



基于图数据库 Neo4j 的宋代镇江诗词 知识图谱构建研究

□李永卉* 周树斌 周宇婷 卢章平

摘要 为使诗词研究顺应大数据环境下的发展,有效分析诗词领域内知识间的关系,本研究将本体与图数据库技术相结合,提出诗词领域内知识图谱的构建方案。文章以宋代镇江诗词为例,在对诗词相关数据进行采集与分析的基础上,借助 Protégé 工具对诗词本体进行构建并生成 RDF 数据。使用语义插件完成 RDF 数据到图数据结构的映射后,通过 Neo4j 图数据库对诗词本体数据进行存储,完成宋代镇江诗词知识图谱的构建。结果表明,构建宋代镇江诗词的知识图谱,将诗词、作者、地点、时间、典故等数据以图数据的形式存储,实现了检索、分析与利用的功能,为今后诗词领域知识图谱的深度开发与数据存储奠定了基础。

关键词 知识图谱 图数据库 宋代镇江诗词 Neo4j

分类号 G122; G250.7

DOI 10.16603/j.issn1002-1027.2021.02.007

1 引言

镇江是长江下游的重镇,“雄峙江南,群山拱卫”,“襟吴带楚,控南拒北”^[1],历来为兵家必争之地,清代著名学者阮元曰:“京口自东晋以来,屹为重镇。”^[1]尤其“自大运河开凿后,镇江成为江河交汇处江南运河段的起点”^[2],成为沟通南北的交通枢纽,无数文人墨客驻足。两宋时期,镇江因为特殊的地理位置和历史文化积淀,产生了一批非常有特色的诗词。这些诗词在一定程度上展现了宋代镇江的风土人情和历史文化变迁,具有重要的文化和历史价值。然而在海量诗词数据背景下,传统研究方法往往需要花费大量的时间和精力去收集、处理和分析数据,较难满足大数据时代的研究和应用需求。

在数字人文背景下,人文学科与现代信息技术关系愈发紧密,文化与科技融合程度日益加深^[3],为诗词研究的现代化转型创造了契机。在大数据时代,深度学习、机器学习等人工智能技术在处理海量数据方面发挥着重要作用,在数据预测和决策制定上具有一定的优势,但是在数据描述上有所不足^[4],知识图谱则具有强大的数据描述能力,可弥补这方面的不足。知识图谱概念最早由谷歌公司于 2012

年提出,是一种基于图的数据结构,由节点和边构成,可用于描述客观世界存在的实体概念及其相互关系^[5]。在知识图谱中,各节点表示客观实体,节点间的边则用来表示实体与实体间存在的关系。可以说,知识图谱是关系的最有效表示方式,它把异种信息连结起来形成关系网络,从而可以从关系的角度去分析处理知识实体单元,这种表示方式非常适用于诗词领域的概念表达。

目前知识图谱已在众多领域中应用,在诗词研究中,也有了一定的探索实践。如崔京菁等运用知识图谱对小学语文古诗词翻转课堂教学模式与应用的探讨^[6];冯俐通过 Neo4j 图数据库对中学语文诗词的知识图谱进行了构建^[7];李畅、贾丙静等提出了构建唐代诗词知识图谱的简单方案^[8-9];周莉娜等基于知识图谱完成 KnowPoetry 平台的设计,支持和提供唐诗知识探索、时空轨迹、语义查询等知识服务^[10];刘昱彤等基于诗词的注释和词语的词典解释,从词语的维度构建古诗词图谱,进行相关推理与分析实践^[11]。此外,北京师范大学的“唐诗别苑”也是基于知识图谱实现的全唐诗语义检索与可视化平台^[12];北京大学—哈佛大学数字人文开放实验室则

* 通讯作者:李永卉,ORCID:0000-0002-7992-303X,邮箱:15862973682@163.com。



形成了宋元学案知识图谱及宋代学术传承知识图谱等项目成果^[13]。这些研究验证了知识图谱与诗词研究相结合的可行性,但无论从广度还是深度上都尚待提高和深化。虽然诗词结合知识图谱的研究成果颇丰,但是目前将诗词与知识图谱结合的相关研究主要是宏观层面的把控,缺少微观、精细化研究,更多关注和描述了诗词的外部特征,对其内容关注不足,如对诗词资源中的地域属性关注不够,对诗词的背景、典故、情感等内容挖掘不足,缺乏对相关资源的关联化研究。因此结合知识图谱,深度挖掘宋代诗词背景、内容、地域特征等,符合当前诗词研究的需要,具有较强的现实意义。

目前,学界将关联数据视为一种语义知识图谱,而将基于图数据库实现的知识图谱视为广义上的知识图谱,前者注重知识的发布与关联,后者则是知识挖掘与计算能力突出^[14]。二者虽有一定差异,但总体而言都包括知识的抽取、融合与加工等典型知识图谱构建过程^[15],一定程度上两类方案通过相互结合可以打破二者的局限与制约,从而实现优势互补。基于图数据库实现的知识图谱和关联数据的结合,无疑会成为知识图谱领域内研究的新增长点,这也为诗词领域的知识图谱构建提供了新的启示。之前已有通过本体与关联数据技术进行宋代镇江诗词资源关联组织方案的研究^[16],本文引入基于图数据库实现的知识图谱技术则是对以往研究的进一步深化,实际上实现了将关联数据与广义知识图谱的结合。一方面,本研究对数据的存储结构进行了优化,由关系型数据的存储模式转为非关系型的图数据存储结构,充分利用图数据结构的优点,在保证数据关联性的同时,提高了查询和计算的效率;另一方面,借助知识图谱进行数据间关系的揭示和知识推理,挖掘隐性知识,并在表现形式上进行改进,用知识图谱的结构对领域内知识元进行更为直观的展示,便于用户理解;同时,在内容上也更加深入,在“诗”“人”“地”的基础上进行维度的拓展,引入“时”和“事”,即增加了在时间层面的具体描述,并引入典故构成事件元素。因此。本文提出通过建立诗词知识图谱数据库,实现以图数据的形式对宋代镇江诗词资源进行存储,探索将图数据库技术应用于诗词研究的可行性,有利于在大数据背景下进一步探讨诗词资源的深度开发与利用。

本论文首先广泛收集宋代镇江诗词数据,包括确

定数据来源、采集和分析数据。然后通过 Protégé 构建本体,以本体为基础,运用 Protégé 软件进行诗词领域的本体构建,分析其中的知识结构,明确领域内的本体概念,针对领域内逻辑概念进行限定,确定关系属性,利用定义规则的方式导入实例,形成 RDF 数据,完成对知识与数据的组织,从而为知识图谱的构建奠定基础。最后通过 Neo4j 图数据库实现对于 RDF 数据的存储,构建诗词知识图谱,利用 Cypher 语言对诗词中的实体、关系以及属性进行设计、编译,使诗词知识图谱可以呈现清晰的节点关系结构,并且利用该图数据库进行交互式查询和关联化推理。这样就为诗词的知识组织提供了方法基础,为今后诗词资源的研究和利用提供了模型上的借鉴参考。

2 数据获取与分析

2.1 数据来源与采集

本研究所需数据主要包括镇江地名数据、通过地名数据确定的诗词数据、以及诗词相关的作者数据。数据采集主要以谭其骧先生编的《中国历史地图集》第六册《宋辽金时期》标注的地名为据,并利用《中国行政区划通史》《嘉定镇江志》《至顺镇江志》和其他相关史籍、地方文献,系统搜集镇江在宋代及以前的相关地名数据,并对数据进行分类和地点间关系的归纳,包括异名、旧称、简称以及地名间的从属关系。将镇江主要地名确定后,再进行诗词相关数据的采集,宋诗的题目和内容数据来源于全宋诗分析系统,宋词来源于诗词在线网、古诗文网等网站,网络收集后,通过权威文献,如北京大学出版社 1998 年版《全宋诗》和中华书局 1999 年版《全宋词》等书籍进行校对。人物数据主要采集自国家图书馆规范档、中国基本古籍库、全宋诗分析系统、国家图书馆的历史人物数据库等,同时辅以《宋史》《宋会要辑稿》《续资治通鉴长编》《建炎以来系年要录》等宋代基本史籍,以及《全宋文》、作者个人的诗文集、地方志、碑刻等重要文献。对初步采集的数据进行逐条筛选,剔除无关数据,使得搜集的数据准确可靠,具有一定的学术价值。

数据采集完成后,将采集到的数据根据属性特点和彼此间的关系信息,结构化保存并赋予相应标签,录为 CSV 格式文件存储。其中,诗词数据包括题目、正文、作者、相关地点、创作时间、创作背景、用典、诗词 ID 等内容,词数据还包括词牌内容;人物数



据包括姓名、朝代、字、号、生年、卒年、籍贯、人物小传、诗词 ID 等内容;地点数据则包括地名及地名 ID;此外对相关数据字段还有文献来源版本的标注。共计收集到与宋代镇江相关诗 1005 首、词 145 首、作者 341 位、地名 89 处。

2.2 数据分析

宋代镇江诗词的数据收集完成后,基于南宋与北宋的时代背景,对采集到的诗词和人物数据分析,由于部分作者生卒年份不详,本研究选取了具有明确时代背景的人物及其诗词作品来分析,共选出 380 首北宋诗、30 首北宋词、472 首南宋诗、87 首南宋词。分析来源数据,对诗词作者进行统计发现,北宋作者主要包括苏轼、梅尧臣、王安石、蔡肇、苏颂、吴则礼等,南宋作者主要包括刘宰、杨万里、陆游、文天祥、章甫、戴复古等,作者与镇江的关系主要包含途经镇江、籍贯为镇江、在镇江任职、徙居镇江等。

诗词内容方面,使用 ShapeWordle 工具^[17]对两宋镇江诗词内容进行分词与可视化展示,具体如图 1 所示(其中图 1(a)为北宋诗的关键词,图 1(b)为南宋诗的关键词,图 1(c)为北宋词的关键词,图 1(d)为南宋词的关键词)。去除镇江典型地名词语(如京口、金山、海门等)后发现:北宋诗的核心词语为“江山”“楼台”“不见”“不知”“万里”等,北宋词的核心词语为“夕阳”“灯火”“行乐”“东风”“扬州”等;南宋诗的核心词语为“长江”“万里”“中流”“不知”“江山”等,南宋词的核心词语为“英雄”“人间”“江

山”“万里”“长江”等,从中可以看出两宋诗词风格具有一定的差异性。

诗词的情感方面,使用 ROSTCM 工具^[18]对两宋的诗词分别作情感分析发现,北宋诗积极情绪占比 21.40%,中性情绪占比 46.29%,消极情绪占比 32.31%;南宋诗积极情绪占比 21.93%,中性情绪占比 46.71%,消极情绪占比 31.36%;北宋词积极情绪占比 13.36%,中性情绪占比 60.37%,消极情绪占比 26.27%;南宋词积极情绪占比 18.67%,中性情绪占比 52.40%,消极情绪占比 28.93%,总体而言,两宋镇江诗词的情感表达倾向较为接近,均为消极倾向大于积极倾向。通过进一步考察诗词内容和两宋镇江的政治地理环境^[19],发现其情感产生原因不尽相同,这是由南宋和北宋时镇江所处的地理位置和政治局势决定的。北宋时期,镇江为南北交通枢纽,官员谪宦旅途的必经之地,北宋文人吟咏镇江多是注重秀美的山河风光,旅途中抒发的也多是送别或羁旅行役中的感伤之情,更多的是注重个体情感的阐述。到了南宋,镇江属于前沿阵地,是抗击金和伪齐的战区,诗词中对镇江地理形势和战略地位给予了更多关注,情感态势也是更加偏向家国情怀,如此就形成了浓厚的慷慨悲凉、忧国忧民的诗词风格^[20]。综上,对所采集诗词的数据分析,虽然不能做到百分百全面,但已具有较强的代表性,一定程度上能够反映两宋三百余年镇江的社会风貌和学士文人之思想感情,对其进行研究具有重要的学术和实用意义。

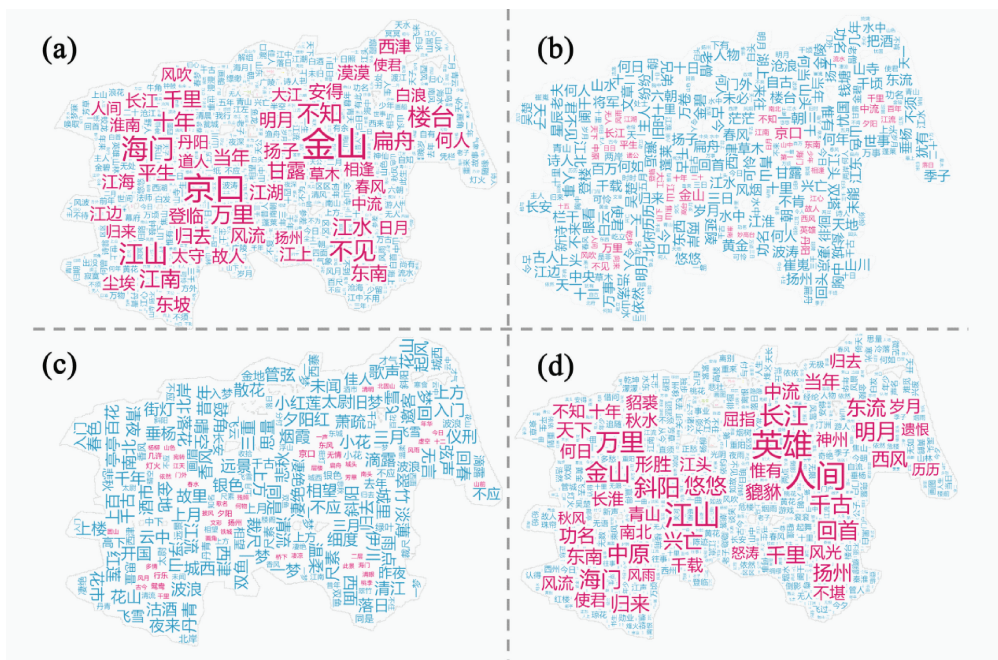


图 1 两宋镇江诗词内容关键词可视化



3 宋代镇江诗词领域本体建模

本研究首先应用 Protégé 本体软件工具,进行宋代镇江诗词本体建模、实例的导入,以及模型的可视化呈现,形成 RDF 三元组数据,完成对诗词知识与数据的组织,成为领域内知识图谱构建的基础。

3.1 核心类的定义

基于上文对于宋代镇江诗词数据的提取和分析,结合数据的结构特点,拟使用 Protégé 软件工具对宋代镇江诗词本体进行构建与描述。为保证语义推理的正常进行,需对各本体概念进行定义说明。调研发现,目前可供利用的本体模型无法满足诗词资源组织描述的需要,故通过复用现有本体词表与自建本体词表相结合的方式对本体的构建。基于本体的通用性原则,通过 direct imports 的方式对 dc,dcterms,foaf,geonames^[21]、上海图书馆开放数据平台本体(包括人名规范库、中国历史纪年表、地理名词表等)^[22]等国内外通用本体进行复用,同时自定义词表实现对诗词核心类和属性的描述。不同本体间相同的实体属性通过 equivalent to 进行等同处理,从而实现多本体间类和属性的共享与关联。关于自建本体词表的设计,首先自定义 IRI 为 http://www.region_poem_ontology.com,通过 Protégé 中 class 标签对本体中的核心类进行定义与规范说明,为充分描述诗词的时空与事件属性,将时间与事件单独设为两类,此外增设版本来源类用以交代相关数据字段出处,故利用 class 标签创建诗词(poem)、人物(person)、地点(place)、时间(time)、历史事件或典故(allusion)、版本来源(editonSource),类下包括子类,如诗词包括诗、词两部分,地点包括人工建筑(building)、自然景观(naturalLandscape)与地域(region)三部分,且可以进一步细化,如时间包括朝代(temporal)、年号(reignTitle)、公元纪年(commonEra)、历史年号纪年(reignTitleEra)等。最终的核心概念层级结构以结构树的形式展现,如图 2 所示。

3.2 属性创建与实例导入

宋代镇江诗词领域本体类创建完成后,需要对各类属性做出定义和描述。属性由对象属性和数据属性两部分组成,其中对象属性用于定义类之间的关系,定义域和值域均为类;数据属性则用以描述本体的数据类型,定义域为一种或多种类,值域为某一指定的数据类型^[23]。在 Protégé 软件工具中,通

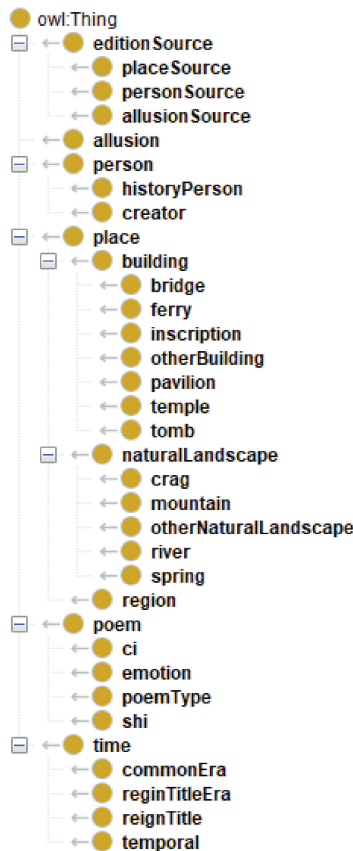


图 2 本体概念层级结构

过 Object Properties 标签对镇江诗词对象属性进行构建,以各类间的关系为依据进行相关对象属性的定义,对象间的关系可以是异类间的,也可以是同类间的,如 hasWriter,其定义域(Domain)为 poem,值域(Range)为 person,表示诗词的作者是某人;placeOtherName,其定义域与值域均为 place,表示了同一地点的异名或别称;placeShortName 与 placeFullname 定义域与值域为 place,表示了同一地点的简称与全称,二者具有相反(inverse)的关系;visited 与 wasVisitedBy 则互为相反关系,代表了人物(person)与地点(place)间到访与被到访的关系,对象属性的具体设置如图 3 所示;数据属性的创建则通过 Data Properties 标签来进行,根据采集诗词数据的实际情况,定义了 25 个数据属性,设置数据属性如图 4 所示。属性设置完成后,利用 Protégé 的 OntoGraf 插件进行本体模型可视化展示,如图 5 所示。

建立本体模型架构并对类和属性进行定义后,制定规则,导入诗词、人物、地名、时间、事件等实例(individuals)数据,形成领域内知识组织架构体系。以地点(place)数据中建筑(building)类别中的亭台



楼阁(pavilion)的实例数据为例,进行可视化展示,如图6所示。其中包含了 pavilion 类中各实例数据,各地点数据的全称、简称、异名均单独设定了对象实体,并通过对象属性的设置来建立各实体间的联系;图右以“北固亭”为例展示了其相关属性信息,如 URI,对象属性包括所属地点、相关诗词、异名等,数据属性如地名、地点 ID。数据的实例化完成后便可生成 RDF 数据,成为宋代镇江诗词知识图谱的理论及数据基础,为领域内本体开放、共享及互操作等应用提供底层逻辑支持。

Object Property	Domain	Range
owl:topObjectProperty		
belongsTo	owl:Thing	owl:Thing
belongsToPlace	place	place
belongsToTemporal	editionSource, time, allusion, person	temporal
relatedPoem	place	poem
relatedShi	place	shi
relatedCi	place	ci
happenedTime1	poem, allusion	reginTitleEra
include	owl:Thing	owl:Thing
includePlace	place	place
relatedPlace	poem	place
titleRelatedPlace	poem	place
contentRelatedPlace	poem	place
wasVisitedBy	place	person
created	creator	poem
emotions	poem	emotion
placeOtherName	place	place
relatedPerson	owl:Thing	person
placeShortName	place	place
poemType	poem	poemType
relatedAllusion	poem, person	allusion
hasWriter	poem	creator
happenedTime	poem, allusion	commonEra
placeFullName	place	place
visited	person	place
sourceFrom	person, allusion, poem, place	placeSource, allusionSource, personSource

图3 对象属性设置

4 宋代镇江诗词知识图谱

本文使用的 Neo4j 是一种基于 Java 的开源图数据库,其采用具有自由邻接的图存储结构,比传统的关系型数据库具有更好的数据分析处理能力,在查询方式与效率方面明显优于基于 OWL 文档的查询与存储方式^[24],同时 Cypher 查询语言为其提供了简便有效的表达方式^[25]。故本文使用 Neo4j 图数

Data Property	Domain	Range
owl:topDataProperty		
edition_source	editionSource	rdf:PlainLiteral
edition_name	editionSource	rdf:PlainLiteral
allusion_name	allusion	rdf:PlainLiteral
creator	creator	rdf:PlainLiteral
creator	creator	rdf:PlainLiteral
poem_id	poem	rdf:PlainLiteral
creator	creator	rdf:PlainLiteral
creator	creator	rdf:PlainLiteral
creator	creator	rdf:PlainLiteral
ci	ci	rdf:PlainLiteral
creator	creator	rdf:PlainLiteral
creator	creator	rdf:PlainLiteral
poem	poem	rdf:PlainLiteral
creator	creator	rdf:PlainLiteral
poem_title	poem	rdf:PlainLiteral
poem_coutesyName	poem	rdf:PlainLiteral
poem_background	poem	rdf:PlainLiteral
person_id	creator	rdf:PlainLiteral
place_id	place	rdf:PlainLiteral
person_nativePlace	creator	rdf:PlainLiteral
place_name	place	rdf:PlainLiteral
person_biography3_type	creator	rdf:PlainLiteral
poem_content	poem	rdf:PlainLiteral
title	owl:Thing	rdf:PlainLiteral
person_name	creator, person	rdf:PlainLiteral
person_pseudonym	creator	rdf:PlainLiteral
allusion_content	allusion	rdf:PlainLiteral

图4 数据属性设置

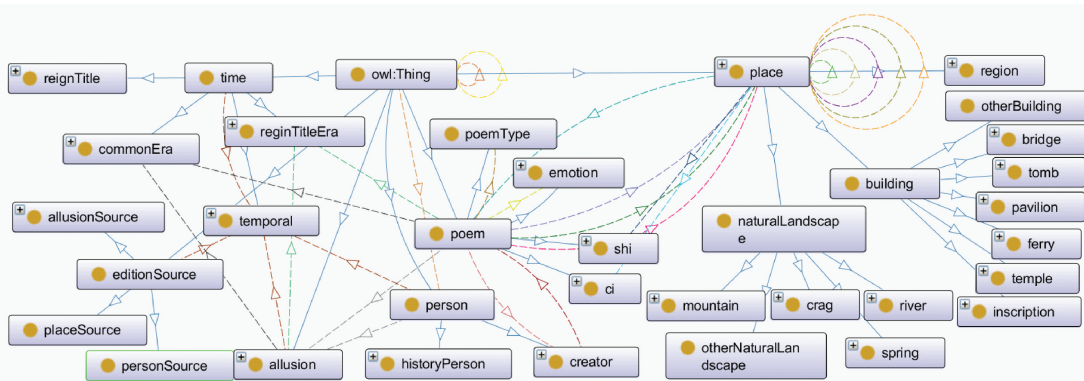


图5 本体模型可视化

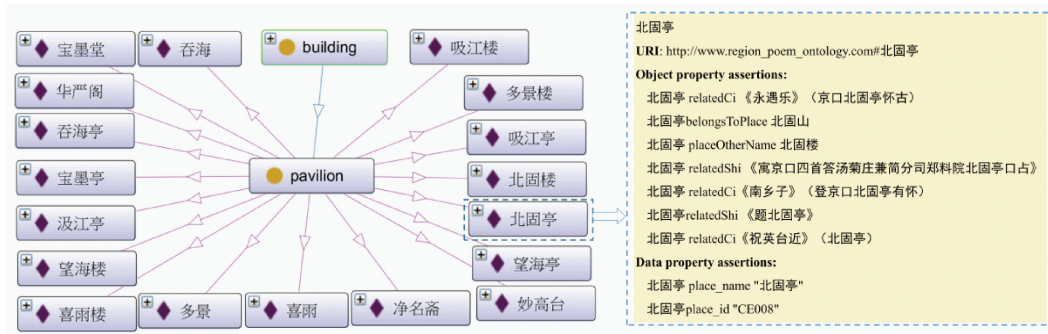


图6 地名实例数据展示



数据库来完成 RDF 数据的存储,构建宋代镇江诗词的知识图谱,基于对已有节点间关系的设定,通过计算(如图遍历、路径计算)和推理的方式可以得出所需的知识单元内容。

4.1 映射匹配分析与本体知识存储

上文通过 Protégé 构建宋代镇江诗词本体模型并生成 RDF 数据,从而建立起领域内知识框架及本体类与属性和实例间的映射关系。RDF 数据由多个三元组构成,三元组 = <主语,谓语,宾语>^[26],本身便具有语义网络结构,为知识图谱构建提供了具有统一标准规范的表现形式^[27]。RDF 数据能够与图数据结构相互匹配映射,可通过图数据库实现对宋代镇江诗词本体知识的存储。Neo4j 图数据库主要包含节点(node)、关系(relationship)、属性(property)、标签(label)和路径(path)等五类元素,节点和关系皆可以创建多种属性元素,属性通过键值对表示且都是独立存储的^[28]。本文的诗词 RDF 数据与 Neo4j 模型中的元素主要从节点、关系及属性三个方面进行对应,Neo4j 中的节点与本体中的类名和实例名一一对应,其他类型节点则根据已定义的语义框架创建多个键值对;关系与本体中的对象属性相对应,如类与实例间的“instance-of”关系,诗词与诗人间的“hasWriter”“created”关系;属性则是与本体中的数据属性进行对应,例如,“person”对象的数据属性主要包括“person_name”“person_ID”等,“poem”属性包括“poem_title”“poem_ID”等。这样就做到了本体数据到 Neo4j 存储的完整映射。

本文所使用的 Neo4j 版本为 Neo4j 社区版(3.5.13),RDF 作为用于数据交换的 W3C 标准模型,Neo4j 本身对 RDF 数据格式是不支持的,通过使用插件 NSMNTX^[29]实现在 Neo4j 中使用 RDF,无损地将 RDF 数据存储于 Neo4j 中,可随时导出导入 RDF 而不会出现数据丢失的问题,并且可在 Neo4j 上进行模型映射和推理。完成插件的配置后,通过 semantics.import RDF 方法导入 RDF 数据,该方法以其最基本的形式仅使用 URL 字符串来访问 RDF 数据和序列化格式,导入 URL 所返回的三元组并将其持久化到 Neo4j 中。

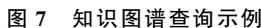
Neo4j 导入 RDF 数据后,对各属性关系进行调节,对可视化效果如节点与关系的颜色、显示的主属性等进行调整,就可以对宋代镇江诗词知识图谱进行构建。Neo4j 一方面可以通过对各实体节点进行

图遍历及节点间路径关系的计算,若已知节点间的关系,可直接依托于节点间的关系对所需内容进行查询,在未知节点间关系的情况下,可以通过查询路径的方式对节点间的关系进行挖掘,路径设置的深度影响所得结果的深度;另一方面,可以根据已有的逻辑关系实现隐性数据关系的发现与挖掘,完成对节点间关系的推理。通过知识图谱展现的方式极具整体性,一目了然,且其结构属性适用于对检索结果的进一步挖掘和深入,在得到查询结果的基础上,可以推理出检索结果以外的隐性知识内容,以关联的形态对诗词领域内的知识元进行充分的调动和利用。

4.2 结果展示和查询推理

完成 Neo4j 对本体知识的存储后,可使用 Cypher 进行查询和推理。通过 Cypher 语言可以对数据库中的数据进行增(CREATE)、删(DELETE)、改(SET)、查(MATCH)操作,方便地实现对诗词、作者、地点等实体及相关关系的检索、遍历和导航功能。生成的知识图谱通过颜色差异将不同实体类型的节点区别开来,节点间的边代表了不同实体间的关系,如连接作者实体与诗词实体之间的有向边,表示诗词与作者间存在的创作及被创作关系。对构建的知识图谱进行查询,如,查询地点“西津渡”相关的诗词内容,通过 Cypher 语句查询后,可得到“西津渡”诗内容,点击节点后可得到距离其路径为 1 的全部节点,拓展了 Neo4j 查询的易操作性。本文对查询相关内容进一步延伸,得到“西津渡”相关的全部知识图谱内容,如图 7 所示,不仅得到了与“西津渡”直接相关的诗词内容,还发现“西津渡”具有“西津”的简称,具有“蒜山渡”的异名,且可以进一步发现与“西津”和“蒜山渡”相关的诗词,且图中的各节点关系的属性可以进一步展示。

通过推导实体间的路径关系,可探讨实体间的联系,如查询“辛弃疾”关于“北固山”的诗词,结果如图 8 所示,可以发现“辛弃疾”并未创作内容直接涵盖“北固山”的诗词,但是两首词与“北固山”密切相关,其中“《永遇乐》(京口北固亭怀古)”一词题目与“北固亭”相关,而“北固亭”地属“北固山”;而“《南乡子》(登京口北固亭有怀)”一词题目与“北固亭”相关,内容与“北固楼”相关,而“北固亭”与“北固楼”互为异名,为同一地点(《嘉定镇江志》云:“北固楼或名为亭”)^[30]。由此,“辛弃疾”“北固山”及其相关诗词间的关联便可一目了然。



对诗词内容进一步探析,还是以“辛弃疾”所作的“北固山”相关的两首词为例,在诗词、人物、地点的基础上生成典故、相关历史人物、时间、诗词类型、版本来源等节点,具体如图 10 所示。可以发现,“辛

弃疾”的数据来源包括《宋史·辛弃疾传》《宋会要辑稿·职官》《全宋诗》中的作者小传;辛氏的两首词创作时间相近,其中《永遇乐》的创作时间为开禧元年(1205),而《南乡子》的创作具体年份为嘉泰四年(1204)或开禧元年(1205)^[31]。京口地区在南宋属两浙西路镇江府^[32],在诗词的创作背景属性中可以发现,这一时间段,辛弃疾以六十五岁高龄出任镇江知府(嘉泰四年三月到任,开禧元年六月改知隆兴府)^[33-34]。镇江的北固山濒临大江,以险固得名,三国时期,与曹操争衡的孙权曾在此处屯兵。南北朝时期,曾经北伐中原取得胜利的宋武帝刘裕也在此起家;后刘裕之子在没有充分准备的情况下举兵伐魏,结果兵败,北魏太武帝反攻,曾一度挺进镇江对岸的瓜州(扬州)。到南宋绍兴四十一年(1161),金主完颜亮攻陷扬州直逼采石,镇江一度面临严重危



机,直到虞允文在采石矶胜利,才解除危机。“这些旧事,远的已有千年,近的也快半个世纪了,但在作者的脑海中都被这一座饱经沧桑的名城联系了起来”^[35]。所以作品中用典较多,通过“孙权”“刘裕”“刘义隆”“王玄谟”“拓跋焘”“廉颇”等诸多历史人物事件,表达了辛弃疾主张北伐又反对草率行事以免重蹈刘义隆“仓皇北顾”覆辙的思想,且自比廉颇抒发了老骥伏枥的悲愤,后者则通过赞扬孙权的英雄气概讽刺了南宋统治者的无能,二者皆展现了诗人的爱国主义情怀^[36],作品“风格明快,情调昂扬,充满了鼓舞人心的力量”。因此,通过辛氏的两首词也直接地加深了“北固亭”或“北固楼”这一地名意象的文化内涵,对于“北固山”的文化特色建设也颇有益处。通过对诗词的时空背景、典故或相关历史事件、相关历史人物等内容的多元化呈现,一种历史的画面感油然而生,诗词所蕴含的情感以及作者的创作意图跃然于眼前,进一步突出了诗词中包含的地方文化底蕴,知识图谱的优势由此可见一斑。

综上,本研究主要通过 Protégé 结合 Neo4j 来实现本体与知识图谱的结合,将其应用于诗词研究,突破了现有研究中对诗词内容属性揭示不足的局限,对诗词进行了“诗”“人”“地”“时”“事”的全方位探讨和多维度分析,从诗词知识内容的背景、典故、情感、地域特征等方面进行深入的挖掘与组织。同

时,充分利用 Neo4j 架构灵活、性能可靠、可扩展性强且可用性高等优势,使用 Cypher 语言高效准确地完成对数据的检索与推理,克服了现有研究中使用关系型数据库存在的低效与不足,且速度比基于本体语言的查询推理更快。因此,本研究验证了将图数据库应用于诗词知识图谱构建的优势,为未来诗词领域的发展研究带来更多可能。

5 结语

本文通过对宋代镇江诗词资源进行调研与采集,对采集到的数据中的诗词、诗人、地点、时间、典故等实体属性逐一进行提取,在本体构建的基础上,通过图数据库完成对诗词知识的结构化存储,实现对宋代镇江诗词知识图谱的构建,并可进行多种查询检索和推理,细粒度地实现了对各部分实体属性的关联,方便对于诗词资源的调用。本文研究的不足之处在于,宋代镇江诗词知识图谱构建过程中,辅助了大量人工手段对各部分数据进行筛选与甄别以保证数据的准确,对于命名实体识别的任务仍无法自动完成,研究的维度有待进一步加强。后续研究需建立古诗词文本标注的语料库,在诗词知识单元上进一步细化拓展,同时可将知识图谱与更多的数字人文技术框架结合使用,共同来完成对古诗词的知识组织与开发利用。

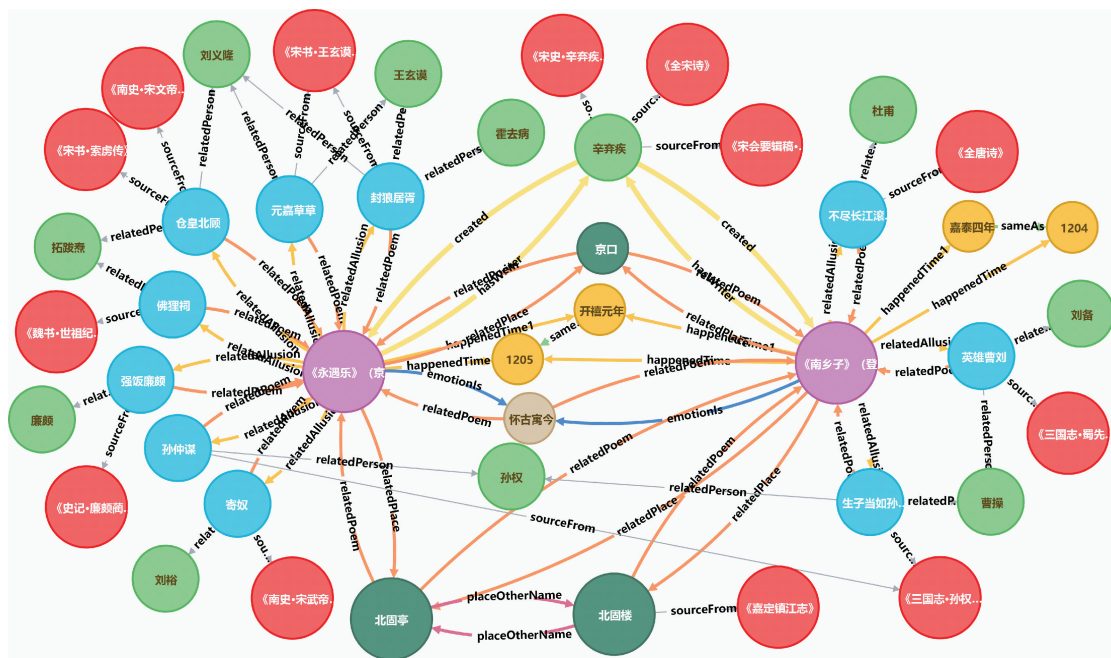


图 10 诗词内容示例



本文立足于宋代镇江诗词资源完成对诗词领域知识图谱的构建,但其应用价值不拘泥于对宋代镇江这一时一地诗词的知识组织。古诗词内部属性和外部特征结构具有一定的共性,故在现有数据的基础上从内涵和外延均可进一步扩展,所采用的方法与技术路线对于整个诗词领域具有普适的借鉴意义。构建诗词领域的知识图谱有利于完善诗词资源的整合利用,促进诗词知识的关联共享,在应用场景上既可为技术人员提供技术支持,也可为专家学者提供可靠又全面的诗词知识搜索服务,还能为社会大众提供通俗易懂的诗词文化普及教育,从而推动地方历史文化、旅游事业的发展。本文工作的意义在于,拓展了宋代镇江诗词的研究方法和范畴,更好地对诗词本体数据进行了展示和存储,可为诗词文化领域与科技的深度融合演进提供参考。

参考文献

- 俞希鲁.至顺镇江志[M].南京:江苏古籍出版社,1999.
- 严其林,程建.京口文化[M].南京:南京大学出版社,2001.
- 柳执一.“文化+科技”的破壁创新及融合发展[J].人民论坛,2019(35):138-139.
- 赵凯,王华星,施娜,等.基于 Neo4j 桂枝汤类方知识图谱的研究与实现[J].世界中医药,2019,14(10):2636-2639,2646.
- 刘峤,李杨,段宏,等.知识图谱构建技术综述[J].计算机研究与发展,2016,53(3):582-600.
- 崔京菁,马宁,余胜泉.基于知识图谱的翻转课堂教学模式及其应用——以小学语文古诗词教学为例[J].现代教育技术,2018,28(7):44-50.
- 冯俐.基于 Neo4j 图数据库构建中学语文诗词知识图谱[D].西安:陕西师范大学,2019.
- 李畅.基于知识图谱的唐代诗词平台构建[J].科技视界,2019(34):131-132.
- 贾丙静,马润.基于实体对齐的知识图谱构建研究[J].佳木斯大学学报(自然科学版),2018,36(3):453-455,464.
- 周莉娜,洪亮,高子阳.唐诗知识图谱的构建及其智能知识服务设计[J].图书情报工作,2019,63(2):24-33.
- 刘昱彤,吴斌,白婷.古诗词图谱的构建及分析研究[J].计算机研究与发展,2020,57(6):1252-1268.
- 李艳燕,张香玲,李新,等.面向智慧教育的学科知识图谱构建与创新应用[J].电化教育研究,2019,40(8):60-69.
- 北京大学数字人文实验室.项目展示[EB/OL].[2020-07-16].<https://pkudh.org/program>.
- 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.
- 刘峤,李杨,段宏,等.知识图谱构建技术综述[J].计算机研究与发展,2016,53(3):582-600.
- 李永卉,周树斌,黄柏芝.地方诗词资源的关联组织研究——以苏轼镇江诗词为例[J].数字图书馆论坛,2020(2):28-36.
- Wang Y H, Chu X W, Zhang K Y, et al. ShapeWordle: tailoring wordles using shape-aware Archimedean spirals[J]. IEEE Transactions on Visualization & Computer Graphics, 2020, 26(1): 991-1000.
- 吴潇,李鑫,赵炜.基于唐诗文本挖掘的关中地区人文景观格局研究[J].风景园林,2019,26(12):52-57.
- 余蔚.两宋政治地理格局比较研究[J].中国社会科学,2006(6):171-183,207.
- 胡玉尺.两宋京口诗歌研究[D].湘潭:湖南科技大学,2016.
- 姚天泓,陈艳梅,刘革,等.基于 CIDOC-CRM 的数字人文史料资源语义文化知识组织研究——以张学良史料资源为例[J].图书馆学刊,2019,41(7):35-43.
- 上海图书馆开放数据平台[EB/OL].[2020-08-16].<https://data.library.sh.cn/index>.
- 潘理虎,张佳宇,张英俊,等.煤矿领域知识图谱构建[J].计算机应用与软件,2019,36(8):47-54,59.
- 黄奇,钱韵洁,袁勤俭,等.基于图形数据库的产品分类本体存储研究[J].情报学报,2019,38(8):849-859.
- 林启胜,王磊,周喜,等.基于图数据库的文献检索方法优化与实现[J].微电子学与计算机,2017,34(10):63-67.
- 王飞,易绵竹,谭新,等.基于 Neo4j 对涉藏领域本体的存储方法研究[J].郑州大学学报(理学版),2019,51(2):60-65.
- 沈志宏,姚畅,侯艳飞,等.关联大数据管理技术:挑战、对策与实践[J].数据分析与知识发现,2018,2(1):9-20.
- 施国良,谢泽宇,杨小莉.高校图书馆复杂网络构建与智慧化应用探索[J].图书情报工作,2019,63(23):106-112.
- Neo4j. NSMNTX-Neo4j RDF & semantics toolkit[EB/OL].[2020-05-21].<https://neo4j.com/labs/nsmtx-rdf>.
- 卢宪.嘉定镇江志:卷十二[M].南京:江苏古籍出版社,1988:2.
- 朱德才,薛祥生,邓红梅.辛弃疾词新释辑评.下[M].北京:中国书店,2006:1449,1458.
- 谭其骧.中国历史地图集:第6册.宋·辽·金时期[M].北京:中国地图出版社,1982.
- 脱脱等.宋史[M].北京:中华书局,1977:12165.
- 邓广铭.辛弃疾传·辛稼轩年谱[M].北京:三联书店,2007:253-258.
- 陈千帆,吴新雷.两宋文学史[G]//陈千帆全集:第十三卷.石家庄:河北教育出版社,2001:358.
- 朱德才.辛弃疾词选[M].北京:人民文学出版社,1988:220-223.

作者单位:李永卉,江苏大学科技信息研究所,江苏大学廉政法治研究与评估中心,江苏镇江,212013
周树斌、周宇婷、卢章平,江苏大学科技信息研究所,江苏镇江,212013

收稿日期:2020年6月9日

修回日期:2020年11月19日

(责任编辑:支娟)



Research and Implementation on Knowledge Graph of Zhenjiang Poetry in Song Dynasty Based on Graph Database Neo4j

Li Yonghui Zhou Shubin Zhou Yuting Lu Zhangping

Abstract: In order to adapt the research in the field of poetry to the development of big data, to effectively analyze the relationship between knowledge in the field of poetry, and to combine the ontology technology with graph database technology, a construction plan of knowledge graph in the field of poetry was proposed. The study takes Zhenjiang poetry in Song Dynasty as an example. Based on the collection and analysis of relevant data of Zhenjiang poetry in Song Dynasty, the Protégé tool was used to construct the ontology in the field and to generate RDF data. After using the semantic plug-in to complete the mapping of RDF data to the graph data structure, the poetry ontology data is stored through the Neo4j graph database to complete the construction of the Song Dynasty Zhenjiang poetry knowledge graph. It shows that The Neo4j graph database was used to construct the knowledge graph of Zhenjiang poetry in the Song Dynasty, which realized the functions of storing poetry, poet, place, time and allusions in the form of graph data for analysis and retrieval. The research results prove the feasibility of the method and lay a foundation for the further development and data storage of knowledge graph in the field of poetry.

Keywords: Knowledge Graph; Graph Database; Zhenjiang Poetry in Song Dynasty; Neo4j



(接第 51 页)

An Exploratory Analysis on the Interdisciplinary Knowledge Diffusion in the International Studies for Digital Humanities

Wang Jingjing Ye Ying

Abstract: Digital humanities, as an emerging interdisciplinary research field, has received great attention from many disciplines. The interdisciplinary knowledge diffusion of digital humanities is explored in this article. Using dataset from Web of Science, we conduct interdisciplinary analysis through the betweenness centralization, network density, network coreness, and DIV indicator. For knowledge diffusion, we use papers cited at least once and their citing papers for analysis, and then we use the Sankey diagram for visualization. The results show that interdisciplinarity has become stronger and stronger in the field of international digital humanities, where the disciplines such as arts & humanities, engineering have occupied an important position. However, the network coreness of information science & library science has shown a downward trend. The decline of the network coreness in the field of digital humanities is conducive to establishing more connections between digital humanities and new disciplines, to absorb other knowledge and expand the scope better. In the process of knowledge diffusion, knowledge flows in the two disciplines of computer science, information & library science are relatively frequent, while the papers in the field of linguistics have received extensive attention from other disciplines.

Keywords: Digital Humanities; Knowledge Diffusion; Interdisciplinarity