统计人物的度中心性。如表5所示，以Cypher语句“Match（p1）-[：residenceIn]->（e：Place）Where e.name=‘新加坡’Match（p1）-[r]-（other：Person）Return other.name As name，count（r）as centrality Order By centrality Desc Limit 5；”统计得到新加坡华侨华人度中心性TOP 5。检索华侨华人研究文献可印证，这5位人物均为新加坡华侨华人的杰出代表，他们在各自的领域中都取得了显著的成就。

中介中心性即一个节点在网络中的所有最短路径中出现的频率，中介中心性高的人物通常在网络中扮演“中间人”角色。Neo4j图数据科学库（Graph Data Science Playground，GDS）可统计中介中心性。表6显示了基于GDS统计的华侨华人中介中心性TOP 5人物，中介中心性得分最高的人物节点为孙中山。在40年革命历程中，孙中山先生约有一半的时间奔走于海外华侨社团、留学生和侨领之间，在海外华侨中做教育启蒙、宣传鼓动、组织策划的革命工作，在华侨华人中扮演着“连接者”或“桥梁”的角色；其他得分较高的人物也均对华侨华人间的信息传播有着重要的影响。

在社会网络中，社区是指关系更为紧密、交互更为频繁的子网络群体。社区探测是使用特定算法来识别和揭示社区结构的过程，有助于揭示社会资本如何通过社区内部和社区间的联系而积累和流通。本研究利用GDS的Louvain算法[24]对知识图谱中华侨华人的

表5　新加坡华侨华人度中心性TOP 5

序号	节点名称
1	陈嘉庚
2	薛佛记
3	陈若锦
4	陈金钟
5	张永福

表6　华侨华人中介中心性TOP 5

序号	节点名称
1	孙中山
2	何香凝
3	陈友仁
4	林文庆
5	陈嘉庚

社区结构进行划分，其中人数排名前5的社区如表7所示，其中：社区1中的人物大多与孙中山及其领导的中国同盟会有着密切的联系，他们大多是孙中山的追随者，支持或资助孙中山的革命活动；社区2则以陈嘉庚为中心，涵盖了与他有直接或间接关联的人物，如王雨亭、庄西言、庄世平和庄希泉等；社区3和社区4是基于亲缘或友谊关系聚集的群体，显示了家族世代间的紧密联系和传承；社区5则为菲律宾华侨意见领袖。

表7　华侨华人社区探测结果

序号	人数	关键节点
1	9	孙中山、伍廷芳、张民达、林文英、郑智勇、戴愧生、黄隆生、张静江、郑振秀
2	7	陈嘉庚、王雨亭、庄世平、庄西言、庄希泉、李光前、侯西反
3	6	郭恒节、黄仲涵、黄志信、黄宗孝、黄宗宜、黄宗宣
4	5	庄清泉、庄材润、庄长泰、庄杰华、庄允顺
5	5	吴克诚、薛芬士、薛敏佬、薛华诚、杨孔莺

4.3　基于认知资本的语义检索与可视化

在共同的经历与文化背景下，人们能够形成共同的认知和行为模式，孕育相似的思维方式和价值观，进一步增强彼此之间的信任和协作。认知资本侧重于强调群体内部的共享解释、表达和意义系统，其生成依赖于行动者之间的共同语言、共享经历和愿景。

相同的地域文化是形成认知资本的重要条件，同一出生地、同一侨居地的华侨华人可能面临相似的跨文化适应和融入挑战，需要应对语言、文化和社会制度等方面的问题，同时也需要积极适应并融入当地社会，这种跨文化适应和融入的经验可以成为他们互助基础。华侨华人知识图谱中存储了华侨华人的祖籍地和居住地信息，借助Neo4j的NeoMap地理可视化插件[25]，可以直观展示华侨华人的祖籍地的宏观分布。例如，为输出泰国华侨华人的祖籍地分布，可在NeoMap中以“Match（a：Person）-[：residenceIn]->（：Place）-[：country]->（：Country{name：‘泰国’}）Match（a：Person）-[：ancestralPlace]->（：Place）-[：city]->（c：City）Return c.lat AS latitude，c.lng AS longitude，c.name AS tooltip；”为检索语句，将图谱类型设置为热力图，得出泰国华侨华人的祖籍地主要分布在广东、福建、海南等地。在此基础上，可以从微观视角进一步输出同祖

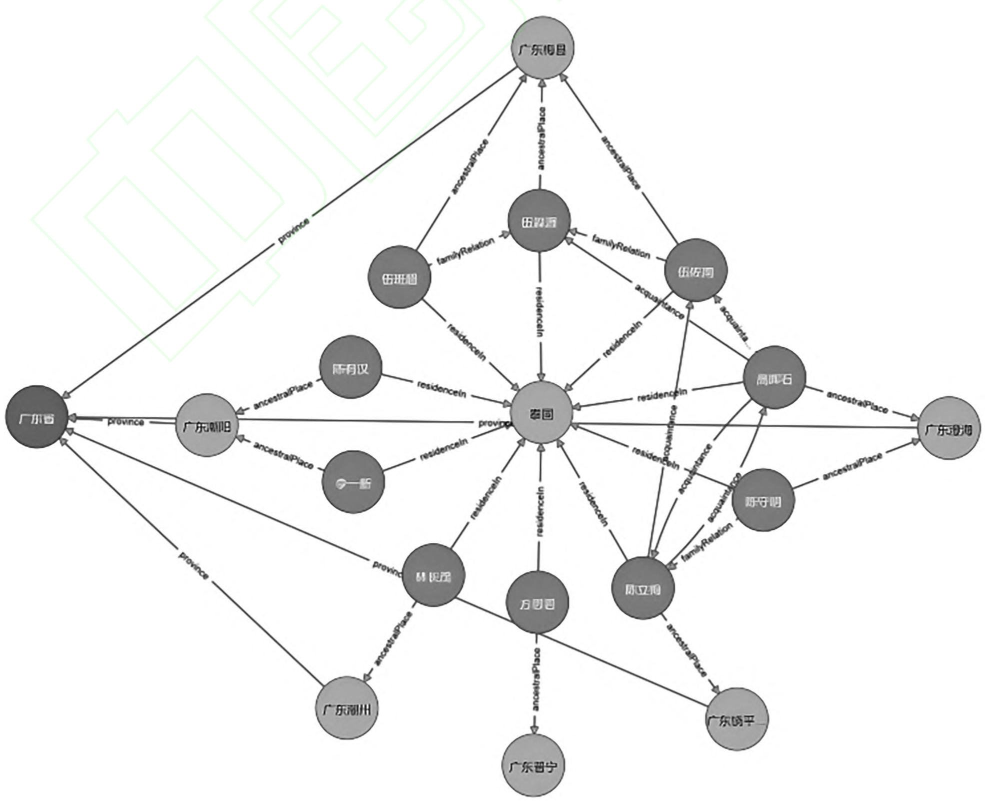

图7　祖籍地为广东、侨居地为泰国的重要华侨华人

籍地、同居住地的华侨华人。如图7所示，以Cypher语句“Match（p：Person）-[：residenceIn]-＞（：Place）-[：country]-＞（a：Country{name：‘泰国’}）Match（p：Person）-[：ancestralPlace]-＞（：Place）-[：province]-＞（b：Province{name：‘广东省’}）Return p，a，b；”输出祖籍地为广东、侨居地为泰国的部分重要华侨华人。

当华侨华人参与相同事件时，可以通过协作和相互支持来建立信任。这些经历塑造了他们的身份认同、文化价值观和社会网络，同时也可间接反映出他们与祖国的联系和互动方式。在华侨华人知识图谱中，可以从某一历史事件出发，检索共同参与事件的华侨华人。如图8所示，以“Match（p1：Person）-[：participateIn]-＞（e1：Event {name：[“辛亥革命”]}）Match（p2：Person）-[：participateIn]-＞（e2：Event{name：[“抗日救亡运动”]}）Return p1，e1，p2，e2；”查询得到参与辛亥革命和抗日救亡运动的部分关键华侨华人。

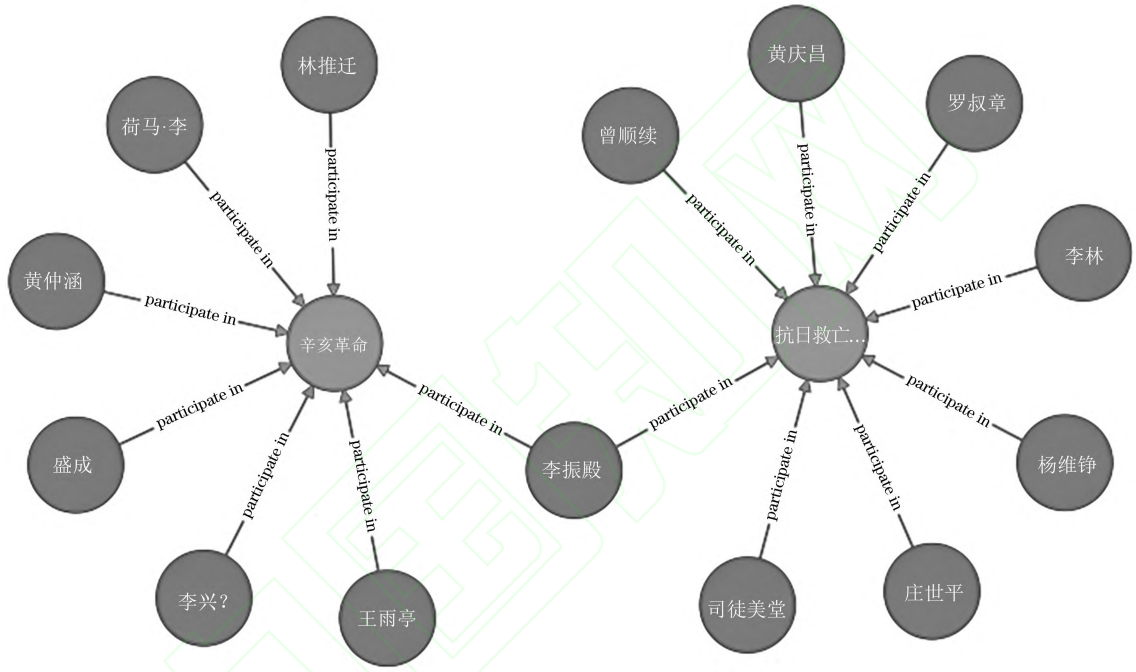

图8　参与辛亥革命和抗日救亡运动的重要华侨华人

5　华侨华人知识问答系统的实现

近年来，以GPT为代表的大型语言模型已成为问答系统关键支撑技术。尽管大型语言模型对于大多数常识性问题可以生成相对准确的回答，但对于未在预训练语料中出现的事实，大型语言模型易生成错误甚至完全虚构的回答，即所谓的“幻觉”问题。因此，为大型语言模型赋予专业知识，提升其在领域问答系统中的表现，是当前的关键挑战。

为了应对“幻觉”问题，支持检索增强生成（Retrieval Augmented Generation，RAG）技术的大型语言模型在回答问题前，先会从大量文档中检索出上下文提示，基于上下文提示生成回答，以提高回答的质量。目前多数RAG系统使用向量数据库作为单一的上下文来源。知识图谱和向量数据库均为信息的组织和表征方式，综合两者对于构建高质量知识问答系统而言十分必要：一方面，知识图谱可对问题涉及的实体关系进行精准刻画，召回与问题相关的“关系链”或“关系网”上的一系列关键实体及其相互关系；另一方面，向量数据库可以在此基础上，检索与这些关键实体高度相关的文本片段，作为大型语言模型生成答案的上下文依据。二者的结合既保证了问题和事实匹配的精确性，又可提升相关事实的召回率，进而提升知识问答系统的综合表现。

基于此，本研究提出一种融合华侨华人知识图谱、向量数据库和大型语言模型的华侨华人知识问答系统架构。如图9所示，该架构的核心组件包含两个基于GPT的代理，分别为“Cypher生成代理”和“向量查询代理”。

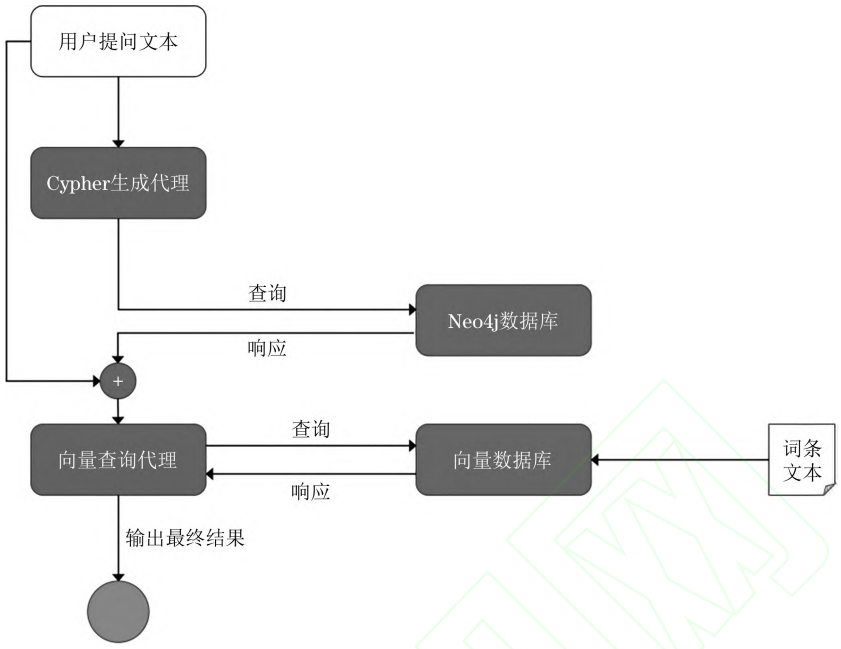

图9　华侨华人知识问答系统的设计思路

Cypher生成代理的主要功能是将自然语言问题转化为面向知识图谱的Cypher查询语句。该代理的提示提供了华侨华人知识图谱中所有的节点标签、关系类型、属性及其含义，同时，基于小样本学习范式，提供少数典型问题到Cypher语句的转换示例。据此，Cypher生成代理将用户提问与华侨华人知识图谱的定义（即节点类型、关系类型、属性等）相匹配，生成对应的Cypher查询语句。该Cypher语句被Python程序通过应用程序编程接口传入Neo4j数据库，进行语义查询，从而得到基于华侨华人知识图谱的语义查询结果。值得注意的是，由于知识图谱中仅包含结构化信息，仅依赖Neo4j语义查询结果难以完全满足用户的信息需求。因此，有必要通过向量查询代理引入向量语义检索机制，补充更加丰富的非结构化背景信息。

向量查询代理负责利用向量数据库，检索与问题相关的上下文片段，并基于上下文片段输出最终回答。具体而言，本研究预先对数据源中所有华侨华人词条的合并文本进行切片，对于每一切片，生成向量表示，并存入向量数据库。向量查询代理工作时，首先将用户提问文本与Neo4j语义查询结果中包含的实体名称、关系名称拼接，形成向量数据库的查询变量，并生成该查询变量的向量表示；然后，通过余弦相似度算法，从向量数据库检索TOP N（本研究中N=10）向量，进而获取这N个向量对应的上下文文本片段。这些文本片段即为向量查询代理用于生成最终结果的上下文提示。

通过两个代理的协同配合，可实现对结构化语义知识和非结构化文本的综合利用。这种融合知识图谱和向量数据库的混合方法，赋予了大型语言模型与问题高度相关的大量领域知识，有助于生成对华侨华人领域问题的高质量回答。

基于上述思路，本研究设计并实现了一个针对华侨华人知识的问答原型系统。该系统整合了知识图谱和向量数据库，旨在回答涉及华侨华人社群关系及其社会活动的相关问题，如特定国家华侨华人的祖籍地分布、华侨华人共同参与的历史性事件，以及华侨华人的投资活动模式等。

表8显示了对于一系列代表性问题，仅使用向量数据库的RAG方法以及结合知识图谱和向量数据库的RAG方法生成答案的差异。通过对领域文献的考证可以发现，仅使用向量数据库作为上下文来源生成的回答仍存在着一定程度的“幻觉”问题，语言流畅性也有所欠缺；综合方法总体而言能够提供更为丰富的上下文信息，生成与事实更为吻合的流畅回答。

6　结语

目前，对华侨华人文献的开发仍处在以数字化为主的阶段，构建华侨华人知识图谱有助于在知识单元层面推动华侨华人领域知识的细粒度开发和利用。本研

表8　两种RAG方法对代表性问题的回答质量对比

代表性问题	仅使用向量数据库	结合知识图谱和向量数据库
与陈嘉庚相关的华侨华人有？	与陈嘉庚相关的华侨华人包括在七七事变后响应陈嘉庚号召组织槟城华侨筹赈祖国难民委员会的侨领之一，他们为主要负责人之一，日夜奔波劝募捐款。此外，1961年8月陈嘉庚在北京逝世后，……	与陈嘉庚相关的华侨华人包括王雨亭、陈敬贤、侯西反、孙崇瑜、李光前和庄西言等。这些人物或是与陈嘉庚有直接的家族关系，如陈敬贤是陈嘉庚的七弟；或是在商业、社会活动中与他有合作，比如李光前不仅是……
印尼华侨华人的投资地有？	印尼华侨华人的投资地包括了多个行业和领域，如纺织业、种植业、木材业、地产、旅游和金融等。在80年代，他们还收购了美国加利福尼亚州的一家中等银行。90年代初期，他们开始在厦门等地进行投资……	印尼华侨华人的投资地包括雅加达、新加坡和厦门等地。他们在这些地方设立了分支机构，经营书刊、侨汇业务以及其他贸易活动。例如，蒋维泰在雅加达等多个城市设立了专营各种书刊及教科书的分店……
印尼较为有名的华侨华人有？	印尼较为有名的华侨华人包括麦少彭，他是一位日本华侨富商，在1880年赴日经营火柴输出贸易，并在梁启超的协助下创办了神户中华同文学校。还有麦汉锦，马来西亚政界领导人……	印尼较为有名的华侨华人包括张国基、曾国正、陈性初、房延凌、吴德疆、吴元盛、汤龙飞、施进卿和俞昌檀等。这些人物在印尼社会中具有一定的影响力，他们可能在商业、教育或社团领导等方面做出了突出贡献。例如……

究提出了一套华侨华人知识图谱构建与应用方法，研究创新之处在于：①将知识图谱技术应用于华侨华人领域文献的知识组织中，构建出具有应用价值的华侨华人知识图谱；②从关系资本、结构资本、认知资本3个维度出发，提出了华侨华人知识图谱的语义检索策略；③将知识图谱与大型语言模型、向量数据库相结合，提出了华侨华人领域知识问答系统的设计方法。

本研究也存在着一定的不足：首先，以华侨华人工具书为语料来源，尚不能完全反映华侨华人的全面特征和多元文化背景；其次，在结构资本的中心性计算和社区划分中，仅考虑家人关系、相识关系等强人物关系，并未考虑同乡、校友等弱关系；最后，华侨华人领域知识广泛多样，且大型语言模型的输出内容具有一定的随机性，这为对问答系统质量的评价带来较大的挑战，通过原型系统实例，尚无法全面评估华侨华人知识问答的质量。在未来的工作中，将尝试引入更多反映华侨华人工作、生活的来源语料，引入弱关系进一步分析华侨华人结构资本维度内容，并采用更为系统化的评价框架来评估与优化华侨华人问答系统性能。

参考文献

[1] 中华全国归国华侨联合会. 国务院侨务办公室关于印发《关于界定华侨外籍华人归侨侨眷身份的规定》的通知[EB/OL]. [2023-09-26]. http://www.chinaql.org/n1/2019/0322/c420275-30990528.html.

[2] 周南京. 世界华侨华人词典[M]. 北京：北京大学出版社，1993.

[3] 周南京，黄昆章. 华侨华人百科全书[M]. 北京：中国华侨出版社，1999.

[4] 陈奭琛. “申遗”成功后泉州侨批文献的保护和利用[J]. 泉州师范学院学报，2014，32（4）：117-120.

[5] 陈苑琼. 广东梅州客家侨批传承保护探讨[J]. 档案学研究，2018（2）：85-88.

[6] 宫毅敏. 传统文化创新视角下侨批档案开发利用与对策研究[J]. 山西档案，2019（5）：112-119.

[7] 张亚兵. 侨批档案资源开发利用存在问题及对策研究[D]. 郑州：郑州大学，2021.

[8] 陈初昇，李丹阳，李楚薇，等. 华侨华人网络、企业战略激进度与对外直接投资[J]. 华侨大学学报（哲学社会科学版），2023（4）：83-99.

[9] LESKINEN P, HYVÖNEN E. Reconciling and using historical person registers as linked open data in the AcademySampo portal and data service[C]//International Semantic Web Conference. Cham: Springer, 2021: 714-730.

[10] 徐彤阳，黄映思. 名人年谱资源的知识图谱构建：以徐朔方《晚明曲家年谱》为例[J]. 数字图书馆论坛，2022（12）：29-36.

[11] 沈雪莹，欧石燕，卢彤彤. 中国古代文人生平知识图谱构建与应用：以李白和杜甫为例[J]. 数字图书馆论坛，2023，19（8）：1-14.

[12] 张强，高颖，刘飞，等. 基于知识重组的红色历史人物智能服务研究[J]. 现代情报，2023，43（7）：96-108.

[13] 程结晶，王璞钰. 古籍中人物史料的关联组织研究：以《汉书·艺文志》中西汉经学家群体为例[J]. 图书馆论坛，2023，43（3）：64-74.

[14] 周济南，罗依平. 城市社区合作治理失灵的矫正：一个社会资本理论的分析框架[J]. 湖湘论坛，2021，34（4）：118-128.

[15] 杜晶晶，王涛，郝喜玲，等. 数字生态系统中创业机会的形成与发展：基于社会资本理论的探究[J]. 心理科学进展，2022，30（6）：1205-1215.

[16] 付少雄，朱梦蝶，郑德俊，等. 基于社会资本理论的在线医疗

社区医生知识贡献行为动因研究[J]. 情报资料工作，2022，43（3）：67-74.

[17] FIELD J. Social capital[M]. London：Taylor & Francis，2016：3.

[18] 张广利，陈仕中. 社会资本理论发展的瓶颈：定义及测量问题探讨[J]. 社会科学研究，2006（2）：102-106.

[19] 卢燕平. 社会资本的来源及测量[J]. 求索，2007（5）：5-8.

[20] CAREY S，LAWSON B，KRAUSE D R. Social capital configuration，legal bonds and performance in buyer-supplier relationships[J]. Journal of Operations Management，2011，29（4）：277-288.

[21] Information Resources Management Association. Operations and service management：concepts，methodologies，tools，and applications[M]. Hershey：IGI Global，2017：32.

[22] SUN Y，WANG S H，FENG S K，et al. ERNIE 3.0：large-scale knowledge enhanced pre-training for language understanding and generation[EB/OL]. [2023-04-10]. https://arxiv.org/abs/2107.02137.

[23] Microsoft. Geodata API[EB/OL]. [2023-07-23]. https://learn.microsoft.com/en-us/bingmaps/spatial-data-services/geodata-api.

[24] LU H，HALAPPANAVAR M，KALYANARAMAN A. Parallel heuristics for scalable community detection[J]. Parallel Computing，2015，47：19-37.

[25] NeoMap[EB/OL]. [2023-08-01]. https://github.com/stellasia/neomap.

作者简介

洪宝惜，女，硕士研究生，研究方向：知识图谱。
林泽斐，男，博士，副教授，通信作者，研究方向：知识图谱、数字人文，E-mail：linzf@fjnu.edu.cn。

Construction and Application of Knowledge Graph for Overseas Chinese Based on Social Capital Theory

HONG BaoXi　LIN ZeFei
(College of Sociology and History, Fujian Normal University, Fuzhou 350117, P. R. China)

Abstract: The overseas Chinese community is a vital force in driving modernization and national rejuvenation in our country. Grounded in social capital theory, this paper proposes the construction and application methods of the knowledge graph for overseas Chinese, providing data support for knowledge services aimed at this population. First, by identifying the sources of social capital, we analyze relevant concepts in the overseas Chinese domain and engage in ontological modeling. Next, knowledge units are extracted from the literature pertaining to overseas Chinese and stored in the Neo4j graph database. Finally, semantic search and question-answering strategies for the overseas Chinese knowledge graph are designed from the three dimensions of social capital theory. The overseas Chinese knowledge graph designed based on the sources of social capital reveals the complex and diverse relationships between knowledge domains of overseas Chinese. The semantic search and question-answering system developed from the three dimensions of social capital theory enables the fine-grained and multi-dimensional presentation of knowledge in the field of overseas Chinese.

Keywords: Social Capital Theory; Overseas Chinese; Knowledge Graph

（责任编辑：王玮）

待作者解决的问题：

1）图7和图8模糊，请更换清晰的图片