

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

沈雪莹 欧石燕 卢彤彤
(南京大学信息管理学院, 南京 210023)

摘要: 中国古代文人是传统文化的创造者和传播者, 对他们的传记以及作品的内涵进行语义组织有助于提升历史文化知识服务水平。采用本体与语义网技术对文人生平内容进行语义表示, 重点关注文人的生平事件与作品内涵, 从文人生平的时空维度出发, 构建文人的生平事件以及作品信息之间的语义关联, 以期通过可视化技术对文人生平经历以及思想演化特征进行呈现和分析。实例部分, 以李白和杜甫为对象, 基于构建的古代文人语义模型, 对他们的生平和作品信息进行半自动标注, 构建“李杜”生平知识图谱, 并存入neo4j图数据库, 采用可视化技术展示和分析两位文人的 interpersonal 关系、生平事件、时空轨迹、作品风格演变。关注文人生平的时空信息, 将他们的生平事件与作品特征关联起来, 从而直观地观察他们的生平经历与思想感情变化之间的联系, 为传统人文研究助力。

关键词: 传记语义组织; 生平知识图谱; 生平可视化; 作品风格演化呈现

中图分类号: G250.7 **DOI:** 10.3772/j.issn.1673-2286.2023.08.001

引文格式: 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(8): 1-14.

历史人物塑造历史, 对其进行研究是研究历史文化的必要通道, 使用数字人文技术将历史进程中的历史人物数字化、数据化和智能化, 有助于文本中的历史人物“活起来”, 传承中华优秀传统文化。近年来, 随着数字技术以及互联网的应用普及, 图书馆、档案馆建设了大量含有历史人物数据的资源库, 其中有非结构化的古代典籍如《史记》等, 也有(半)结构化的重要历史人物传记数据库, 如英国牛津国家传记词典(Oxford Dictionary of National Biography, ODNB)^[1]、美国国家传记(American National Biography Online, ANBO)^[2]、芬兰国家传记(National Biography of Finland, NBF)^[3]以及中国历代人物传记资料库(China Biographical Database, CBDB)^[4]等, 还有基于关联数据的历史人物知识图谱, 如北京大学数字人文研究中心研究平台的朱熹年谱时空数据^[5]、“宋元学案”知识图谱^[6]等。但是, 目

前历史人物数据库主要存储历史人物的元数据, 较少深入探索历史人物生平。

中国古代文人是传统历史文化的创造者和记录者, 他们的生平有着很高的历史研究价值。因此, 本文以中国古代文人为研究对象, 采用本体与语义网技术对其生平内容进行统一语义化表示, 以期实现从传记内容出发的文人生平知识组织与分析。在构建模型时重点考虑以下三点。①从时空维度出发。人物的生平信息都是随着时间的演进而发生变化的, 从时空维度出发对文人的生平内容进行语义表示, 可以实现文人生平内容的细粒度表示。②重点关注古代文人的作品内涵表示。作品内涵反映着文人在不同时期表达的思想情感, 是最重要的体现文人生平的数据。因此, 重点关注文人的作品内涵表示, “细读”文人生平内容。③以生平事件为核心。人物生平是由一系列相互关联的事件组成的^[7-8],

以文人生平事件为中心,不仅可以实现文人生平信息的时空语义建模,还可以将作品与文人生平事件进行语义关联,呈现出文人在不同时空下的人生经历和思想^[9]。选取我国历史上生平时空特征明显的两位著名诗人李白和杜甫作为实证对象,对李杜生平内容、作品等信息进行时空语义化表示,构建李杜生平知识图谱,采用可视化技术探索两位诗人的生平画像,从而更好地理解两位诗人的生平经历与作品时代特征,以期为研究历史人物的人文学者提供新的研究范式。

1 语义网技术下的人物传记相关研究

随着数字人文的广泛应用,采用语义网和知识图谱技术对人物生平信息进行语义表示和组织已成为人物传记学新的研究范式。语义网技术下的人物传记研究可分为人物本体构建和关联数据技术下的人物传记学研究。

1.1 人物本体构建相关研究

人物生平信息语义组织的前提是人物本体构建,目前人物本体主要经历了3个发展阶段,分别是人物基本信息本体、人物关系本体和人物生平事件本体。

人物基本信息本体只描述人物姓名、性别、生卒年等基本属性,代表性本体有FOAF本体^[10]以及schema.org中的人物实体标志方案^[11]。之后,学者开始关注人物关系并定义了不同类型的人物关系,譬如: Davis等^[12]对FOAF本体中的foaf:knows属性进行扩展,构建了描述各类人物关系的Relationship词表;英国伦敦国王学院数字人文系以叙利亚传记为研究对象,将人物关系分为10类,并细分了54种人物关系^[13];等等。

随着人物描述的深入,一些研究者开始在人物静态信息基础上关注动态事件信息。首先,对生平事件进行了分类描述。譬如, Ramos^[14]构建了BLO (Biography Light Ontology)对人物生平事件进行语义描述,将事件分为4类,即健康相关(如出生、死亡)、位置变化(如移民)、人物之间关系变动和个人发明创造;位通等^[15]设计了人物年谱时空数据重构模型,将人物生平事件分为5类,即为学、为政、写作、生活和其他;牛力等^[16]则通过事件属性对人物工作经历(讲学事件、为官事件)和人物生活经历(出生事件、求学事件、搬迁事件、亲友变动事件)进行表示;夏翠娟等^[17]构建的家谱

本体对迁徙事件进行描述;CBDB本体^[18]对历代人物的职官事件和迁徙事件进行描述;等等。其次,对生平事件的构成元素进行描述,将事件与参与者、时间和地点等元素进行关联。譬如, Krieger等^[19]设计了用于表示名人传记信息^[20]的Biography.owl轻量级本体。以事件为核心的人物本体能够比较全面地展现人物生平经历,使得对人物生平的描述更加细致完整。

1.2 关联数据技术下的人物传记学研究

随着数字人文的发展,有关人物信息的海量数字资源被开发出来,为方便人物传记学的相关研究,国内外相关机构开始采用关联技术将分散的数据进行关联和集成。譬如,20世纪90年代,英国伦敦国王学院数字人文系创建结构化人物传记数据库开发项目SNAP,该项目使用关联数据技术将古代人物、姓名、头衔等进行关联,建立了一个权威的可查询古代人数据库^[21];2016年, In Concert项目使用关联技术将1750—1800年的伦敦音乐会日程和19世纪伦敦音乐会数据集中的作品、作品风格、音乐家等关联起来,旨在实现基于关联开放数据的音乐家传记研究等^[22];印第安纳哲学本体项目(In-diana Philosophy Ontology Project, InPhO)则以哲学思想家为研究对象,将《斯坦福哲学百科》(Stanford Encyclopedia of Philosophy, SEP)、《网络哲学百科全书》(Internet Encyclopedia of Philosophy, IEP)等中的思想家、组织以及作品等进行关联,方便用户检索和浏览哲学学者、思想和作品之间的关系^[23]。国内,夏翠娟等^[17]采用关联数据技术将散落在不同家谱中的有关胡适的人、地、时、事信息进行关联,构建了家谱知识图谱;位通等^[15]则基于时空共现,实现了对朱熹年谱中编年和系地信息的可视化分析。关联数据技术很好地促进了分散数据的集成,提升了语义检索的效率。

对于古代文人,部分学者在构建本体时也考虑了作品信息,如徐彤阳等^[24]构建的晚明曲家年谱本体、程结晶等^[25]构建的西汉经学家本体模型均将人物与其作品进行了关联。然而,这些本体只对作品的书目信息进行简单描述,并没有深入其内容和风格。罗凤珠等^[26]认为文学作品风格受地理环境的影响,将李白、杜甫和韩愈的诗作与他们游历的地点关联,对3位诗人的行吟地图进行了呈现;高劲松等^[27]则将人物与作品的情感相关联,并采用地理信息系统(Geographic Information System, GIS)技术展示了诗人的时空情感轨迹。但是,

古代文人在其作品中表达的思想情感往往与其生平经历有关, 因此通过将作品与生平事件相关联, 可以实现对文人生平内容以及作品演变特征等的直观观察。

2 中国古代文人生平本体模型设计

历史由一系列连续事件构成, 通过将其划分为连续发生的故事块, 可更好地进行历史研究或教学^[28]。在文人不同的生命阶段, 他们的作品以及参与的事件通常会反映不同的思想。基于此, 将文人生平看作“小历史”, 将其划分为多个不同时期, 并结合作品探索文人在其生平不同阶段的思想情感演变。一般基于某些重大变化划分时期, 如人物角色变化(从求学到入仕)、人物所处地点环境变化(从长安到江南)、人物所处时代变化(安史之乱前后)等。通过对古代文人的传记内容进行分析, 发现人文学者主要根据人物角色变化来划分文人生平。因此, 将引发文人角色变化的重大事件(如科举、任官等)作为他们生平时期的分割点。通过对生平不同时期的划分, 对不同阶段的作品以

及事件进行描述和关联, 分析文人不同时期文学作品的风格、主题、情感等, 进而更好地揭示文人生平中的思想演变。

重点参考国际文献工作委员会的概念参考模型CIDOC CRM设计中国古代文人生平本体(Chinese Ancient Literati's Biography Ontology, LBO)。CIDOC CRM是文化遗产领域的经典本体模型, 主要用于对不同领域文化遗产资源的描述和整合^[29]。该模型的最大特点是以事件为中心, 而古代文人生平的核心也是事件, 因此主要复用该模型对文人生平事件进行语义描述。

首先将文人人生看作一个时空卷(cidoc:Spacetime Volume), 将其生平经历置于该时空卷中, 然后依据引发文人角色发生变化的重大事件将时空卷划分为多个不同时期(cidoc:Period), 每个时期包含一个或多个生平事件(cidoc:Event), 并以事件为中心将时间和地点与时空卷进行间接关联, 以表示人物一生涉及的时间范围和地理区域(如出生地、居住地、死亡地等)。中国古代文人生平本体的整体架构见图1。

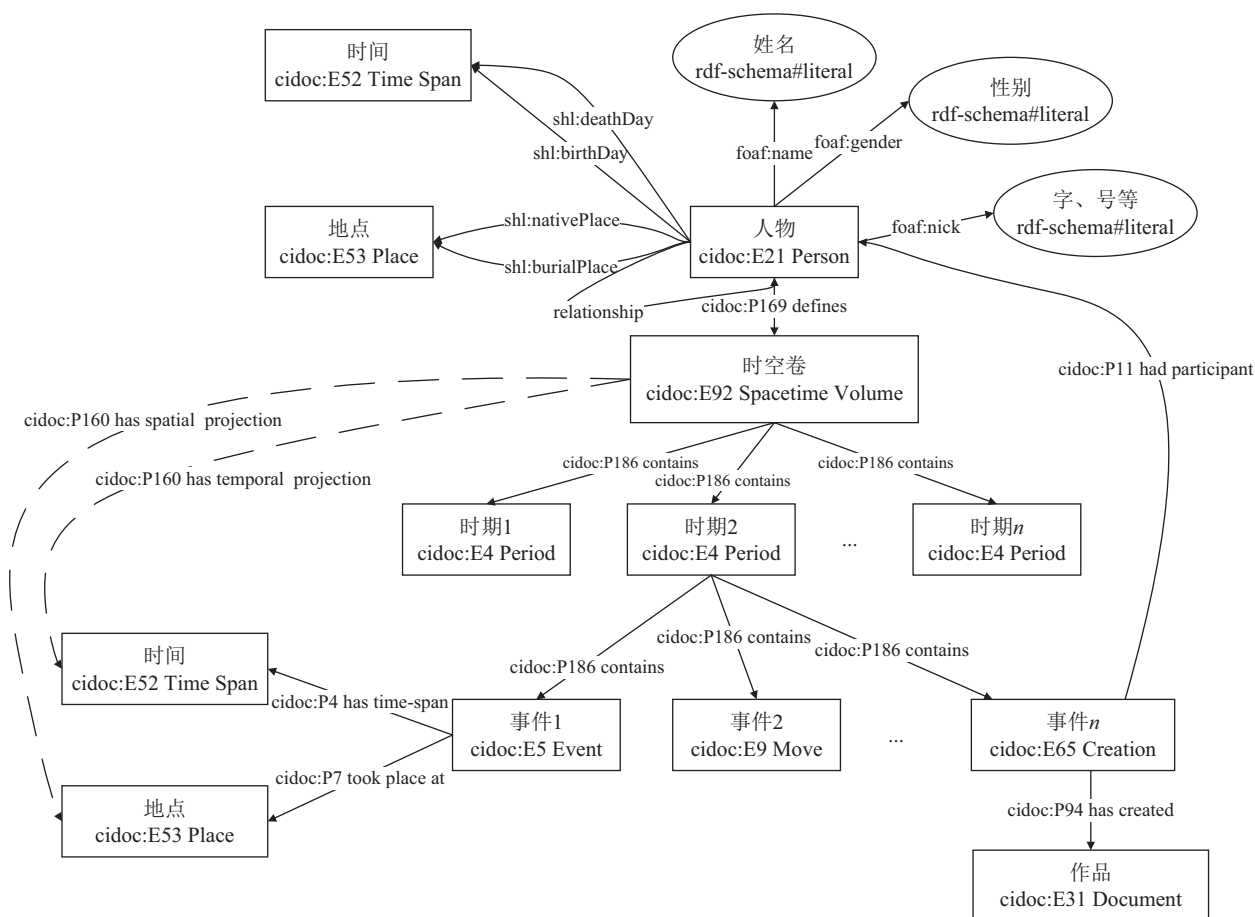


图1 中国古代文人生平本体的整体架构

2.1 文人基本属性

《中国文学家大辞典》^[30]、CBDB等对文人基本信息的描述主要包含姓名、字、生卒年/地等静态属性。为实现描述的规范统一性,复用FOAF本体(前缀foaf)、schema.org中关于人物实体的标注方案(前缀schema)以及上海图书馆构建的CBDB本体(前缀shl)中的词汇来描述人物基本信息,并新增部分自定义词汇(前缀lbo),描述的属性有:姓名(foaf:name)、字(shl:courtesyName)、号(foaf:nick)、性别(schema:gender)、姓(schema:familyName)、名(foaf:givenName)、籍贯(shl:nativePlace)、朝代(shl:dynasty)、出生日期(schema:birthDay)、出生地点(schema:birthPlace)、逝世日期(schema:deathDay)、去世地点(schema:deathPlace)、埋葬地点(shl:burialPlace)等。

借鉴陈涛等^[31]在CBDB本体中对人物关系的分类,将文人的人物关系分为家族亲属关系(lbo:

kinshipRelation)和社会关系(lbo:socialRelation)两大类。在此基础上,复任由Davis等^[12]设计的Relationship词表(前缀rel)对文人的人物关系进行细分,其中:家族亲属关系包括父母(rel:parentOf)、子女(rel:childOf)、配偶(rel:spouseOf)、后代(rel:descendantOf)、先祖(rel:ancestorOf)等;社会关系包括老师(lbo:teacherOf)、学生(lbo:studentOf)、敌人(rel:enemyOf)、朋友(rel:friendOf)、同僚(rel:colleagueOf)等。

2.2 生平事件子模型

人物生平是由一系列相互关联的事件组成的^[7],事件作为人物生平最小的知识单元,包含时间、地点与人物角色等要素。因此,主要复用CIDOC CRM及其衍生模型BIO CRM来描述中国古代文人生平事件子模型,结构如图2所示。

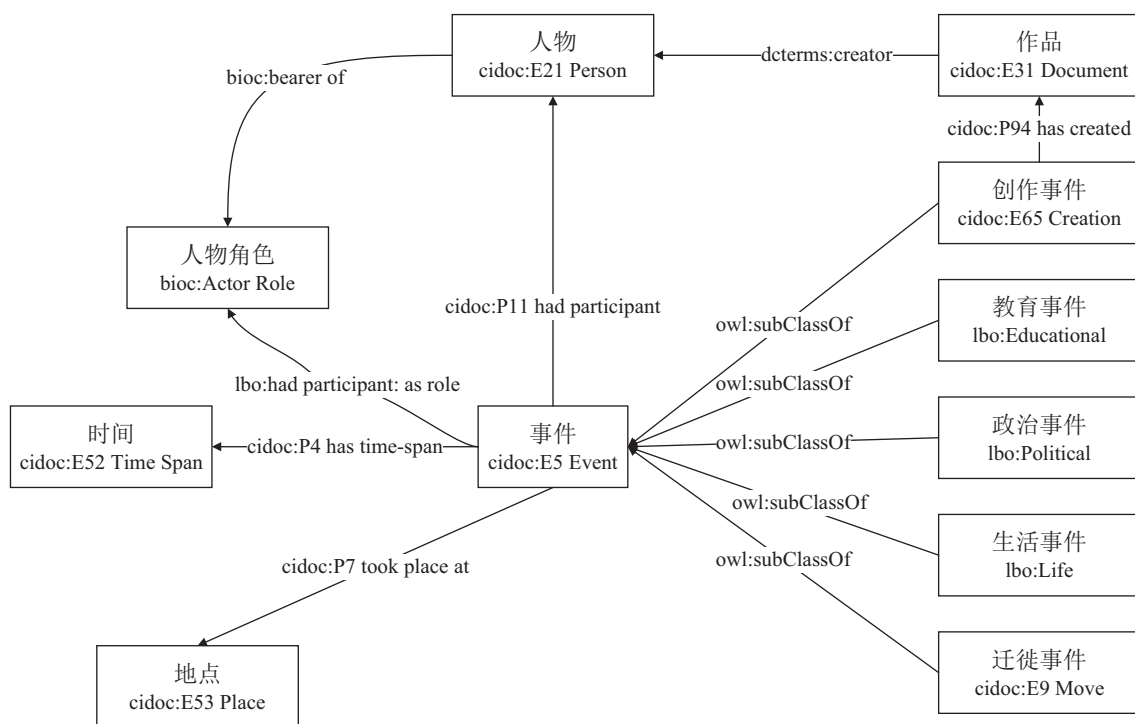


图2 中国古代文人生平事件子模型

人物生平还涉及多种不同类型的事件。在牛力等^[32]及BIO CRM对人物生平事件分类的基础上,根据文人在事件中担任的角色,将古代文人参与的事件划分为教育事件(lbo:Education)、政治事件(lbo:Politics)、

创作事件(cidoc:Creation)、生活事件(lbo:Life)以及迁徙事件(cidoc:Move)。其中:教育事件是指入塾(bio:Enrolment)、出塾(bio:Graduation)或传道授业解惑(lbo:Instruction)等与教育有关的

事件; 政治事件是指为官 (lbo:Appoinment)、贬官 (bio:Demotion)、升官 (bio:Promotion) 和入狱 (lbo:Imprison) 等与政治活动有关的事件, 如“杜甫在至德二载五月十六于凤翔授左拾遗”; 创作事件是产生、创造或发表某个文学作品 (如诗作) 等, 主要涉及创作事件的产物——著作; 生活事件则是指文人在生活中遇到的出生 (cidoc:Birth)、死亡 (cidoc:Death)、结婚 (bio:Marriage) 和离婚 (bio:Divorce) 等事件; 迁徙事件主要指文人居家地址的变化, 主要涉及迁出地、迁入地等, 如“杜甫一家在天宝十三载秋从洛阳移居到奉先”。

事件的参与者不仅有传主也有其他人物, 如“李白让高力士脱靴”这一事件中包含传主“李白”和其他人物“高力士”。由于构建的本体是对传主生平的描述, 将事件的参与者分为传主和其他人物, 并只考虑传主在不同类型事件中的角色, 如“杜甫被肃宗授为左拾遗”这一政治事件中, 传主杜甫的角色是“左拾遗”, “肃宗”则是其他人物。CIDOC CRM采用了属性cidoc:P11描述事件参与者, 但是无法描述参与者在事件中承担的角色。芬兰语义计算研究组为解决事件中人物角色问题, 在CIDOC CRM基础上构建了BIO CRM, 添加了人物角色类 (bioc:Actor Role), 将属性cidoc:P11的值域改为人物角色类, 并添加属性bioc:bearer of来连接人物类 (cidoc:Person) 与人物角色类, 以表示人物承担的角色。基于此, 仍采用CIDOC CRM的cidoc:P11属性连

接事件类和人物类, 以表示事件的参与者。为描述传主在生平事件中承担的角色, 在属性cidoc:P11的基础上扩展了子属性lbo:had participant:as role, 连接事件类和人物角色类。仍使用BIO CRM中的bioc:bearer of属性连接人物类与人物角色类。

地点类与时间类描述了人物生平的时空信息。通过对文人生平事件关联的时间和地点信息进行描述能够很好地挖掘生平涉及的编年、系地要素, 进而方便研究者探索文人生平的时空特征。关于地点, 复用全球地理数据库GeoNames本体^[33]以及WGS84地理位置本体^[34]来表示地点对应的古今地名以及经纬度; 关于时间, 复用Time本体^[35]来对事件发生的时间进行限定, 并使用属性time:has TRS来表示时间的参照系。

2.3 文学作品子模型

文学作品是文人在创作事件中产生的成果, 是反映他们思想和情感的重要工具。参考CIDOC CRM, 将作品视为创作事件的产物。由于研究的实证对象是唐代诗人, 作品形式是诗歌, 主要对诗歌作品进行语义描述。唐诗主要由形式和内容两大部分组成, 基于周莉娜等^[36]设计的唐诗知识本体模型, 复用通用元数据标准DCTERMS、书目本体BIBO等构建古代诗歌本体子模型, 如图3所示。

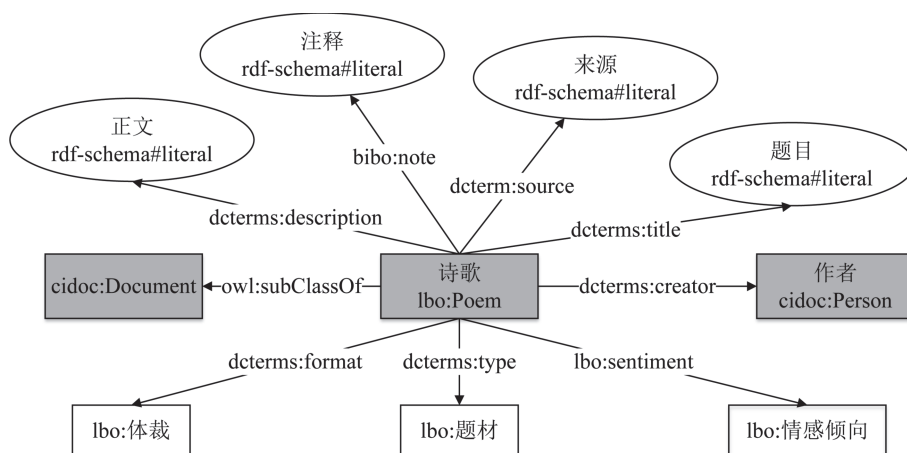


图3 古代诗歌本体子模型

从形式上看, 诗歌具有标题、作者、体裁、题材等属性。其中, 诗歌体裁一般是指其外在形式的格律, 采用属性dcterms:format描述。根据蒋绍愚^[37]对诗歌

格律的梳理, 将唐诗分为古体诗和近体诗两大类, 近体诗又可以分为律诗和绝句。诗歌题材可以看作诗歌的一种类型, 采用属性dcterms:type描述。遵循张曙霞^[38]

对诗歌题材的划分,将唐诗题材分为6类,即山水田园诗、怀古咏史诗、即事(景)感怀诗、边塞战争诗、咏物诗和闺情诗。从内容上看,诗歌有正文、注释、情感等属性。诗歌往往有着不同的情感倾向^[39],如表达报效国家的正面情感倾向和报国无路的负面情感倾向。因此,将诗歌情感倾向分为正面、中性和负面三大类,采用属性lbo:sentiment来描述,其中:正面情感一般是指诗人热爱自然与生活、心怀抱负等情感;负面情感则主要是指诗人怀才不遇、忧国忧民、相思别离与飘零孤苦等情感;中性情感主要是指诗人淡泊宁静的处世态度。

3 李杜生平知识图谱构建

众所周知,诗仙李白和诗圣杜甫是我国历史上著名文人的代表,他们生活在唐朝由盛到衰的历史转折时期,都经历了始于天宝十四年的安史之乱,人生经历发生了巨大的变化,留下了大量具有时代背景的诗歌类作品,为文人作品风格分析提供了资料来源。因此,以李杜为实例对象,基于中国古代文人生平本体对语料中的李杜生平信息进行整理,抽取出人物、事件、地点、时间等实体及其关系,重点实现以事件为中心的李杜生平经历语义标注,从而生成李杜生平知识图谱。

3.1 语料采集

选用多种形式的文本作为李杜生平信息的语料集,见表1。采用光学字符识别(Optical Character Recognition, OCR)技术,将来源语料均转换为纯文本格式,以供后续数据分析和处理。

表1 语料集介绍

语料	形式	语料描述
《李白传》 ^[40] 、 《杜甫传》 ^[41]	非结构化文本	对李杜的生平经历进行详细描述
《全唐诗》 ^[42] 包含的李杜诗歌	半结构化文本	含有李白诗歌836首,杜甫诗歌1 149首
《李白大辞典》 ^[43] 、 《杜甫大辞典》 ^[44]	半结构化文本	描述李杜的基本信息以及李杜诗歌的现代文注释
CBDB	结构化文本	包含李杜基本信息以及人物关系信息

3.2 文本预处理

主要采取半自动方法对相关语料进行处理。首先,基于CBDB和《李白大辞典》《杜甫大辞典》中有关地名、官职、相关人物等的总结,人工构建唐代地名、官职和人名词表作为自定义词典,为后续实体识别提供支持;然后,分别使用现代汉语分词工具jieba和古汉语自然语言处理工具“甲言”对传记语料和李杜诗歌进行分词处理。

3.3 人物生平知识识别

3.3.1 李杜基本属性和关系识别

主要从《李白大辞典》《杜甫大辞典》和CBDB中人工获取李杜的基本属性,如姓名、字、号、生卒年/地等。主要从《李白传》《杜甫传》语料和CBDB中获取李杜人物关系:首先基于人名词典在文本分析平台MARKUS对人名进行自动标记,然后根据人名所在的句子位置人工判定李杜与其他人物的关系,最后基于CBDB对李杜的人物关系进行补充。

3.3.2 李杜生平不同时期划分

通过分析《李白传》和《杜甫传》对李杜生平信息的描述,将引发李杜两人角色变化的事件作为分割点,将李白的时空卷划分为5个时期,将杜甫的时空卷划分为4个时期,具体描述见表2。

3.3.3 李杜生平事件识别

主要采用规则的方法识别语料中的事件。首先,以句子为单位,采用哈工大自然语言工具LTP对句子进行句法解析,识别句子中的谓词作为事件触发词,总结得到的事件类型以及其对应的线索词见表3;然后,基于线索词识别语料中的事件。

针对地点,基于构建的地名词表,使用字符串匹配法从语料中识别地名,并借助中国历史地理信息平台地名查询功能^[45]获取地点所对应的经纬度。

针对时间,采用正则表达式构建抽取规则:年号+()/年/春/孟春/仲春/暮春/季春/孟夏/仲夏/季夏/秋/孟秋/仲秋/暮秋/季秋/冬/孟冬/仲冬/季冬……

表2 李白和杜甫生平不同时期的划分与描述

时空卷	时 期	时间范围	描 述
李白时空卷	李白蜀中读书和任侠时期	705—723年	李白在蜀中读书的青年时期
	李白追求功业时期	724—742年	李白希望通过干谒入仕的时期
	李白供奉翰林时期	742—744年	李白被召入长安供奉翰林时期
	李白漫游时期	744—756年	李白辞官离开长安, 开始漫游时期
	李白报国蒙冤时期	756—762年	李白入永王幕僚, 后获罪流放的晚年时期
杜甫时空卷	杜甫漫游时期	730—745年	杜甫年少游历吴越、齐赵时期
	杜甫困居长安时期	746—756年	杜甫在长安10年, 希望求得官职时期
	杜甫为官时期	756—759年	杜甫终得官职, 任职左拾遗、被贬华州等时期
	杜甫西南漂泊时期	759—770年	杜甫辞官之后在蜀中、潭州等西南一带漂泊时期

注: 时间范围为大概范围, 因此年份有重合。

表3 不同事件类型与其对应的线索词

事件类型	线索词
创作事件	作、撰、写、《》
教育事件	入塾、应试、考试、落第
政治事件	任职、奉召、贬、授、弃官、罢官、疏救、表荐、入……幕、获罪、遭……罚、流放、旋代、谏
生活事件	干谒、省家、投奔、奔赴、至/去+地点、登+地点、漂泊、病逝、卒
迁徙事件	居、住、逃往、移居、移家、搬

从语料中抽取出的时间通常是年号纪年, 且描述精度不一: 有的以年为单位, 如“乾元二年(759)”; 有的以季节为单位, 如“乾元二年春”或“乾元二年孟春”; 有的以月为单位, 如“乾元元年五月”。因此, 统一将其转换为“年-月-日”(YYYY-MM-DD)形式的公元纪年, 且参照对中国古代四季(又称四时)的描述^[46]将四季转换为对应的月份, 即将春季转换为1—3月, 将夏季转换为4—6月, 将秋季转换为7—9月, 将冬季转换为10—12月。

针对事件的产物——唐诗, 从《全唐诗》中获取李杜的诗歌, 并基于规则的方法从《李白大辞典》和《杜甫大辞典》中识别出两人诗歌的创作地点和时间, 共得到1 144首杜甫诗歌和538首李白诗歌的创作时间和地点信息。

针对传主在事件中的人物角色, 需要结合上下文信息和背景知识进行判断和分析。为保证识别的准确性, 人工判断传主角色。另外, 在事件识别的基础上, 为后续清楚展示人物的生平事件, 采用lbo:next对李杜参与事件的先后顺序进行标注。

基于此, 对李白“赐金放还”事件实例进行语义化表示, 如图4所示。

图4表示的李白“赐金放还”事件所属时期是“李白供奉翰林时期”, 事件类型是“政治事件”, 该事件也是其生平中的重大转折事件, 标志着李白政治角色的结束。事件参与者有传主“李白”和其他人物“唐玄宗”, 其中传主担任的角色是“翰林”。事件发生时间是“天宝三载三月”, 对应公元时间是“744-03”。发生地点是“长安”, 对应的当今官方地名是“西安”, 经纬度为(108.906 97, 34.246 42)。

3.3.4 唐诗属性识别

语料中关于诗歌的题材、体裁以及情感倾向等的描述较少, 主要从百度百科的诗歌词条中获取唐诗的题材和体裁信息, 并采用情感词匹配法获取唐诗的情感倾向信息。首先采用八爪鱼爬虫软件, 自动从百度百科中爬取诗歌的相应词条, 分别获得481首李白诗歌和993首杜甫诗歌的词条, 然后对词条Infobox中的诗歌题材和体裁进行整理。

为有效识别唐诗的情感倾向, 人工构建了诗歌情感词表。首先, 随机选择150首李杜诗歌, 人工提取其中的情感词并进行情感倾向分类, 共得到包含76个正面情感词、81个负面情感词和42个中性情感词的诗歌情感词表, 见表4。然后, 将诗歌情感词表与分词后的唐诗进行匹配, 计算每首诗歌中出现的正面、负面和中性情感词数量, 以数量最多的词的情感倾向代表诗歌的情感倾向。为验证该方法的精确性, 随机选取62首诗歌, 邀

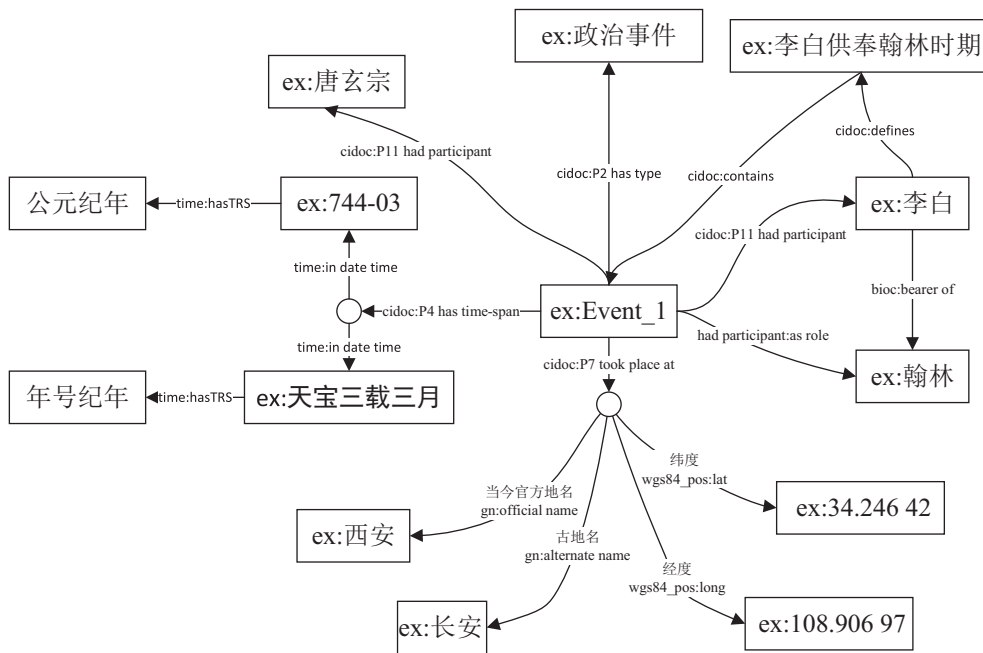


图4 李白生平中“赐金放还”事件的语义化表示

请两位古汉语专业研究生对诗歌情感倾向进行人工标注。测试结果表明，基于情感词识别诗歌情感倾向的准确率为74%。最后，邀请一位古汉语专业研究生对诗歌的题材、体裁及情感倾向进行完善，以保证标注数据的准确性。

表4 诗歌情感倾向及其相对应的情感词示例

情感倾向	情感词示例
正面	彩错、彩云、元气、桃花、春江、绿水、大鹏、白鹭、九霄、潇洒、清风、翠色、开颜、桃李、秀气、喜
负面	悲(悲风、独悲)、病客、耆老、白首、干戈、老病、浮云、萧瑟、愁(愁云、愁苦、愁心)、暮秋、叹息、断肠、蹉跎、惆怅、岁月、多病
中性	仙人、躬耕、陶潜、陶然、果园、汀烟、隐居、隐逸、隐者、畦蔬、幽径、老农、神仙、仙草、蓬莱

3.4 知识图谱存储

将抽取出的李杜生平信息以资源描述框架(Resource Description Framework, RDF)的三元组形式表示，并以CSV格式导入neo4j图数据库进行存储。neo4j是以图结构进行存储和查询的图数据库，其中实体表示为节点(Node)，关系表示为节点之间的边(Edge)。通过

neo4j提供的查询语言cypher，可以对李杜生平知识图谱中的特定信息进行查询和可视化分析。

4 基于生平知识图谱的文人信息可视化分析

李杜生平知识图谱中不仅包含两位诗人的基本信息，还包含着大量时空数据以及不同时期产生的作品。因此，首先对李杜的人物关系网络 and 生平事件进行可视化呈现，然后基于时空信息和诗歌来探索李杜的时空画像，包括李杜不同时期的时空轨迹和李杜创作风格演化分析。

4.1 李杜人物关系可视化分析

在neo4j图数据库中查询李杜的人物关系，使用Gephi软件对两人的人物关系网络进行可视化呈现，如图5所示。

整体来看，李白和杜甫两人的人脉非常广泛，且有着共同的好友，如“贾至”“高适”“张镐”等。从李白的关系网络可以发现，其妻子“许氏”“宗氏”分别是担任过宰相之职的“许圉师”“宗楚客”的孙女；其朋友有当朝重臣“孔巢父”“贺知章”等，也有著名诗人“王昌龄”“孟浩然”“杜甫”等，还有道友“元丹丘”

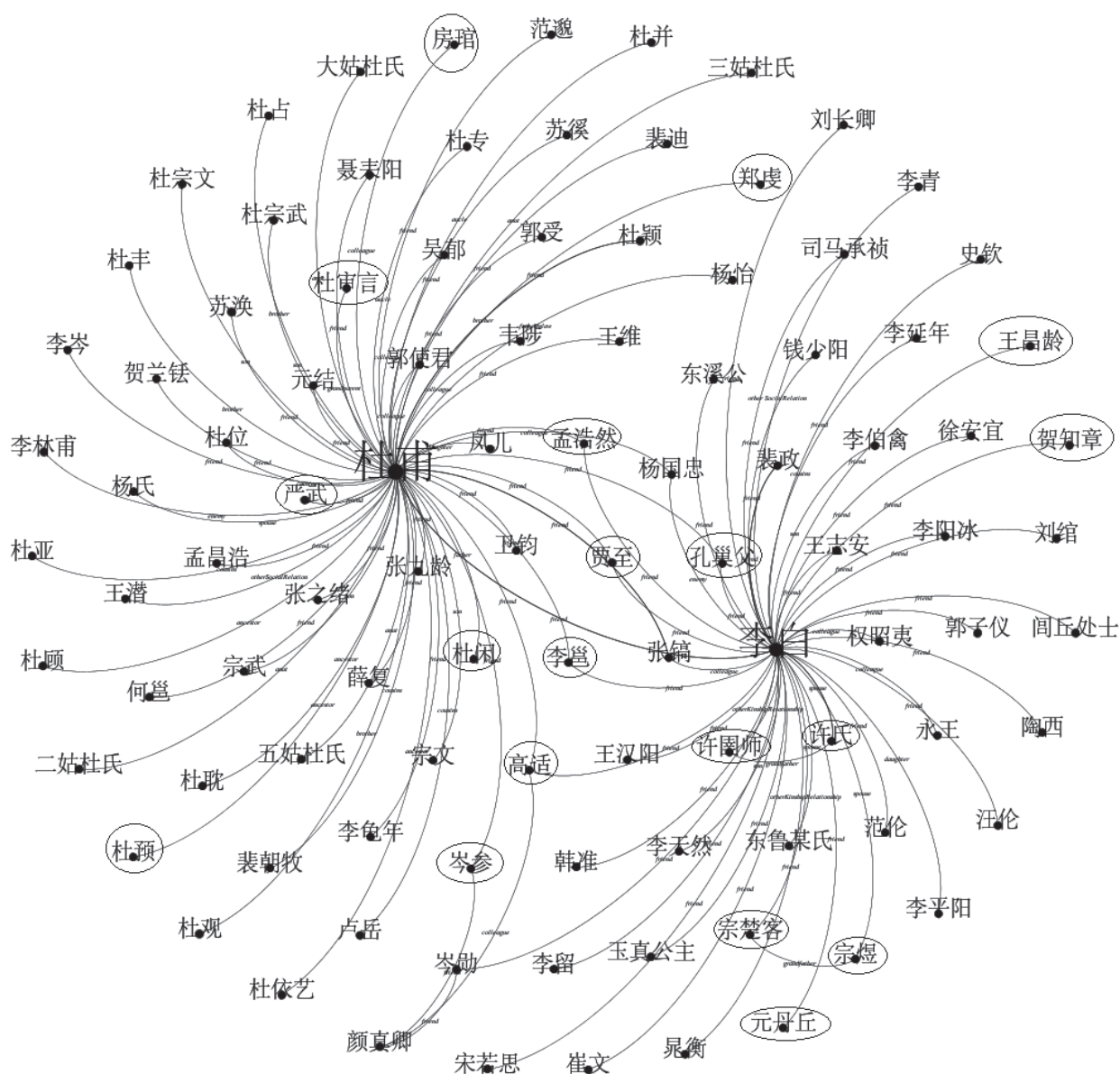


图5 李杜人物关系网络可视化结果

等,交友范围广泛。从杜甫的人物关系网络可以发现,家族人物记载完整,且其家族为仕宦家族,如先祖“杜预”、祖父“杜审言”和父亲“杜闲”均有为官记载;在社会关系方面,其朋友多是分布在各地的官员,如“严武”“房琯”“贾至”“高适”“岑参”“郑虔”等,其朋友集中在官宦阶层。

4.2 文人生平事件可视化

通过cypher查询neo4j中的文人的生平事件,以时间轴的形式对其进行可视化,可以直观地观

察到文人在各个时期参与的事件。以李白为例,对其在各个时期参与的事件进行可视化,结果见图6。

从图6可以看出,李白生平时期的重要分割点是由政治事件引发的,且其在不同时期创作的作品与生平事件密切相关。譬如,742年“奉召入京”是供奉翰林时期的开始,“被赐金放还”是该时期的结束和漫游时期的开始;该时期其多参与政治事件,创作的作品(如《清平调》《宫中行乐词》)多与宫中活动相关。在“白帝城遇赦”之后,立即创作了《早发白帝城》一诗以表达其喜悦的心情。

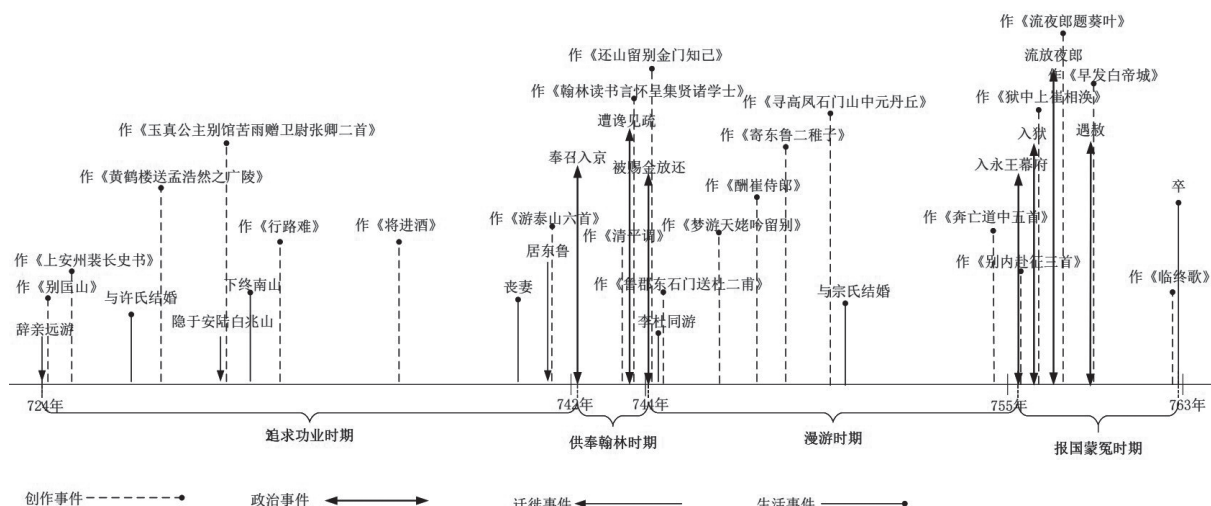


图6 李白各时期参与事件可视化结果（部分）

4.3 李杜时空画像分析

4.3.1 李杜生平时空轨迹分析

基于生平事件发生地点的经纬坐标，采用GIS技术将李杜时空轨迹分别可视化，见图7（a）和图7（b）。

从图7（a）可以看出，李白一生涉足很多地点，且其多以某个地区为中心开展地域性活动：蜀中读书和任侠时期在蜀中一带活动，追求功业时期在安陆的周围活动，供奉翰林时期在长安一带活动，漫游时期则在江南一带游历，报国蒙冤时期在浔阳和洞庭湖一带漂泊，于当涂去世。

从图7（b）可以看出，杜甫在前期开展地域性活动，后期其活动则是线性的，且受安史之乱的影响明显：安史之乱之前，在漫游时期在东都等地游历，在困居长安时期则主要在长安活动；安史之乱之后，杜甫开始了流亡，在为官时期主要在凤翔、华州活动，之后在西南流离漂泊，病逝于潭州一带。

4.3.2 李杜不同生平时期创作风格演化分析

将李杜诗歌与其生平经历结合，从诗歌形式和内容两个角度出发，展示李杜不同生平时期诗歌创作风格的演变情况，进而展示文人作品与其生平经历的关联。

首先，对neo4j中李杜两人各时期的诗歌作品及其对应的题材、体裁以及情感倾向进行统计与可视化，图8显示了李杜两人不同生平时期的诗歌题材和体裁，

图9显示了李杜不同生平时期诗歌的情感倾向。

整体来看，李杜两人在不同生平时期创作的诗歌题材有着较为明显的区别，但是诗歌体裁并没有明显变化，且随着时间演变，诗歌的情感倾向发生变化。具体来说，李白在蜀中读书和任侠时期，创作的诗歌体裁以赋为主，题材以山水田园诗居多，诗歌情感均为正面，这个时期的李白对未来充满了美好愿景。在追求功业时期，李白的干谒诗和闺情诗占比增多，诗歌情感倾向则是喜忧参半：有想通过干谒以施展才华的正面情感，也有因干谒活动不理想而苦闷的负面情感，其中正面情感多于负面，这也表明了这个时期的李白是胸怀抱负的。在供奉翰林时期，其山水田园诗占比明显减少，诗歌情感倾向也是喜忧参半：有其初入长安得到皇帝赏识的意气风发，也有其在长安因抱负不得施展而产生的悲伤与无奈，遂产生了归隐之心（中性）。在漫游时期和报国蒙冤时期，安史之乱爆发，边塞战争诗占比明显增多，诗歌中负面情感较多，其因入狱、流放事件等产生负面情感。

在漫游时期，杜甫诗歌题材以山水田园诗为主，其情感倾向多为正面：在困居长安时期，干谒诗占比增加，山水田园诗占比减少，负面情感占比开始增加，主要包含困居长安时经历的人生辛酸以及社会动荡下的忧民等情感。在为官时期，安史之乱爆发，边塞战争诗占比明显增加，负面情感占比也明显增加，诗歌主要表达对民生、国家等的担忧。在西南漂泊时期，其诗歌多是即事（景）感怀诗，整体上诗歌以负面情感为主：有其老病孤愁的痛苦、流离漂泊中的相思别离，以及国破家亡的离愁别恨等。其中，中性情感诗歌占比相对来说

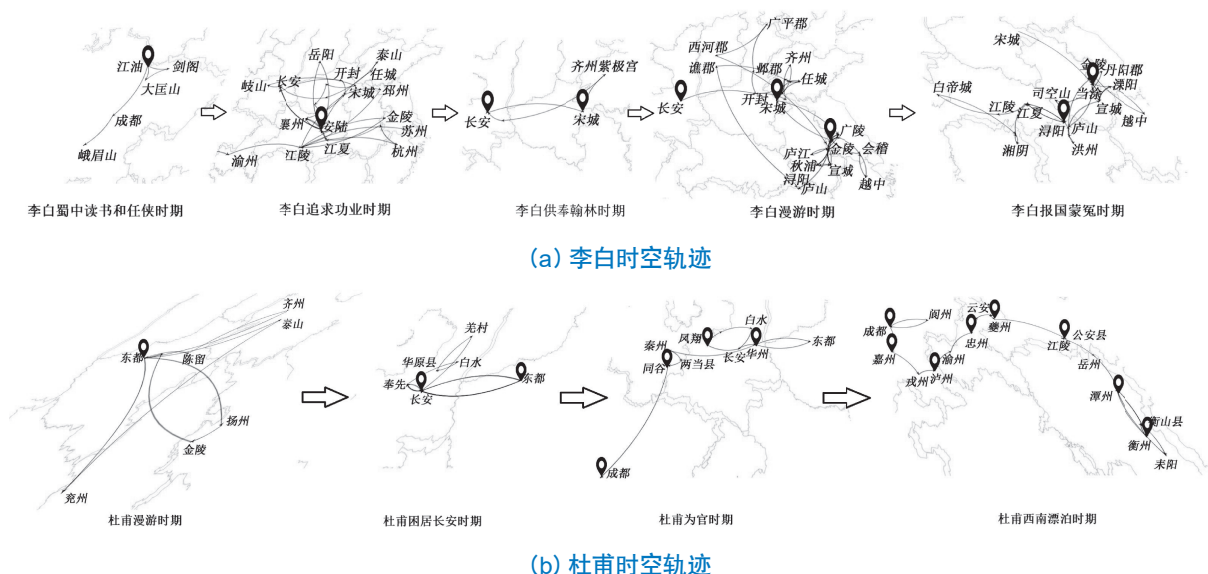


图7 李杜时空轨迹可视化结果

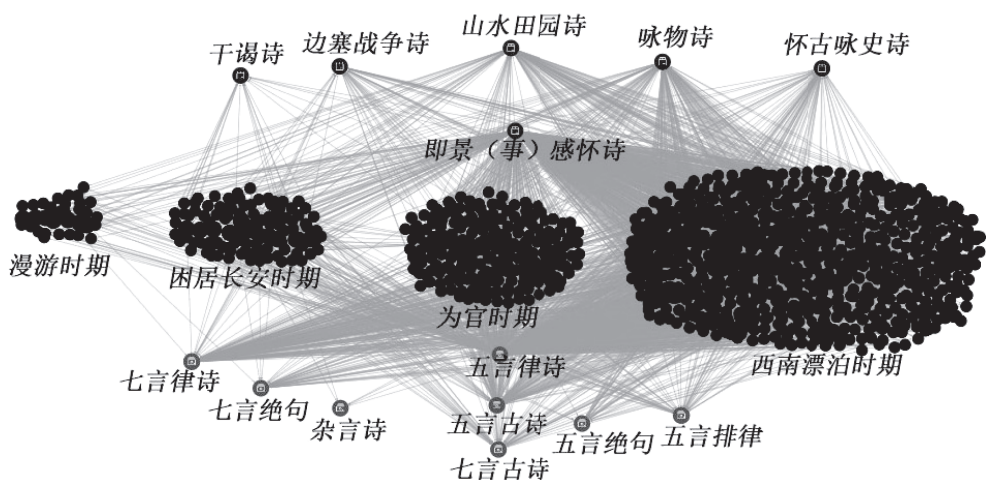
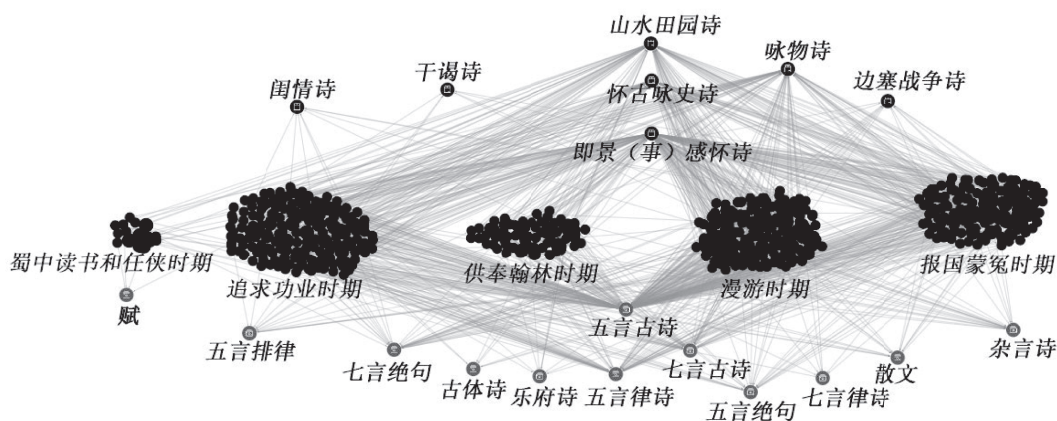


图8 李杜不同生平时期的诗歌题材和体裁

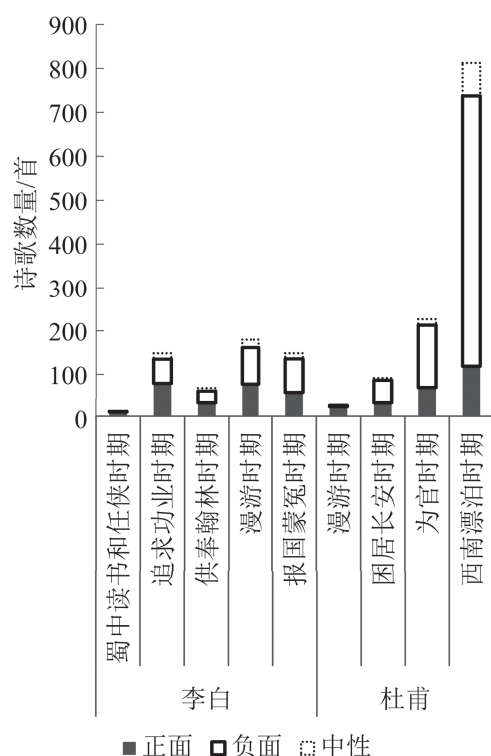


图9 李杜不同生平时期的诗歌情感倾向统计

有所增加,这表明其辞官入蜀后对生活的热爱。总体来看,杜甫后期创作诗歌的情感基调是悲伤的。

李杜两人创作的诗歌体裁基本是五言诗和七言诗,在他们整个生平时期没有明显变化,这说明五言诗和七言诗在盛唐时期盛行。

综上,通过将李杜两人的作品与其生平经历进行关联,对李杜两人不同生平时期的诗歌题材和情感倾向进行呈现,可以观察文人的生平经历与文学作品的联系以及文学作品的内容和风格反映的生平经历和诗人所在社会环境的联系。

5 结语

本研究以中国古代文人为对象,采用语义网和知识图谱技术对文人传记内容进行语义描述与组织,利用可视化技术对文人的生平信息进行可视化分析,以呈现文人的生平画像,进而帮助读者更好地理解文人的生平经历与作品时代特征。设计了描述中国古代文人传记内容的概念模型:从文人生平的时空维度出发,将他们的生平事件与作品进行时空语义关联,并纳入文人生平本体,进而探索文人作品与其生平经历之间的相关关系。在此基础上,以李白和杜甫为例,构建了李杜生

平知识图谱,并采用Gephi和GIS等可视化技术对李杜的人物关系、时空轨迹、生平事件、诗歌创作风格进行了可视化呈现与分析。结果表明,通过融合事件与作品信息对文人生平进行语义描述,可以直观呈现文人生平,进而直观地观察文人生平事件与文学作品创作风格的联系。

本研究主要采用半自动的方法对文人生平信息以及作品进行标注,后续将探索基于深度学习的自动化方法来构建古代文人生平知识图谱。

参考文献

- [1] Oxford Dictionary of National Biography[EB/OL]. [2023-02-17]. <https://www.oxforddnb.com/>.
- [2] American National Biography Online[EB/OL]. [2023-02-17]. <https://www.anb.org/>.
- [3] National Biography of Finland[EB/OL]. [2023-04-13]. <https://seco.cs.aalto.fi/projects/biografiasampo/>.
- [4] 中国历代人物传记资料库[EB/OL]. [2023-02-17]. <https://www.dhlib.cn/site/cyberinfrastructure/database/1760.html>.
- [5] 朱子年谱可视化系统[EB/OL]. [2023-01-19]. <https://nianpu.pkudh.org/>.
- [6] 宋元学案知识图谱[EB/OL]. [2023-01-21]. <https://syxa.pkudh.org/>.
- [7] DAVIS I, GALBRAITH D. BIO: a vocabulary for biographical information[EB/OL]. [2022-09-01]. <https://vocab.org/bio/#event>.
- [8] TUOMINEN J, HYVONEN E, LESKINE P. BIO CRM: a data model for representing biographical data for prosopographical research[C]//Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017). CEUR Workshop Proceedings, 2018: 59-66.
- [9] MAYR E, SALISU S M, FILIPOV V, et al. Visualizing biographical trajectories by historical artifacts: a case study based on the photography collection of Charles W. Cushman[C]//Proceedings of the Biographical Data in a Digital World 2019, 2019.
- [10] FOAF vocabulary specification[EB/OL]. [2023-02-17]. <http://xmlns.com/foaf/0.1/>.
- [11] Person-A schema.org type[EB/OL]. [2023-02-17]. <https://>

- schema.org/Person.
- [12] DAVIS I, VITIELLO E Jr. Relationship: a vocabulary for describing relationships between people[EB/OL]. [2023-02-27]. <https://vocab.org/relationship/>.
- [13] SCHWARTZ D L, GIBSON N P, TORABI K. Modeling a born-digital factoid prosopography using the TEI and linked data[EB/OL]. [2023-05-06]. <https://journals.openedition.org/jtei/3979>.
- [14] RAMOS M R. Biography Light Ontology: an open vocabulary for encoding biographic texts[J]. Technical Report, 2009: 1-23.
- [15] 位通, 桑宇辰, 史睿. 基于知识重构的年谱时空可视化呈现: 以《朱熹年谱长编》为例[J]. 中国图书馆学报, 2022, 48 (2): 62-75.
- [16] 牛力, 高晨翔, 张宇锋, 等. 发现、重构与故事化: 数字人文视角下档案研究的路径与方法[J]. 中国图书馆学报, 2021, 47 (1): 88-107.
- [17] 夏翠娟, 张磊. 关联数据在家谱数字人文服务中的应用[J]. 图书馆杂志, 2016, 35 (10): 26-34.
- [18] 上海图书馆. 上海图书馆CBDB本体[EB/OL]. [2023-03-12]. <http://www.usources.cn/ont/ontology/tree?g=http://ont.library.sh.cn/graph/cbdb>.
- [19] KRIEGER H, DECLERCK T. An OWL ontology for biographical knowledge: representing time-dependent factual knowledge[C]//Proceedings of the First Conference on Biographical Data in a Digital World 2015, 2015: 101-110.
- [20] ADOLPHS P, CHENG X, KLÜWER T, et al. Question answering biographic information and social network powered by the semantic web[C]//Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010.
- [21] KCL's Department of Digital Humanities. SNAP: DRGN project[EB/OL]. [2022-11-27]. <https://snapdrgn.net/about.html>.
- [22] In Concert[EB/OL]. [2022-09-07]. <https://inconcert.datatodata.com/>.
- [23] BUCKNER C, NIEPERT M, ALLEN C. From encyclopedia to ontology: toward dynamic representation of the discipline of philosophy[J]. Synthese, 2011, 182 (2): 205-233.
- [24] 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.
- [25] 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.
- [26] 罗凤珠, 郑锦全, 范毅军, 等. 唐代诗人行吟地图建构: 李白、杜甫、韩愈[C]//第四届中国古籍数字化国际学术研讨会, 2013.
- [27] 高劲松, 张强, 李帅珂, 等. 数字人文视域下诗人的时空情感轨迹研究: 以李白为例[J]. 数据分析与知识发现, 2022, 6 (9): 27-39.
- [28] SHAW R B. Events and Periods as Concepts for Organizing Historical Knowledge[M]. Berkeley: University of California, Berkeley, 2010.
- [29] CIDOC Conceptual Reference Model (CRM) [EB/OL]. [2023-01-23]. <https://cidoc-crm.org/>.
- [30] 李时人. 中国文学家大辞典(明代卷)[M]. 北京: 中华书局, 2018.
- [31] 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.
- [32] 牛力, 蒋菲, 曾静怡. 面向数字记忆的数字文档资源描述框架构建研究[J]. 档案学研究, 2019 (4): 40-49.
- [33] GeoNames Ontology[EB/OL]. [2023-01-23]. <https://www.geonames.org/ontology#>.
- [34] wgs84_pos[EB/OL]. [2023-01-23]. http://www.w3.org/2003/01/geo/wgs84_pos#.
- [35] Time Ontology in OWL[EB/OL]. [2023-03-01]. <https://www.w3.org/TR/owl-time/>.
- [36] 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.
- [37] 蒋绍愚. 唐诗语言研究[M]. 北京: 语文出版社, 2008: 5-6.
- [38] 张曙霞. 唐诗百首解读[M]. 天津: 天津古籍出版社, 2011.
- [39] 杨旭辉. 唐诗鉴赏大辞典[M]. 北京: 中华书局, 2011.
- [40] 安旗. 李白传[M]. 北京: 人民文学出版社, 2019.
- [41] 冯至. 杜甫传[M]. 天津: 百花文艺出版社, 2007.
- [42] 彭定求, 沈三曾, 杨中讷. 全唐诗[M]. 北京: 中华书局, 1960.
- [43] 郁贤皓. 李白大辞典[M]. 南宁: 广西教育出版社, 1995.
- [44] 张忠纲. 杜甫大辞典[M]. 济南: 山东教育出版社, 2009.
- [45] China Historical GIS[EB/OL]. [2023-04-02]. <https://chgis.fairbank.fas.harvard.edu/>.
- [46] 刘桂华. 论四时物候与中国古代文学四季抒情模式[J]. 荆楚学刊, 2013, 14 (6): 78-83.

作者简介

沈雪莹, 女, 博士研究生, 研究方向: 数字人文、文本挖掘、知识组织。

欧石燕, 女, 博士, 教授, 通信作者, 研究方向: 文本挖掘、数字人文、引文分析, E-mail: oushiyan@nju.edu.cn。

卢彤彤, 女, 硕士研究生, 研究方向: 数字人文、文本挖掘。

Construction and Application of Knowledge Graphs of Lives of Ancient Chinese Literati: Taking Li Bai and Du Fu as Examples

SHEN XueYing OU ShiYan LU TongTong

(School of Information Management, Nanjing University, Nanjing 210023, P. R. China)

Abstract: Ancient Chinese literati play a pivotal role as creators and disseminators of traditional culture. Organizing the semantic content of their biographies and works is instrumental in elevating the knowledge services related to historical culture. This paper utilizes ontology and semantic web technologies to semantically represent the life stories of literati, with a focus on their life events and the content of their works. Starting from the spatiotemporal dimension of literati's lives, it constructs semantic associations between the life events and work information of literati. The aim is to present and analyze the life experiences and evolutionary characteristics of their thoughts through visualization techniques. Using Li Bai and Du Fu as examples, a semi-automatic annotation of the life and work information of literati is performed based on the constructed semantic model of ancient literati, creating a biographical knowledge graph of "Li-Du", which is then stored in neo4j graph database. Visualization techniques are used to showcase and analyze interpersonal relationships, life events, temporal-spatial trajectories, and the evolution of the style of their works. By emphasizing the spatiotemporal aspects of literati's life experiences and connecting them with the characteristics of their works, we can intuitively observe the connection between their life experiences and changes in thoughts and emotions, so as to support traditional humanities studies.

Keywords: Biographical Semantic Organization; Biographical Knowledge Graph; Biography Visualization; Presentation of Style Evolution of Works

(责任编辑: 王玮)