DOI:10.12154/j.qbzlgz.2021.06.010

面向非遗文本的知识组织模式及人文图谱构建研究*

张 卫[1,2] 王 昊[1,2] 李跃艳[1,2] 邓三鸿[1,2]
([1]南京大学信息管理学院 江苏 210023; [2]江苏省数据工程与知识服务重点实验室(南京大学) 南京 210023)

摘 要: **[目的/意义]**目前,非物质文化遗产领域内细粒度的人文性知识(情感、观念、思维、风格等)广泛散布在多源异构的非结构化文本中,尚未得到有效组织,如何设计一套由大规模非遗文本向人文性知识自动转化的模式,对于人文知识图谱的构建及其应用具有重要指导作用。**[方法/过程]**本文面向非遗非结构化文本探索领域知识组织模式,重点结合知识图谱在文本语义关联解析与数据语义链接层面的技术特质,一方面引入自然语言处理技术对非遗文本内人文性知识进行语义关联解析,另一方面在语义融合的基础上推动数据的语义链接与知识服务。随后,以联合国非遗名录中的"古琴艺术"为案例进行实现路径分析,包括:元数据特征解析与人文本体语义建模、"冷启动"下融入汉字语言特征的文本语义关联解析、语义融合下多源知识的语义链接、基于语义知识图谱的非遗人文知识服务。**[结果/结论]**本文提出了非遗非结构化文本到结构化知识至开放共享知识库的一整套知识组织模式与自动化实现路径,语义技术的有效利用细化了领域知识的粒度,知识的语义描述与链接提升了数据关联的范畴,进而为更深层次的非遗人文知识服务打下基础。

关键词: 非物质文化遗产 知识组织 语义知识图谱 人文性知识 自然语言处理

Research on Intangible Cultural Heritage Text-Oriented Knowledge Organization Model and Humanistic Atlas Construction

Zhang Wei[1,2] Wang Hao[1,2] Li Yueyan[1,2] Deng Sanhong[1,2]
([1]School of Information Management of Nanjing University, Jiangsu, 210023;
[2]Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing University, Nanjing, 210023)

Abstract: **[Purpose/significance]** At present, the fine-grained humanistic knowledge (emotions, concepts, thinking, styles, etc.) in the field of intangible cultural heritage is widely scattered in unstructured texts of multiple sources, and has not been effectively organized. How to design a model for automatic transformation from large-scale intangible heritage texts to humanistic knowledge is an important guide for the construction of humanistic knowledge graphand its application.**[Method/process]** In this paper, we explore the domain knowledge organization model for unstructured ICH texts, focusing on the technical characteristics of knowledge graphs in semantic association parsing of texts and semantic data linking, introducing natural language processing techniques for semantic association parsing of humanistic knowledge in ICH texts, and promoting semantic data linking and knowledge services based on semantic fusion.A technical path is demonstrated by taking "Guqin Art" in the ICH list of the United Nations as an example, including: metadata feature analysis and humanistic ontology modeling, semantic association parsing of text incorporating linguistic features of Chinese characters under "cold start", multi-source knowledge link under semantic fusion, and deduction and knowledge discovery basing knowledge ontology. **[Result/conclusion]**We aim at constructing a complete set of knowledge organization model and automated realization path from unstructured text to structured knowledge and open shared knowledge base. The effective use of semantic technology refines the granularity of domain knowledge, and the semantic description and association of knowledge enhance the scope of data linking, thus laying the foundation for deeper ICH humanistic knowledge services.

Keywords: intangible cultural heritage knowledge organization semantic knowledge graph humanistic knowledge natural language processing

*本文系国家社科基金重点项目"大数据环境下领域知识加工与组织模式研究"(项目编号:20ATQ006)、江苏省研究生科研创新计划"面向心理健康的医学文本语义解析与知识图谱构建研究"(项目编号:KYCX21_0026)和中央高校基本科研项目"面向人文计算的方志文本的语义分析和知识图谱研究"(项目编号:010814370113)的研究成果,并受江苏青年社科英才和南京大学仲英青年学者等人才培养计划的支持。

1 引言

非物质文化遗产(简称“非遗”),是一种濒临消失的、以非物质形态存在的传统文化,被公认为是促进世界文明发展与激发人类创造力的宝贵财富。目前,我国非遗文本涉及领域之广,蕴含知识之丰富,难以通过现有主题、地域、级别等[1]外部粗粒度知识组织方式进行有效管理。因此,当下亟待突破的关键问题便是从语义关联的视角对非遗信息资源进行深层次的知识组织与应用。

知识图谱是一种利用图数据结构进行知识表示、组织与管理的载体,在各方场景下应用的侧重不一。一方面,学界多聚焦于对已有知识的语义描述与链接[2],旨在实现关联数据在网络标准下的语义发布与开放共享,故有学者称之为语义知识图谱[3]。目前,非遗领域的知识组织工作多围绕其展开。然而,非遗知识的语义链接往往诉诸已有的“传承人”“地点”“机构”等常规知识,未能深入大规模非遗文本内部,解析文化对象的所含情感、所系风格、所表态度、所述观念等具有人文内涵的知识,这些细粒度的人文性知识散布在各类非遗平台、百科网页、研究文献等多源异构的非结构化文本中,亟待语义层面的知识组织。另一方面,工业界的知识图谱多关注非结构化文本中结构化知识的自动提取以及图存储[4],主要借助实体识别、关系抽取等自然语言处理技术对文本进行语义关联解析获取细粒度的领域知识。不过,此类方法在当下非遗知识组织中的应用尚不深入,知识图谱建构的可行性和准确性均有待完善。

因此,本文将结合知识图谱在语义关联解析与语义链接共享两方面的技术特质,针对零散、碎片化的非遗文本,深入信息资源内部探索细粒度的人文知识的解析方法及其优化方案,以人文知识图谱为导向构建非结构化文本到结构化知识至可开放共享知识库的一整套非遗知识组织模式与技术实现路径,进而推进非遗领域知识的语义链接共享与语义知识服务。

2 相关研究工作

近年来,非遗的知识组织工作正从传统主题、地域、级别等外部粗粒度方式趋向于信息资源内部特征的整合,主要包括:(1)本体理论。滕春娥等[5]通过统计层级关系和添加资源实例组织赫哲族非遗知识资源。(2)主题图技术。王蒙等[6]对京剧、昆曲资源进行组织以构建领域主题图模型。(3)关联数据。翟姗姗[7]利用关联数据完成了非遗文本中实体的RDF化、关联、存储与可视化,由于其能够有效实现数据集间的语义链接和共享,故逐渐成为学界的研究热点。然而,关联数据在非遗知识组织中的应用多侧重于对已有数据集(非遗平台等)或非遗地点、机构、传承人等常规知识进行语义描述和开放,而对于非遗文化所承载的风格、态度、情感等人文性知识的组织,目前尚未形成较为完善的解决方案。

随着人工智能的兴起,知识图谱在关联数据的基础上发展而来,其以“实体-关系-实体”通用三元组为基本单位,重点关注对非结构化文本进行语义关联解析,进而将细粒度的结构化知识存储形成图数据库[8],以工业界应用居多。为作区分,领域学者将图数据库命名为广义知识图谱,关联数据为语义知识图谱[3],前者侧重非结构化文本到结构化知识的自动化流程[4],后者多考虑结构化知识在数据集间的语义链接与开放共享[2]。因此,领域文本的语义关联解析可有效推动非遗语义知识图谱对人文性知识的获取。

在知识图谱的技术体系中,文本语义关联解析主要涉及实体识别与关系抽取两大关键性自然语言处理技术[9]。实体识别是指从文本当中获取关键词、命名实体、术语等,而关系抽取旨在提取出实体间的语义关系,包括分类及非分类关系,演化路径如图1所示。

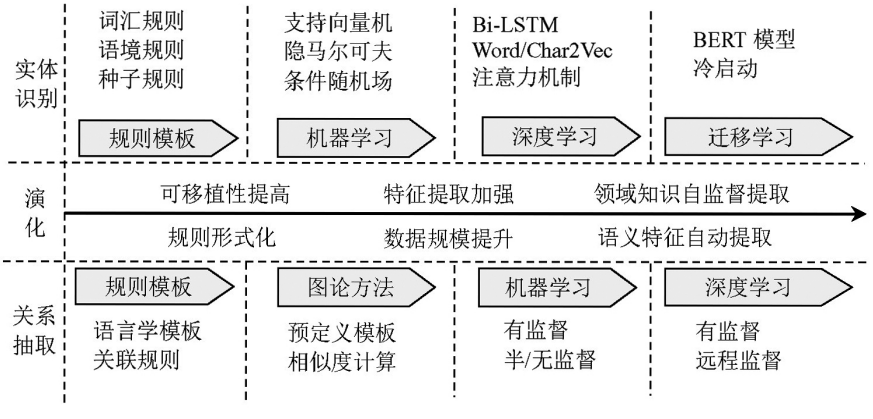

图1 文本语义关联解析关键技术的演化路径

由图1可知，文本语义关联解析关键技术的演化路径展现了实体识别与关系抽取从模板化至自动化、语义化的发展历程。早期，主要利用模式匹配标记文本中各类实体[10]，通过词形、句法、依存分析结合的语言学模板[11]抽取实体关系。然而，考虑到规则模板在建设周期以及移植性上的不足，条件随机场[12]等有监督机器学习的算法渐受推崇，旨在将实体及实体关系的抽取转换为序列标注或分类任务[13]。近年来，深度学习的兴起弥补了特征工程难以捕获数据深层语义特征的缺陷。其中，字标注模型的字向量映射（Char2Vec）特别适用于中文文本的语义关联解析[14]，驱使学者们在非遗领域进行尝试。任秋彤等[15]针对非遗传统戏剧提出一种效果更优的术语抽取模型，以BERT-BiLSTM-CRF模型为基础加入图卷积网络更好地捕捉句子中远距离词语的约束关系。Dou等[16]采用ID-CNNs-CRF模型识别出我国十二节气文本中的人物、事件、时间和地点等实体资源，并进一步通过Bi-GRU模型抽取出非遗实体之间的关系以构建领域知识图谱。然而，领域学习任务需要攻克在无或少学习语料下进行模型训练的“冷启动”问题以及领域文本特征的提取问题。因此，非遗文本语义关联解析的可行性与准确性均有待优化与完善。

综合来看，现有的非遗知识组织研究尚存在以下三点不足：其一，领域学者虽然趋向于采取知识图谱来对非遗信息资源进行组织，然而在语义关联解析与数据语义链接两方面侧重不一，尚未形成一套完整的非遗语义知识组织体系；其二，非遗知识获取仍然是构建领域知识图谱的根本基础，而领域研究对细粒度人文性知识的挖掘尚不全面，对关键自然语言处理技术的使用尚不成熟；其三，目前越来越多的非遗知识图谱得到了开发，但对非遗知识图谱应用与服务模式的探索还有待深化。因此，本文将综合语义关联解析与数据语义链接两方面的特质设计非遗知识组织模式，重点探索以人文性知识为核心的非遗文本语义解析及其优化方案，以期推动非遗知识图谱的语义链接与知识服务。

3 我国非遗知识组织模式设计与技术分析

非遗领域知识图谱旨在实现人文性知识的链接共享与智能应用。因此，非遗知识组织模式的关键步骤是通过自然语言处理技术的引入与优化，对领域文本中具有人文内涵的知识（人文情感、风格、思维等）进行语义关联解析，并在此基础上完成非遗数据的语义链接与知识服务。其中，相较于传统非遗知识图谱，本模式的自动实现路径旨在知识获取、融合、应用等核心环节内通过语义技术的应用提升知识组织的范畴与效率。此外，部分环节也将保留必要的人工知识进行调整及优化，如图2所示。

由图2可知，本模式以我国非遗文本为语料，将其中细粒度、具有人文内涵的知识囊括入非遗知识本体的语义建模范畴开展文本语义关联解析，将获取的知识与已有（半）结构化知识相整合、清洗，通过链接外部关联数据集实现我国非遗语义知识图谱的构建及开放共享，进而深入非遗的人文内涵与意喻以实现语义知识服务，具体包括四个方面的内容：

（1）非遗文本特征解析，旨在基于非遗文本资源中特有的人文性特征进行本体建模。将非遗文本分类形成结构化、半结构化与非结构化三种形式，并着重针对后两种文本，参照语义网中现有的元数据标准或规范词表，解析非遗文本中细粒度的人文特征，结合领域专家知识构建顶层的人文本体模型。

（2）非遗人文性知识获取，旨在通过自然语言处理技术从非结构化的非遗文本中获取细粒度的领域知

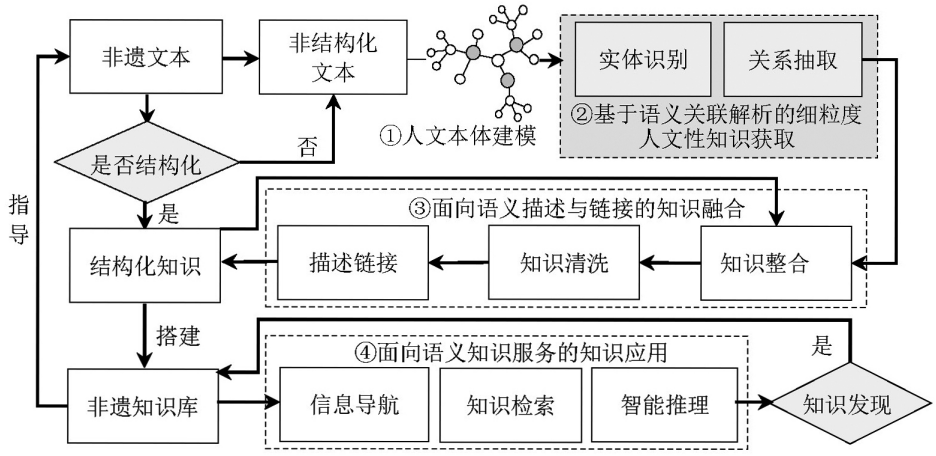

图2 我国非遗文本知识组织模式的设计方案

识。以本体中定义的类、数据/对象属性为指导，通过领域特征的提取与各类语义解析技术的优化识别出非遗文本中实体的分类及非分类关系，并着重探讨非遗文化的所表情感、所含风格、所系来源、所述典故等人文性知识的获取，形成非遗人文知识图谱的资源基础。

(3)非遗知识融合，旨在实现非遗人文性知识的整合、清洗以及语义映射链接。首先，将语义关联所解析出的实体及实体关系与结构化知识相融合形成关系型数据库，随后按照分类与非分类对实体关系进行梳理，对于尚未建立关系的实体进行知识整合，对存在错位的关系进行知识清洗，并利用本体描述语言进行RDF语义映射与跨数据集语义链接，进而存储为可开放共享的非遗语义知识库。

(4)非遗知识应用，旨在从信息行为角度发掘用户对非遗知识库的应用模式，挖掘具有潜在语义关联价值的领域知识，包括：①信息导航，通过RDF链接为用户提供不同数据源间的访问导航；②知识检索，利用SPARQL查询语言从关联数据中检索出知识资源；③智能推理，制定推理规则发掘人文知识间的关联，推动领域知识发现以回馈知识库，进而指导非遗文本知识组织的增量迭代。

为了说明所提出非遗知识组织模式的可行性与合理性，本文以我国被列入人类非遗名录的“古琴艺术”为案例，按照图1中语义建模、知识获取、知识融合、知识应用的技术流程进行论证分析。

3.1 非遗文本元数据特征解析与本体语义建模

传统非遗知识本体大多以非遗项目、传承人、传承机构、地理位置等常规知识为基础进行建模，对此本模式将重点关注领域非结构化文本中可供语义关联的人文性知识实体。

(1)非遗文本元数据特征解析。旨在将领域资源中归纳、提取的细粒度人文语义特征以元数据的形式融合为初步知识框架，继而指导后续领域本体构建，如图3所示。

由图3可知，该过程旨在将非遗文本中具有人文内涵的特征线索进行系统归纳。首先，采集多源非遗文本，考虑到此阶段采集的文本资源除本体建模外还将用于后续模型训练，而主流的深度学习算法需要大规模数据的训练才能取得较好的效果，故在国家级/省级非遗网站基础上将语料渠道扩展至公共服务平台、百科网站、学术文献等获取更大规模非遗文本，并经清洗、汇总后形成资源库；随后，分析并归纳文本内容中实体及实体关系，提炼成为核心人文元数据，如文学流派、收录琴谱、蕴含情感、古琴风格等。接着，将文本中提取的所有元数据(人物、时间、琴曲等)融合形成领域知识框架，进而展现并洞悉框架内部知识结构，以期利用该框架指导后续文本的本体建模与语义关联解析。

本文根据古琴艺术知识构成的核心要素，综合国家非遗网站内的类目体系与非结构文本中所解析的人文元数据，抽象出12个核心类作为本体的基础知识概念(见表1)。

(2)非遗本体语义建模。旨在将类间逻辑关系系统化地组织起来，核心在于规范非遗人文知识概念及概念关系，使其得到有效描述和揭示。为此，本模式在古琴艺术知识概念的基础上，通过专家参与审核，建立

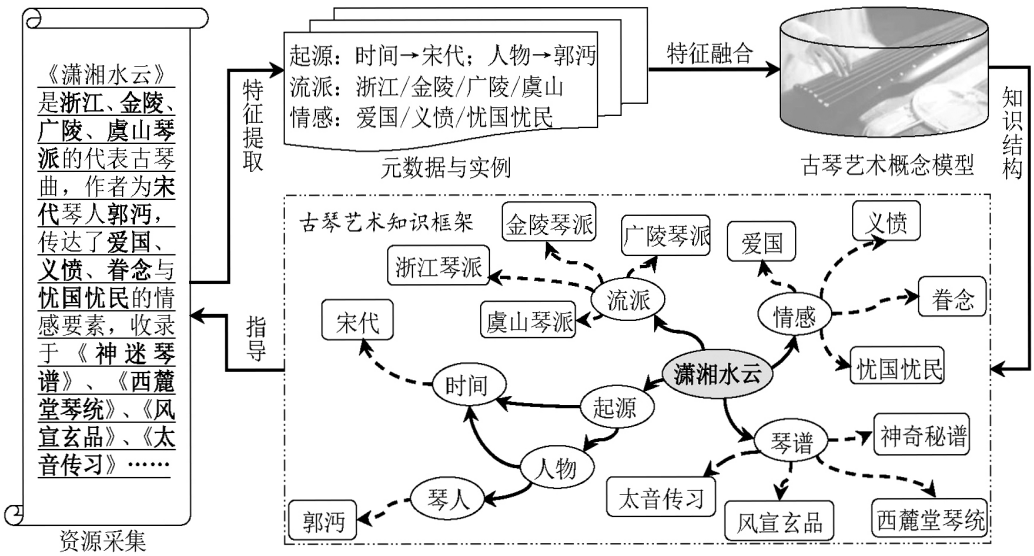

图3　非遗文本的元数据特征解析(以古琴艺术为例)

表1 古琴艺术领域基础类及其定义说明

中文类描述	英文类描述	复用标准	定义	概念来源
流派	mo: Genre	mo: Genre	古琴艺术流派	国家非遗平台
地点	dcterms: Location	dcterms: Location	古琴领域事物的相关地点	国家非遗平台
人物	foaf: Person	foaf: Person	古琴领域的相关人物	文本语义解析
时间	crm: Time	crm: Time	古琴领域事物的相关时间	文本语义解析
琴谱	crm: Collection	crm: Collection	收录古琴曲的琴谱	文本语义解析
琴人	mo: Composer	mo: Composer	古琴曲的作者	文本语义解析
琴曲	mo: Composition	mo: Composition	古琴曲的名称	文本语义解析
情感	kdo: Sentiment	kdo: Sentiment	古琴曲蕴含的情感	文本语义解析
风格	guqin: Style	自定义类	古琴流派的演奏风格	文本语义解析
创始人	guqin: Founder	自定义类	古琴流派的创始人	文本语义解析
传承人	guqin: Inheritor	自定义类	古琴流派的传承人	国家非遗平台
机构	guqin: Institution	自定义类	古琴流派的传承机构	国家非遗平台

类间逻辑关系标准以形成本体概念模型，如图4所示。

由图4可知，古琴艺术本体参考了FOAF、Dublin Core Terms、CRM、MO与KDO等本体和元数据标准，通过“guqin”表示古琴艺术本体命名空间。其中定义数据属性为类名，并着重设计11个类间对象属性用于描述实体之间的语义关系，如表2所示。

由表2可知，通过在本体模型中定义对象属性及其定义域与值域，一方面为后续开展的非遗实体及实体关系的自动化抽取建立了标准，以实现领域常规知识与人文性知识的语义关联；另一方面，在此基础上可通过自定义推理规则实现语义关系的推理进而促成隐性的人文知识发现。

表2 古琴艺术领域类间对象属性及其定义说明

对象属性	复用标准	定义域	值域	定义
kdo: hasSentiment	kdo: hasSentiment	mo: Composition	kdo: Sentiment	某古琴曲蕴含某情感
guqin: hasOriginTime	自定义属性	mo: Genre	crm: Time	某流派起源于某时代
guqin: hasLocation	自定义属性	mo: Genre	dcterms: Location	某流派传承于某地点
guqin:hasInstitution	自定义属性	mo: Genre	guqin: Institution	某流派受某单位保护
guqin: hasStyle	自定义属性	mo: Genre	guqin: Style	某流派具有某类风格
guqin: hasPerson	自定义属性	mo: Genre	foaf: Person	某流派具有某些人物
guqin: hasComposition	自定义属性	mo: Genre	mo: Composition	某流派包含某些琴曲
guqin: hasGenre	自定义属性	mo: Composition	mo: Genre	某琴曲属于某些流派
guqin: hasComposeTime	自定义属性	mo: Composition	crm: Time	某琴曲创作于某时代
guqin: hasComposer	自定义属性	mo: Composition	mo: Composer	某琴曲由某人所创作
guqin: hasCollection	自定义属性	mo: Composition	crm: Collection	某琴曲收录于某琴谱

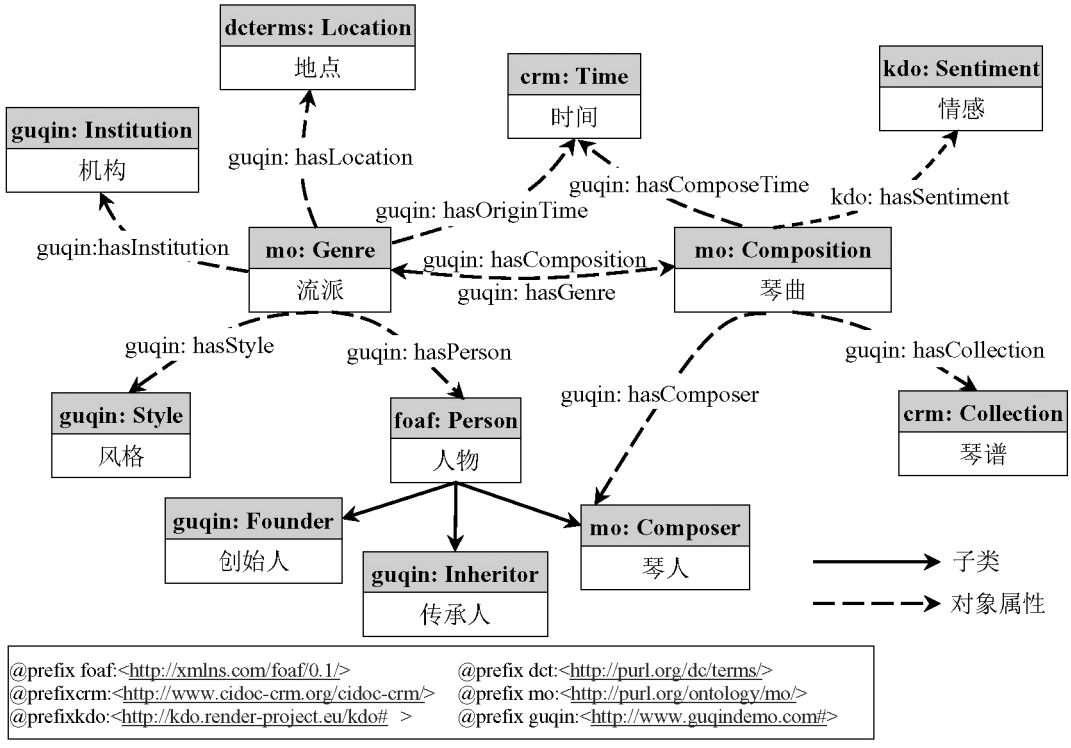

图4 非遗人文知识本体概念模型(以古琴艺术为例)

3.2 基于关键技术的非遗文本语义关联解析

针对已有非遗知识组织研究中对自然语言处理技术应用不深入的问题，本模式重点对非结构化文本中细粒度的非遗人文实体及实体关系进行语义关联解析。其中，本体中所定义的类将用于指导文本中实例（体）的识别，而类间属性将用于支撑实例（体）间关系或属性的抽取。

（1）模式匹配与序列标注相结合的非遗实体识别方法。非遗文本在不同情形下的实体识别方法不尽相同：对具有明显标注符号的文本可诉诸人工制定规则模板，通过模式匹配从文本中自动匹配字符串，标记为琴人、琴曲、琴谱、情感等各类实体，例如琴谱可用“《》”进行匹配；对连续性较强且不具备特定分隔符的文本可转化为序列标注模型[17]，通过数据的序列标注S（字）、B（词首）、M（词中）、E（词尾）、O（词外）获取学习语料，训练领域分类器预测文本序列以识别实体，是获取大规模非遗实体的核心方法。本模式基于谷歌BERT-base预训练模型[18]，在网络层数为12，隐藏层维度为768，多头个数为12，110M参数BERT上进行微调训练实体识别模型。其中，为保证模型学习效果，基于前文所述渠道采集相应规模的数据，采取“。”“？”“！”“；”等中文标点作为断句标准，控制断句后文本序列长度阈值为128个汉字且句子数量在万级以上[19]以适应BERT使用。考虑到获取的非遗文本尚缺乏大规模标注数据，本模式试图引入“冷启动”技术从相关领域迁移标注数据或知识获取学习语料，并通过领域汉字语言特征的优化提升模型的识别精度，进而优化非遗实体识别的效果。

对于古琴曲、情感、风格等人文实体，如琴曲（composition, cn）“潇/B-cn 湘/M-cn 水/M-cn 云/E-cn”，可通过领域词典来自动标注文本序列以获得学习语料，在此基础上引入古琴文本内的汉字语言特征构建实体识别模型。以BERT-BiLSTM-Attention-CRFs为例，BERT层是核心模块，抽取768维字向量（Char2Vec）充分发挥汉字语言知识，重在解决词语错分造成的误差传递与一词多义问题；随后，将领域文本内总结的语言特征（汉字部首、拼音、字素特征等）通过One-Hot编码的形式扩展到BERT语义向量之后，如实体“潇湘水云”的前两个字的部首均为三点水，便可通过该部首实现769维（768+1）的增强字向量；其次，BiLSTM层对BERT层输入的字向量进行上下文信息编码与网络参数训练；再次，注意力机制用于扩展BiLSTM层编码结果的上下文信息，并输入CRFs层语义向量使其解码后输出最终类别。在完成序列标注之后，通过序列标签向领域实体的映射抽取出文本中的古琴实体“潇湘水云”。其中，将学习语料按照4:1的比例分配形成训练集和测试集，评价指标选取召回率(R)、精确率(P)和F_1值。

（2）对于非遗领域现有或识别出的实体，可采取结构与内容相融合的方法抽取其关系。在模式匹配辅助下，从行文结构和语义内容两个角度实现非遗实体关系抽取。对于分类关系，可在基于文本内容聚类形成分类体系总体框架的基础上，附以基于实体结构共现的形式概念分析方法，实现分类关系的修正和优化；对于非分类关系，则基于语言片段结构和概念模型语义定义关系类型，一方面利用本体概念模型中定义的语义关系对基于上下文结构抽取的关系进行标准化和规范化，另一方面利用基于文本结构抽取的关系丰富概念模型中的语义关系，实现非分类关系的优化和完善。

对于非遗领域未知或尚未识别的实体，可基于联合学习模型同时抽取实体及实体关系，这不仅能达到更好的整合效果，亦可兼顾分类与非分类关系。为此，本文提出序列标注方案，基于（1）中断句后的文本，利用远程监督对齐知识库（维基/百度百科）与文本内的关系获取学习语料，在此基础上基于中文语言特征与深度学习算法构建端到端模型以实现非遗实体关系的联合抽取[20]，如图5所示。

在图5中，依据本体模型定义的类间对象属性规定实体关系标签方案，包含四个部分：实体位置角色、实体类型、关系类型以及关系角色。其中前二者与实体识别一致，关系类型从预定义的关系集合中获取，关系角色使用“1”和“2”表示，最终抽取结果用三元组(Entity1,RelationType,Entity2)表示。如琴曲（composition, cn）“潇/B-cn 湘/M-cn 水/M-cn 云/E-cn”由琴人（Composer, cr）“郭/B-cr 沔/E-cr”所作（hasComposer, HCR），故可抽取（潇湘水云1, HCR, 郭沔2）；同时该琴曲蕴含（hasSentiment, HST）情感（sentiment, st）“爱/B-st 国/E-st”，亦可抽取出（潇湘水云1, HST, 爱国2）。此时，琴曲（cn）参与了不同类型的关系（HCT/HST），故将其关系类型标注为“D”以作区分[21]。

在此标签方案基础上构建关系抽取模型。首先，利用BERT预训练模型微调训练字嵌入表示，随后总结实体间距离、关系等特征进行Char2Vec增强[15]，如“潇

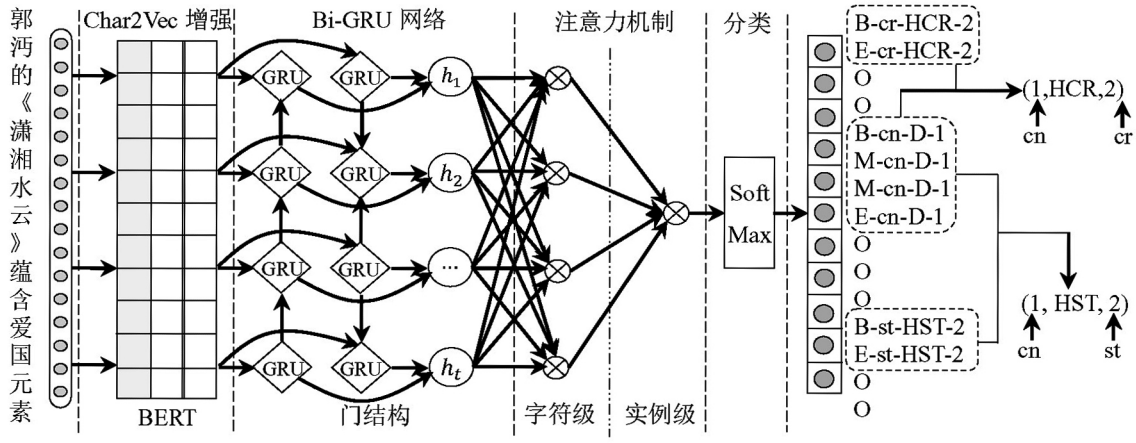

图5 非遗文本实体关系抽取的技术解决方案(以古琴艺术为例)

湘水云”是名词,“爱国”为形容词,二者的先后顺序反映了形容词对名词的解释或修饰作用,故可利用One-Hot编码提取n维语言特征,将其融入原向量形成增强语义表示(768+n),以帮助模型学习深层语义特征;随后,可通过Bi-GRU、Bi-LSTM、CNN等基线模型对比神经网络在特征学习中的性能差异,并使用字符集、实例级多层注意力机制进行优化配置,一方面自动关注对关系抽取起决定作用的汉字,克服单汉字语义信息对于领域任务影响程度的差异性;另一方面纠正语义错位,避免远程标注的数据携带过多噪声。最后,通过关系分类输出非遗实体关系。其中,数据集分配及评价指标与实体识别一致,网络参数可根据经验值设置(如batchsize=32,epoch=30,hidden dim=256,lr=1e-5,dropout=0.1)[22-23],随后在实验中加以验证及优化。

3.3 融合实体及实体关系的非遗知识图谱语义描述与链接

目前对非遗关联数据方面的探索大多诉诸领域已有的结构化数据集,无须深入处理便可投入链接。本模式的知识主要通过自然语言处理技术自动化抽取获得,需要在融合非遗领域的实体及实体关系基础上,进一步整合与清洗领域知识来保证关联数据的语义完整性、精确性,最终通过非遗知识的RDF语义描述与链接,形成具有逻辑概念体系与丰富实例资源的人文知识图谱,如图6所示。

由图6可知,知识获取源于(半)结构化文本中常规知识的整理与非结构化文本中人文性知识的语义关联解析。随后,将不同数据集中的知识进行语义整合、清洗形成关系型数据库,通过语义描述与链接等形成可开放共享的非遗关联数据集,最后展示语义知识图谱并开展应用。其中,非遗知识融合旨在将多源数据集中在同一语义标准下进行关联与链接,主要涉及三个方面的内容。

(1)实体关系整合。非遗实体关系的整合旨在补充并发掘实体关联,将非结构化文本中解析出的实体

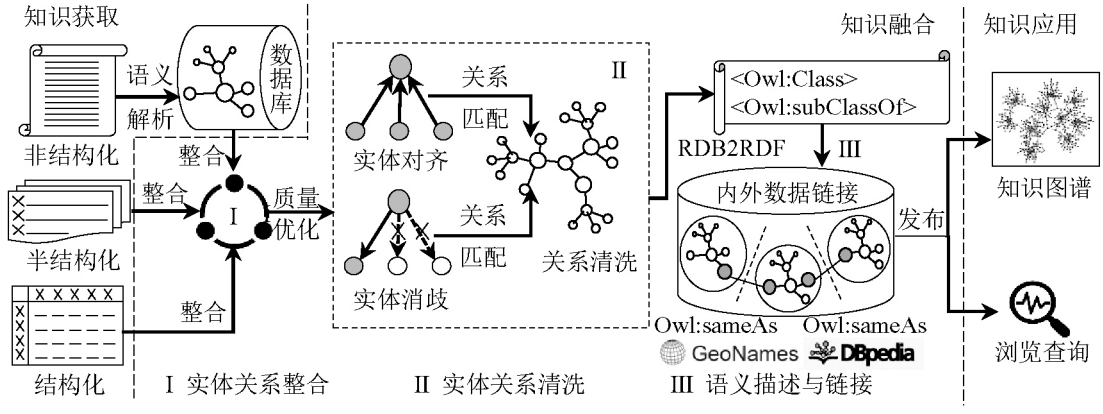

图6 融合实体及实体关系的非遗知识图谱的语义描述与链接

关系同(半)结构化知识相互融合,提高实体关系的完整性。例如国家非遗平台陈列了以古琴艺术流派为核心的结构化数据,包含其与项目类别、传承地点、保护单位等实例的关系;而流派、琴曲、琴谱、人物、情感等实体与实体关系可从非结构化文本中解析得到。因此,通过不同数据集间古琴流派的连接可以实现知识的有机融合。其中,对于一致的实体可采取字符串匹配的方式;对于不一致的实体,可通过共现或表示学习构建特征矩阵,进而采取余弦算法计算实体间相似度,选择相似度前e个实体作为候选连接目标(取值由具体实验设定),经领域专家验证后确定连接节点实现数据库内知识连通。

(2)实体关系清洗。旨在通过实体对齐与消歧技术实现实体关系的正确匹配,以解决领域知识重复、不对应与关联不够明确等问题,进而提高实体关系间的准确性。①实体对齐,旨在解决关联数据集中实体间多词一义的问题。如实体"郭沔""郭楚望",二者文本字符虽不一致,但均是指浙江琴派鼻祖郭沔。对此,可结合传统对齐方法与主流表示学习方法,在融入领域专家知识的基础上通过相似度计算或关系训练确定对齐的领域实体。②实体消歧,意在解决数据集中实体的一词多义问题。如对实体"梅花三弄",既可以表示白琼二胡曲,亦可指向古琴曲,对此可以采取聚类消歧/知识库链接等方法,对于给定指称项M及其链接候选$L=(l_1,l_2,\cdots,l_n)$,计算实体指称项和目标实体间一致性打分$Score(l,M)$,将得分排序前e的实体作为候选链接目标,经审核后选择正确实体完成消歧工作。

(3)语义描述与链接。旨在将经过整合、清洗等工序的非遗知识进行RDF语义映射、跨数据集链接以及关联数据发布。首先,存储非遗实体及实体关系形成关系型数据库RDB,用RDF数据模型描述为其生成HTTPURI及其RDF描述文档。其中,给予每个非遗实体永久唯一的标识符URI(域名+非遗类型+实体名称),保证URI地址能够被遵循HTTP协议的客户端应用程序所解析。而RDB2RDF现多采取D2RQ的MappingLanguage实现语义映射,即通过d2rq:ClassMaps和d2rq:PropertyBridges将数据表、行、列、值映射为RDF数据中的类、资源、属性值等。然而,此方式各阶段较为封闭,可操作性有待优化。本模式基于OWL语言结构利用计算机自动编码,在底层更直接地汇编关联数据集,将不同标准的数据转换成遵循统一标准的结构化数据。如表3中的语法描述了领域文本"《潇湘水云》蕴含爱国情感元素"的三元组形式:(潇湘水云,hasSentiment,爱国)。

表3　基于OWL语法结构的非遗语义知识自动存储(以古琴艺术为例)

◆《潇湘水云》蕴含爱国情感元素(http://www.guqindemo.com#)

```
<!--定义类-->
<Declaration>
    <Class IRI="#情感"/>
</Declaration>
<Declaration>
    <Class IRI="#琴曲"/>
</Declaration>

<!--定义类间对象属性-->
<ObjectPropertyDomain>
    <ObjectProperty IRI="#hasSentiment"/>
    <Class IRI="#琴曲"/>
</ObjectPropertyDomain>
<ObjectPropertyRange>
    <ObjectProperty IRI="#hasSentiment"/>
    <Class IRI="#情感"/>
</ObjectPropertyRange>
```

```
<!--建立类与实例间的关联-->
<ClassAssertion>
    <Class IRI="#琴曲"/>
    <NamedIndividual IRI="#潇湘水云"/>
</ClassAssertion>
<ClassAssertion>
    <Class IRI="#情感"/>
    <NamedIndividual IRI="#爱国"/>
</ClassAssertion>

<!--建立实例与实例间的关联-->
<ObjectPropertyAssertion>
    <ObjectProperty IRI="#hasSentiment"/>
    <NamedIndividual IRI="#潇湘水云"/>
    <NamedIndividual IRI="#爱国"/>
</ObjectPropertyAssertion>
```

随后,通过对内外部开放数据集建立语义链接,以实现多源数据集间的知识聚合。例如,就古琴艺术知识库中的"南京"而言,其能够链接到地理知识库中(江苏,上位,南京)的分类关系,历史知识库中(南京,都城,明朝)的非分类关系,故以"南京"为实体源进行外部语义链接能够丰富与其相关的地理与历史背景知识。本模式选择GeoNames、DBpedia等关联数据集[24],通过OWL内建属性Owl:sameAs将内部知识实体与外部知识库中的实体进行关联。最后,依凭W3C标准和原则配置相应的服务器在Web上发布RDF文档,以期实现我国非遗数据的语义链接与开放共享。

3.4　基于非遗人文知识图谱的语义知识服务

为了实现以用户人文性需求为驱动的非遗语义知识服务,本模式对知识图谱的应用旨在基于知识本体的逻辑框架与语义推理优势,从人文图谱中查询推理出全面、细粒度的人文实体或属性值,以快速洞悉领域知识的整体概貌,准确把握人文对象的细节线索,深入挖掘非遗知识的人文内涵。

(1)人文性知识的脉络梳理溯源,旨在从族性视角剖析本体的人文知识,对具有共同性质或特征的实体概念进行探索查询,如图7所示。

在图7中,非遗知识图谱通过本体中以同一内在特征、不同外在形式存储的实体资源展现古琴艺术的整体知识结构,厘清非遗人文知识的分类体系、类目设置与收录范围。例如,古琴流派有"金陵琴派""浙江琴派""广陵琴派"等,琴曲有"潇湘水云""广陵散""离骚"

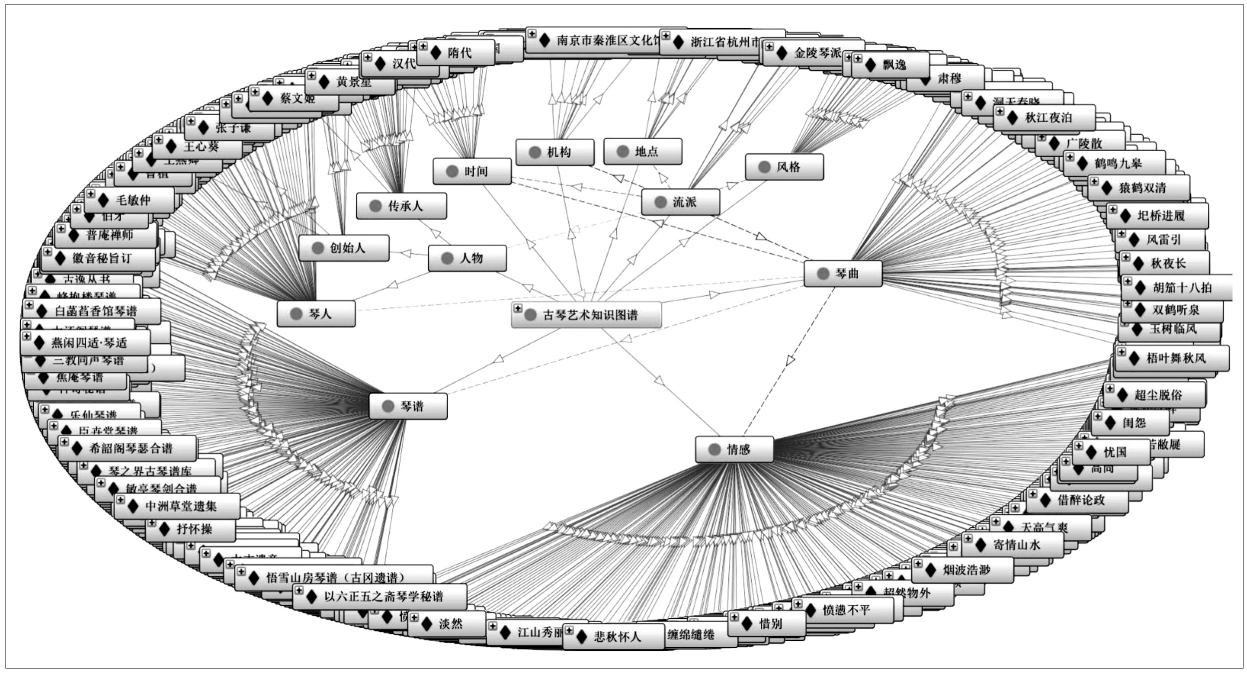

图7 非遗人文性知识脉络梳理溯源(以古琴艺术为例)

等,琴人有“郭沔”“庄臻凤”“毛敏仲”等,琴谱有“神奇秘谱”“太音补遗”“杏庄太音补遗”等。在此基础上,将时间、空间等概念作为主线嵌入古琴艺术的演化历程,即能实现对领域历史的发展溯源与知识样貌的语义分析。

(2)人文性知识的特性解析,旨在从特性视角检索非遗本体的人文知识,根据输入的用户需求进行SPARQL逻辑表示,从知识图谱中查询出实体或属性值,如图8所示。

图8中,以流派(Genre)类的“金陵琴派”为索引词,可检索其对象属性的值域术语,如hasPerson:“徐元白”“庄臻凤”,hasComposition:“潇湘水云”、列子御风”,hasLocation:“江苏省南京市”,hasStyle:“飘逸”“肃穆”,以洞悉“金陵琴派”的关联知识。

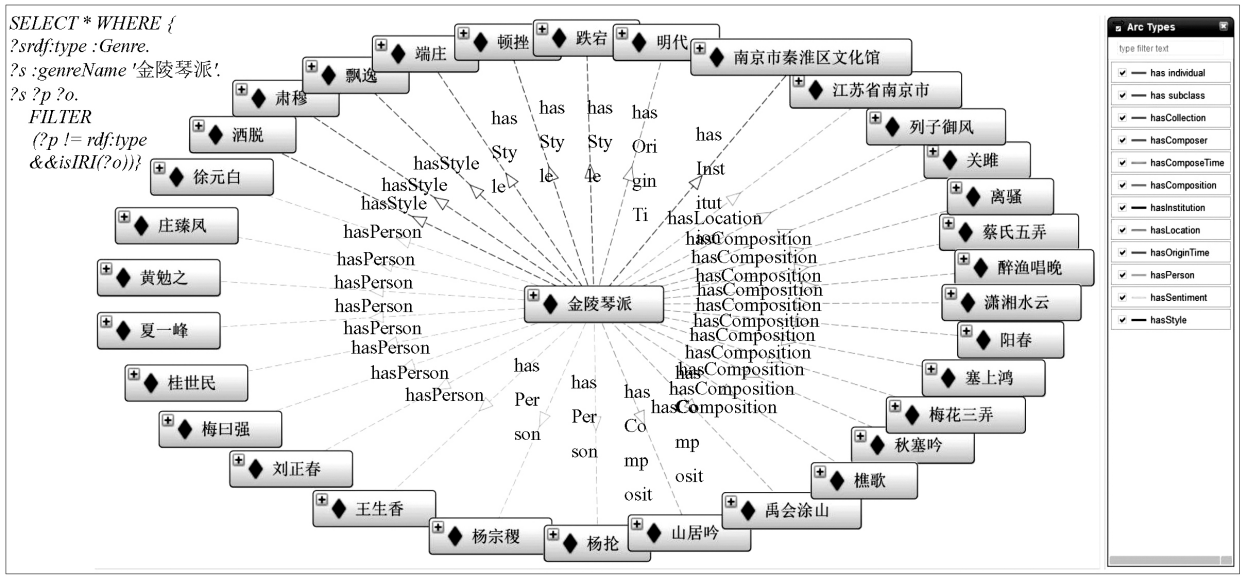

图8 非遗人文性知识特性解析(以“金陵琴派”为检索词)

（3）人文规律的推理估测与空白填补，旨在基于族性与特性双重视角对本体中所蕴含的人文规律进行演绎、推理以至预测，指引研究人员发掘人文领域空白以实现一定程度上的知识填补。在知识本体中可以通过自定义推理规则发现未被数据集存储的新知识，以关系间的推理最为重要，主要包括互逆关系（Inverse functional）、传递关系（Transitive）、对称关系（Symmetric）等。例如在古琴艺术本体中将guqin: hasComposition和guqin: hasGenre设置为一对互逆属性，根据类间对象关系，“金陵琴派”的琴曲有“潇湘水云”，可以推理出未知关系“潇湘水云”的艺术流派为“金陵琴派”。

进一步，传统文化所蕴含的人文情感（Humanity-Sentiment）是驱动人文意识形态形成的重要因素。现有古琴艺术的文学研究多涉及古琴流派与古琴曲间、古琴曲及其传达情感间的关系，鲜有学者探讨古琴流派与人文情感间的关联。因此，通过编制传递、互逆等关系推理规则能够展示流派与情感之间的推理关系（图9）。

图9通过互逆关系展现知识图谱中“情感→琴曲→流派”的推理过程。其中，流派X具备（hasHumanity-Sentiment）情感Z可推理出情感Z来自（hasHumanity-Genre）流派X，具体来说：情感“忧国忧民”“爱国”（Z）源于琴曲（Y），琴曲（Y）属于流派（X），从而发现知识：情感Z归属于（hasHumanityGenre）流派X，实现以人文情感推理古琴流派。通过统计，总结出具备“忧国忧民”“爱国”等情感特质的流派分布，发现更善于运用爱国艺术表现手法的古琴流派为“金陵琴派”与“浙江琴派”。

4　结语

本文面向大规模非遗非结构化文本，以细粒度的人文性知识为核心，综合知识图谱在不同场景下的技术特质，依循本体建构、知识解析、知识链接、知识服务等过程从语义层面构建领域知识组织模式，为非结构化文本到结构化知识以至开放共享知识库的一整套过程提供了理论支撑体系。在该体系的指导下，以联合国非遗名录中的“古琴艺术”为案例，从元数据特征解析与本体建模、文本语义关联解析、实体及实体关系融合、基于知识本体的人文知识服务等角度探讨了我国非遗人文知识图谱的技术实现路径。

本模式以非遗文本中细粒度知识的语义关联为主要驱动力开展领域知识组织。对此，知识本体建模为非遗领域资源提供了概念支撑，语义关联解析深化了非遗知识组织的层次与粒度，知识的语义描述与链接为关联数据的开放共享提供了可行性，语义知识服务能够挖掘知识图谱的人文价值。其中，非遗非结构化文本中人文性知识的语义关联解析是语义知识图谱实现链接共享和知识服务的根本基础。因此，方案中关

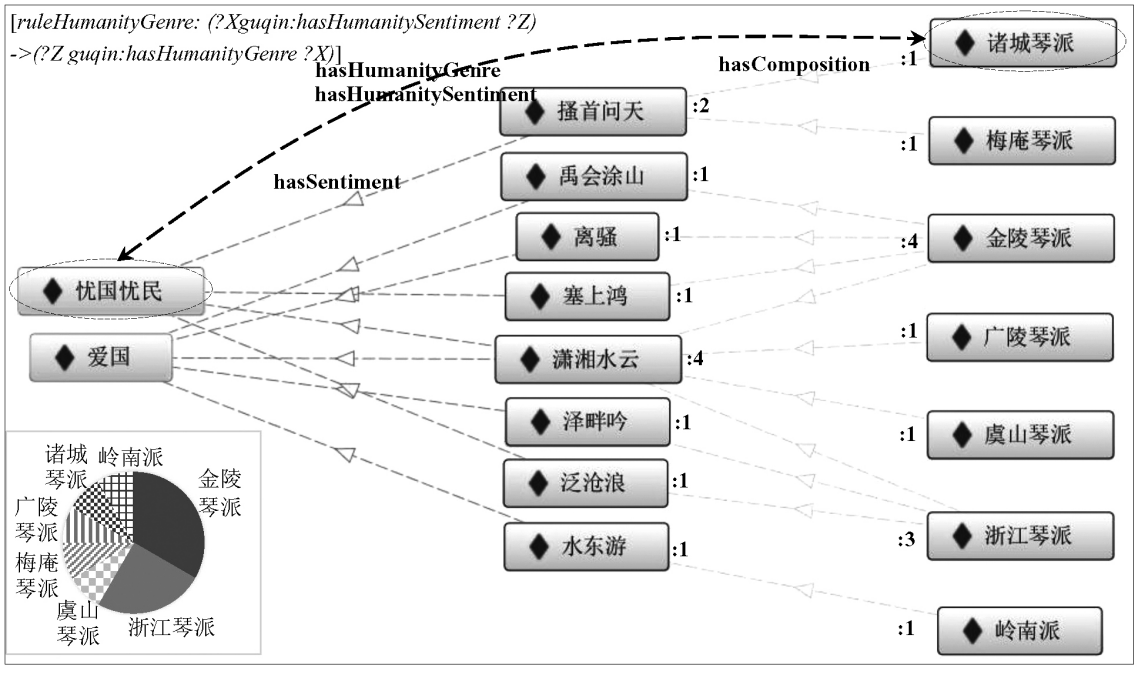

图9　非遗人文性知识发现（以人文情感推理古琴流派）

键自然语言处理技术的实现需要考虑对“冷启动”的噪声控制与文本内部深层特征的提取。所以,下一步研究将以此为切入点着重优化非遗人文性知识解析的完整性与准确性。

参考文献

[1] 侯西龙,谈国新,庄文杰,等.基于关联数据的非物质文化遗产知识管理研究[J].中国图书馆学报,2019,45(2):88-108.

[2] Ravat F, Song J, Teste O, et al. Efficient querying of multidimensional RDF data with aggregates: comparing NoSQL, RDF and relational data stores[J]. International Journal of Information Management, 2020, 54: 102089.

[3] 陈 涛,刘 炜,单蓉蓉,等.知识图谱在数字人文中的应用研究[J].中国图书馆学报,2019,45(6):34-49.

[4] Wang Z, Xu S, Zhu L. Semantic relation extraction aware of N-gram features from unstructured biomedical text[J]. Journal of Biomedical Informatics, 2018, 86: 59-70.

[5] 滕春娥, 王 萍. 非物质文化遗产资源知识组织本体构建研究[J]. 情报科学, 2018, 36(4):160-163.

[6] 王 蒙,许 鑫.主题图技术在非物质文化遗产信息资源组织中的应用研究——以京剧、昆曲为例[J].图书情报工作,2015,59(14):15-21.

[7] 翟姗姗.基于关联数据的非物质文化遗产资源聚合研究[M].北京:科学出版社,2015.

[8] Qiao C, Hu X. A neural knowledge graph evaluator: combining structural and semantic evidence of knowledge graphs for predicting supportive knowledge in scientific QA[J]. Information Processing & Management, 2020, 57(6): 102309.

[9] Bekoulis G, Deleu J, Demeester T, et al. Joint entity recognition and relation extraction as a multi-head selection problem[J]. Expert Systems with Applications, 2018, 114: 34-45.

[10] Liu W, Yu B, Zhang C, et al. Chinese named entity recognition based on rules and conditional random field[C]. Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence,2018:268-272.

[11] 邓 擘,郑彦宁,樊孝忠.信息抽取中实体关系模式的可信度评估[J].情报理论与实践,2009,32(12):103-105.

[12] 王 昊, 王密平, 苏新宁. 面向本体学习的中文专利术语抽取研究[J]. 情报学报, 2016, 35(6):573-585.

[13] Maengsik C, Harksoo K. Social relation extraction from texts using a support-vector-machine-based dependency trigram kernel [J]. Information Processing & Management, 2013, 49(1): 303-311.

[14] Zhang Y, Yang J. Chinese NER using lattice LSTM[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018:1554-1564.

[15] 任秋彤,王 昊,熊 欣,等.融合GCN远距离约束的非遗戏剧术语抽取模型构建及其应用研究[J/OL].数据分析与知识发现:1-19[2021-07-07].http://kns.cnki.net/kcms/detail/10.1478.G2.20210622.1617.002.html.

[16] Dou J, Qin J, Jin Z, et al. Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage[J]. Journal of Visual Languages & Computing, 2018, 48: 19-28.

[17] Mao J, Cui H. Identifying bacterial biotope entities using sequence labeling: performance and feature analysis[J]. Journal of the Association for Information Science and Technology, 2018, 69(9): 1134-1147.

[18] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[19] Liu W, Fu X, Zhang Y, et al. Lexicon enhanced chinese sequence labelling using BERT adapter[J]. arXiv preprint arXiv: 2105.07148, 2021.

[20] Ji D, Tao P, Fei H, et al. An end-to-end joint model for evidence information extraction from court record document[J]. Information Processing & Management, 2020, 57(6): 102305.

[21] 曹明宇,杨志豪,罗 凌,等.基于神经网络的药物实体与关系联合抽取[J].计算机研究与发展,2019,56(7):1432-1440.

[22] Geng Z, Zhang Y, Han Y. Joint entity and relation extraction model based on rich semantics[J]. Neurocomputing, 2021, 429: 132-140.

[23] Wan Q, Wei L, Chen X, et al. A region-based hypergraph network for joint entity-relation extraction[J]. Knowledge-Based Systems, 2021, 228: 107298.

[24] Saquicela V, Vilches-Blázquez L M, Corcho O. Adding semantic annotations into (geospatial) restful services[J]. International Journal on Semantic Web and Information Systems, 2012, 8(2): 51-71.

[作者简介]张 卫,男,1994年生,南京大学信息管理学院博士研究生。

王 昊,男,1981年生,南京大学信息管理学院教授,博士生导师(通讯作者)。

李跃艳,女,1991年生,南京大学信息管理学院博士研究生。

邓三鸿,男,1975年生,南京大学信息管理学院教授,博士生导师。

收稿日期:2021-07-07