DOI:10.13530/j.cnki.jlis.190046

知识图谱在数字人文中的应用研究*

陈 涛 刘 炜 单蓉蓉 朱庆华

摘 要 知识图谱是利用计算机存储、管理和呈现概念及其相互关系的一种技术,一经提出便很快成为工业界和学术界的研究热点,但目前对知识图谱的认知还比较混乱。依据存储方式不同,知识图谱可分为基于 RDF 存储的语义知识图谱(关联数据)和基于图数据库的广义知识图谱。语义知识图谱(关联数据)侧重于知识的发布和链接,广义知识图谱则更侧重于知识的挖掘和计算,两者之间既有共同点,又有不同之处。本文从概念层面和技术层面详细分析了两者之间的异同,指出语义知识图谱(关联数据)才是谷歌知识图谱的延续和发展。随后,提出了将知识图谱应用于数字人文研究的系统框架,并在此基础上构建了中国历代人物传记资料库的关联数据平台(CBDBLD)。该平台借助知识图谱的理念展现了人物之间丰富的亲属及社会关系,形成了特有的社会关系网络,并可通过设置推理规则来实现人物之间隐性关系的挖掘与呈现。广义知识图谱研究中丰富的图运算和关联数据的结合将会成为数字人文领域研究的下一个热点,从而开启数字人文研究的新时代。图 10。表 2。参考文献 25。

关键词 数字人文 知识图谱 关联数据 知识推理 中国历代人物传记资料库

分类号 G251 TP393

Application of Knowledge Graph in Digital Humanities

CHEN Tao, LIU Wei, SHAN Rongrong & ZHU Qinghua

ABSTRACT

Knowledge graph is a technique that uses computers to shore, manage, and present concepts and their relationships. This technique became a research hotspot in industry and academia as soon as it was proposed. However, the concept of knowledge graph was quite chaotic in this field. People often confuse Knowledge Map (KM), Knowledge Graph (KG) and Graph Database (GD). Knowledge map should be regarded more as a metrological method, so there is no detailed discussion in this paper. According to different storage methods, the knowledge graph can be divided into semantic knowledge graph (also called linked data, based on RDF storage), and generalized knowledge graph (due to graph databases). Linked data focuses on the release and linking of knowledge, while the generalized knowledge graph focuses more on the mining and calculation of knowledge. There are both commonalities and differences between the linked data and knowledge graph. This paper analyzes the similarities and differences between the two techniques from the conceptual and technical aspects, and points out that the linked data is the continuation

* 本文系国家社会科学基金项目“数字人文中图像文本资源的语义化建设与开放图谱构建研究”(编号:19BTQ024)的研究成果之一。(This article is an outcome of the project “The Study of Semantic Construction of Image Resources and Open Knowledge Graph in Digital Humanities” (No. 19BTQ024) supported by National Social Science Foundation of China.)

通信作者:刘炜,Email:wliu@libnet.sh.cn,ORCID:0000-0003-2663-7539 (Correspondence should be addressed to LIU Wei, Email: wliu@libnet.sh.cn, ORCID: 0000-0003-2663-7539)

and development of Google's knowledge graph.

In addition, this paper also proposes a system framework for applying knowledge graph to digital humanities research. Simultaneously, we also point out that digital generation, textual conversion, data extraction and intelligent construction are the main stages of research and development in the humanities field. Compared with most humanities research abroad in the textual stage, much humanities research in China are still in the digital stage, which is far from the research stage of smart data.

Based on the theoretical basis of the study of smart data of digital humanities, this paper builds a linked data platform (CBDBLD) of Chinese Biographical Database (CBDB). The seven-step method adopted in the platform construction is representative and has been used in many digital humanities research projects, which can guide the semantic construction of domestic digital humanities research. This platform contains more than 420, 000 biographical data, about 22.7 million triples, and is associated with open related datasets such as Shanghai Library Authority Name Files and VIAF (Virtual International Authority File). CBDBLD dataset contains ten categories of nearly 500 kinds of social relations. Further, this platform uses the concept of knowledge graph and visualization technology to show the rich relatives and social relations between characters. This platform forms a unique social network, and improves the dynamic interaction ability of user's experience and platform.

Knowledge computing and knowledge reasoning are the core technologies involved in the application of knowledge graph, which are widely studied in the application of generalized knowledge maps. However, little research has been done on linked data and digital humanities. Most of the digital humanities research in China uses linked data technology to publish and display metadata, which can be regarded as the basis of knowledge graph application. Nevertheless, it does not represent the whole knowledge graph research. In this paper, the CBDBLD platform uses a general rule reasoner to support user-defined rule-based reasoning which implements the mining and presentation of implicit relationships between characters. Although the current reasoning is relatively simple, it provides a new research direction for digital humanities research. The abundant graph mining and graph computing algorithms in the research of generalized knowledge atlas can be applied to the linked data, which is also the future research and practice direction of this paper's authors.

It can be said that both semantic knowledge graph and generalized knowledge graph can promote the innovation of digital humanities research methods. The combination of the two techniques will become the next hotspot in the field of digital humanities, and brings a new era of digital humanities research. 10 figs. 2 tabs. 25 refs.

KEY WORDS

Digital humanities. Knowledge graph. Linked data. Knowledge inference. China Biographical Database (CBDB).

0 引言

随着互联网的快速发展,网络中的数据内容呈现出爆炸式增长的态势。与此同时,互联网内容的大规模、异质多元、组织结构松散等特点,给人们有效获取信息和知识提出了挑战。而知识图谱则以其强大的语义处理能力和开放

组织能力,为互联网时代的知识化组织和智能应用奠定了基础[1]。知识图谱不仅可以将互联网中的信息表达成更接近人类认知世界的形式,而且提供了一种更好的组织、管理和利用海量信息的方式。其发展得益于多个研究领域的成果,是知识库、自然语言处理、语义网技术、机器学习、数据挖掘等众多知识领域交叉融合的产物。作为人工智能时代最重要的知识表示方式之一,知识图谱能够打破不同场景下的数据隔离,为搜索、推荐、问答、解释与决策等应用提供基础支撑。但目前学界对知识图谱的理解比较混乱,主要存在“知识地图(Knowledge Map,KM)”“知识图谱(Knowledge Graph,KG)”和“图数据库(Graph Database,GD)”三种认知,时常混为一谈。

知识地图(KM)主要是指针对大量科学文献信息,借助于统计学、图论、计算机技术等手段,以可视化的方式来展示科学学科体系的内在结构(主题共现、合作团队、引用关系等)、学科特点、前沿热点、发展趋势等信息的一种计量学方法[2]。严格上讲,知识地图只是作为一种计量学方法,不能称为知识图谱。

谷歌于2012年提出一种在万维网上编码并关联碎片化知识单元的一种方案,该方案本质上是一种由知识点相互连接而成的语义网络,主要用于提升搜索引擎性能,通过描述现实世界中的实体及其关系,让用户能够更快更简单地发现新的信息和知识[3]。知识图谱(KG)要求以RDF三元组模型表达“实体—属性”和属性值(Statement),推荐以规范的词表模式(即Schema.org①)描述各类事物(人、地、事件等),以Microdata、RDFa、JSON-LD等方式进行三元组编码,使相关语义信息能够包含于网页之中并相互关联,并支持搜索引擎进行知识发现、索引以及可视化呈现。在谷歌发布知识图谱之前,Tim Berners-Lee早在2006年提出了“关联数据”概念,这是一种万维网上创建语义关联的方法。关联数据旨在通过URI和本体让机器读懂知识,用于推动数据公开,建立数据之间的链接以形成数据关系网(Web of Data)[4]。关联数据描述了通过可连接的URI发布来链接网络中各类资源的方法,可以看出,知识图谱其实就是在关联数据的基础上提出和发展的。由于知识图谱使用了RDF三元组模型,并支持机器语义描述,因此可看作是基于语义的知识图谱,在学界常被称为“关联数据(Linked Data)”,严格来讲,只有这种图谱才能被称为知识图谱。关联数据常使用RDF数据库(Triplestore)进行存储,本文讨论的知识图谱主要指语义知识图谱。

图数据库是以图形方式表示节点、属性和关系并进行存储和提供管理功能的数据库,如Neo4j、ArangoDB等,属于NoSQL的一种(其他还有键值对Key-Value、列存储数据库、文档型数据库三种),其作为大数据的一种重要支撑技术能够提供完善的图查询语言和丰富的图挖掘算法。图数据库的结构定义相比RDF数据库更为通用,可存储通用的三元组(S,P,O)数据,工业界目前谈论的知识图谱主要属于这一类。学术界和工业界在使用“知识图谱”表述时,往往不严格区分两种存储方案的区别,常常把两者混在一起,统称为知识图谱,因此采用图数据库构建的知识图谱可看成是广义的知识图谱。

知识图谱一经提出便迅速成为工业界和学术界的研究热点,涌现出大量的知识图谱应用和知识库。目前,微软和谷歌拥有全世界最大的通用知识图谱,Facebook拥有全世界最大的社交知识图谱,阿里巴巴和亚马逊则分别构建了庞大的商品知识图谱,百度致力于构建最大最全的中文知识图谱,美团NLP中心正在构建全世界最大的餐饮娱乐知识图谱“美团大脑”。此外,DBpedia、Freebase、Yago等大规模链接数据库(知识图谱)已成为众多知识库链接的首选目

① https://schema.org

标,国内也出现了 CN-DBpedia①、PKUBase②、zhishi.me③、Belief Engine④ 等多个百科全书式知识图谱(Encyclopedia Knowledge Graph),这其中的 CN-DBpedia 和 PKUBase 使用了三元组而非 RDF 标准,因此可看成是广义知识图谱。这些知识图谱可作为自动问答系统的知识来源,由此产生了诸如 Siri、IBM Waston、微软小冰、Google Allo、公子小白等多种成熟的自动问答系统和聊天机器人。在学术界,知识图谱也成为众多学者研究的热点。除了研究各种自动问答系统[5-6]外,知识图谱还被用于构建学术图谱研究中,如清华大学和微软研究院联合发布了全球最大学术图谱"开放学术图谱⑤(OAG)",该知识图谱目前包含 7 亿多条实体数据和 20 亿条关系。此外,清华大学还发布了知识计算开放平台(THUKC),该平台涵盖语言知识、常识知识、世界知识、认知知识等大规模知识图谱以及典型行业知识图谱。上海交通大学 Acemap 团队知识图谱小组则采用了 RDF 进行数据描述,发布了学术知识图谱 AceKG⑥,包含超过 1 亿个学术实体、22 亿条三元组信息[7]。此外,上海乐言科技王昊奋、东南大学漆桂林、浙江大学陈华钧等都是国内知识图谱应用的积极推动者。截至目前,中文知识图谱联盟⑦已有 61 家成员共 88 个数据集,其中包括语义知识图谱和广义知识图谱。

而在图情界和数字人文领域,研究较多的是语义知识图谱,即关联数据技术。欧洲数字图书馆 Europeana、Getty 数字博物馆、威尼斯时光机器项目[8]、芬兰数字人文关联开放数据基础设施(LODI4DH⑧)等都已成为数字人文领域应用关联数据技术的典范。2015 年 6 月 18 日,大英图书馆、新西兰国家图书馆、牛津大学图书馆、哈佛大学等 29 个非营利性图像资源存储机构共同成立国际图像互操作(IIIF⑨)组织,旨在确保全球图像存储的互操作性和可获取性,对以图像为载体的书籍、地图、卷轴、手稿、乐谱、档案资料等在线资源进行统一展示和使用。IIIF 中的一系列 API 都以 JSONLD 格式进行定义,关联数据和 IIIF 这两项开放共享标准已成为 GLAM(艺术馆、图书馆、档案馆和博物馆)的研究热点[9],并将开启数字人文研究的新时代。OCLC CONTENTdm 现已支持 IIIF,使用 CONTENTdm 的图书馆和博物馆可以通过一组常用 API 在不同平台间共享和呈现数字内容[10],威尔士报纸在线⑩、伏尔泰书信⑪、达芬奇手稿⑫等项目都采用这两项技术对其图像资源进行语义组织和发布。在国内,上海图书馆推出的家谱知识库⑬、古籍循证平台⑭、名人手稿知识库⑮等一系列数字人文项目也将关联数据技术和 IIIF 作为核心技术[11];北京大学严承希通过符号分

① http://kw.fudan.edu.cn/cndbpedia/intro
② http://pkubase.gstore-pku.com
③ http://zhishi.me
④ http://www.belief-engine.org/declarative
⑤ https://www.openacademic.ai/oag
⑥ https://www.acemap.info/app/AceKG
⑦ http://www.openkg.cn
⑧ https://seco.cs.aalto.fi/projects/lodi4dh
⑨ https://iiif.io
⑩ https://newspapers.library.wales
⑪ http://scalar.usc.edu/works/voltaire/index
⑫ https://www.vam.ac.uk/articles/explore-leonardo-da-vinci-codex-forster-i
⑬ http://jiapu.library.sh.cn
⑭ http://gj.library.sh.cn
⑮ http://sg.library.sh.cn

析法对 CBDB 数据集中宋代人物政治关系进行可视化分析[12];武汉大学曾子明将关联数据技术应用于敦煌视觉资源关联展示[13];侯西龙等将关联数据技术用于非物质文化遗产知识管理研究中[14]。这些研究大多都使用关联数据技术来进行元数据层面的知识组织和发布,极少使用知识图谱的理念对资源之间的关系进行揭示和知识推理。基于此,本文尝试在构建 CBDBLD①(CBDB 关联数据平台)的基础上,使用知识图谱的方法和推理机制对人物之间的关系进行展示。

1 语义知识图谱(关联数据)与广义知识图谱对比

关联数据和广义知识图谱都可用节点和边来表示实体和关系,因此混淆最多的也是这两种认知。简单来说,关联数据侧重于知识的发布与链接,其特点是将零散的数据进行关联组织,展示资源之间的关联关系,为进一步面向内容和知识的挖掘和计算奠定基础,广义知识图谱更侧重于知识挖掘和计算,其特点是知识存储、推理和计算,发现隐性知识并可视化,实现诸如提问式检索、时空展示等功能,推动人工智能环境下数字人文研究方法的创新。下文从两者的概念和技术层面进行对比,辨析两者异同。

1.1 概念层面的对比

表 1 显示了广义知识图谱和关联数据在概念层面的对比内容。

广义知识图谱用节点和关系所组成的图谱为真实世界的各个场景直观地建模,运用"图"这种基础性、通用性的"语言","高保真"地表达多姿多彩世界的各种关系。该类知识图谱一般以属性图为基本的表示形式,强调的是节点和边,节点上有属性(键值对),边也可以有属性。边有名字和方向,并总有一个开始节点和一个结束节点,节点可以有内部结构(三元组)。三元组数据通常存储在 Neo4j 图数据库中,常用 Cypher 查询语句。广义知识图谱主要用来解决存储和索引问题,并不能解决知识表示和全网络服务问题,因此仍然存在数据孤岛问题。广义知识图谱的主要优点是容易学习和实现,特别适用于社交网络,具有强计算性、运行效率高等特点;其缺点是不同知识图谱之间缺乏统一标准,难以互通,应用中语义模糊。凡是有关系的地方都可以使用到广义知识图谱,目前主要集中在社交网络、金融、保险、电子商务、物流等领域。

表 1 广义知识图谱与关联数据概念层面对比

对比项	广义知识图谱	关联数据
主要功能	存储、搜索	编码、关联
组成	Pattern	Schema
特点	大数据	智慧数据
应用	挖掘	推理
能力	计算	认知
作用域	数据孤岛(Data Silo)	万维网(WWW)
软件举例	Neo4j	OpenLink Virtuoso
应用举例	脸书 Social Graph	谷歌 KG

关联数据表示的语义知识图谱同样存在节点和边的概念,节点用来表示类(实体),边用来表示属性。连接不同类的属性称为对象属性(Object Property),连接类对应的属性值的属性称为数据属性(Datatype Property)。实体必须以 URI 命名,本体用来表示概念之间的关系。不同的图谱之间具有标准的 SPARQL 查询语言,可解决跨域查询。关联数据的主要优点是基于描述逻辑的数学基础,有着标准化的规范词表,图谱之间易于交互;其缺点主要是高复杂性、学习门槛高。目前,关联数据技术已大量应用于数字人文研究、生物医学知识库构建、政府数据开放、规范数据发布等领域。

① http://cbdb.library.sh.cn

1.2 技术层面的对比

构建知识图谱的目的是获取大量的、让计算机可读的知识,在互联网飞速发展的今天,知识大量存在于非结构化的文本数据、大量半结构化的表格和网页以及生产系统的结构化数据中。表 2 中列出了广义知识图谱中常用到的技术:知识建模、知识获取、知识融合、知识存储、知识计算、图挖掘和图计算以及可视化技术[15]。

表 2 广义知识图谱与关联数据技术层面对比

广义知识图谱常用技术	主要作用	关联数据常用技术
知识建模	为知识和数据进行抽象建模	本体构建
知识获取	从不同来源、不同结构的数据中进行知识抽取,将知识存入到知识图谱	RDF 结构化
知识融合	将不同来源、不同结构的数据中抽取的知识融合成一个统一的知识图谱	关联数据
知识存储	用于数据存储,同时支持上层的知识推理、快速检索、图实时计算等应用	RDF 存储
知识计算	通过各种算法,发现其中显式或隐含的知识、模式或规则	知识推理
图挖掘和图计算	图遍历、路径计算与探寻、权威节点分析、族群分析、相似点发现等基于图的分析与计算	图遍历计算
可视化技术	结合可视化工具进行数据分析	可视化技术

下面从关联数据的角度进行解析。

(1)知识建模:设计本体进行知识和数据的组织。

(2)知识获取:根据设计的本体,将不同来源、不同格式的数据转换为 RDF 结构。

(3)知识融合:借助自然语言处理、实体识别、机器学习等算法建立起不同来源实体之间的关联关系。

(4)知识存储:RDF 数据常存储于三元组数据库(Triple Store)中,三元组数据库也可以看成是 NoSQL 数据库的一种。

(5)知识计算:应用逻辑描述推理(TBox 和 ABox)进行隐式数据的推理与发现。

(6)图挖掘和图计算:三元组数据库作为 NoSQL 数据库的一种,可以进行一些基于图的计算,如图遍历、路径计算等。

(7)可视化技术:作为展示层的应用,可以结合 D3.js、ECharts 等可视化插件。当知识图谱以图的形式展现后,信息一目了然,符合人脑对现实世界的认知模型。

2 知识图谱应用于数字人文系统框架

2.1 数字人文研究进程

数字人文研究框架如图 1 所示,主要从研究过程、研究行为和研究方法三个角度来理解数字人文研究。数字人文的研究过程和研究行为两部分和传统的人文研究相似,从研究过程看,人文研究主要包括占有材料、发现事实、分析比较、归纳整理、得出结论和发表交流等步骤;从研究行为看,人文研究主要用于发现、注释、比较、参考、抽样、说明和表示。数字人文的研究方法经历了三个阶段的方法变革。

第一阶段为资源数字化转换阶段,并将结构化描述的信息存入到关系型数据库中。在国内,各个图书馆、档案馆、博物馆馆藏中大量的扫描件就处于数字化建设和研究阶段。这些大规模数字资源中蕴含的丰富知识却长期处于封闭状态,只能通过一些结构化的描述信息来了解一二,严重影响了知识的传播与深层应用。该阶段的数字人文研究需要借助大量的人力参

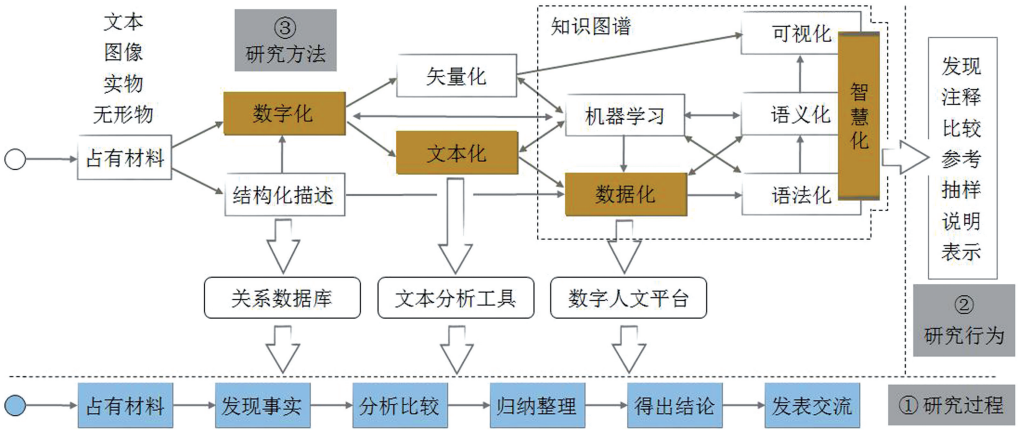

图 1　数字人文研究框架

与,尤其是需要人文学者从事大量低水平且繁杂的资料搜集和整理工作。

第二阶段主要集中在数字资源的文本化建设和研究。目前国外数字人文的研究主要处于第二个阶段,研究内容是对数字化的内容进行文本化识别,并利用文本分析工具对资源内容进行文本分析和内容分析以支持数字人文研究,如对文本内容进行分词、词性标注、文本摘要提取、情感分析等。该阶段的文本化建设可以将人文学者从大量繁杂的工作中解脱出来,也为将来的语义建设提供了基础[16]。

第三阶段为近年来数字人文领域研究的热点,主要是结合本体、(语义)知识图谱、机器学习等语义技术对文本化的资源进行数据化和智慧化建设。数据化需要使用自然语言处理和机器学习方法对文本化资源中的实体进行概念提取,并利用语义化技术对数据化内容进行语义增强;智慧化包括语法化、语义化、可视化等建设过程。在该阶段研究中,知识图谱的引入将给传统的人文研究带来新的方向,赋予新的智慧。数据化的资源需要经过语法化统一、语义化增强,以实现可视化应用,这三者可以作为表 2 中知识图谱研究的七类核心技术的浓缩。语法化统一过程首先需要进行本体的知识建模,并结合本体进行资源的知识获取和存储;知识融合、知识计算、图挖掘和图计算等技术包含在资源的语义化建设之中,并可结合开放的关联数据集进行资源内容的语义化增强;可视化可作为知识图谱最直接的体现,提供直观的人文数据分析和研究。这里也可以看出,知识图谱是一个动态开放的建设过程,包含了多个步骤和流程,而并非狭义上的认为只有最后的可视化才是知识图谱,知识图谱中的“图”更多的是体现在知识组织方面,而不单单是可视化中图表的“图”。

2.2　CBDBLD(CBDB 关联数据平台)构建流程

关联数据要求数据以 RDF 形式存在,因此在进行知识图谱研究和知识推理前,需要采用关联数据技术将非 RDF 的数据转变成 RDF 格式。中国历代人物传记资料库(CBDB)平台拥有 42 万多人的人物数据,这些数据可通过 API 接口获取 JSON 格式。图 2 中结合关联数据生命周期(Linked Data Lifecycle)简单展示了 CBDBLD 平台的建设流程。关于构建关联数据应用的研究有很多,包含关联数据发布规范[17]及发布方法研究[18-19]、本体设计[20]等,相关技术也相对成熟。

(1)数据准备:使用 Python 脚本通过 CBDB 提供的 API 方式获取 CBDB 中的 42 万多人的 JSON 格式数据。

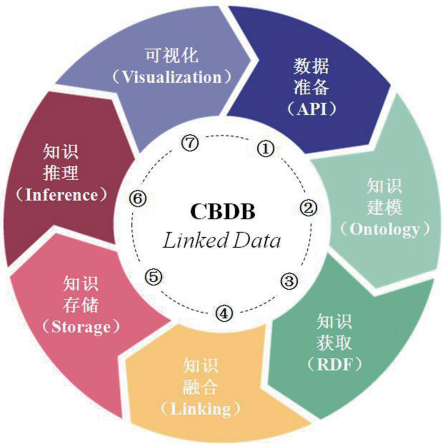

图 2 CBDBLD 平台构建流程

(2)知识建模:构建 CBDB 本体模型,由于 CBDB 数据涉及内容较多,目前仅设计了 CBDB 部分数据的本体结构,本体可在“上海图书馆本体服务中心”平台上获取。

(3)知识获取:通过程序将获取的人物 JSON 数据按照设计的本体结构进行 RDF 转换,RDF 数据格式可看成有格式的文本文件,因此可以通过写一点脚本或者简单的程序来实现。

(4)知识融合:尽管这一步为可选步骤,但关联数据的核心是发布和链接,且“开放数据五星模型”中也将数据之间的关联定义为第五星,因此将转换的 RDF 数据与上海图书馆人名规范库、VIAF、DBPedia 等数据集进行关联。关联时,可采用 SILK① 或者 LIMES② 框架进行关联。

(5)知识存储:上述生成的数据需要进行持久化存储,从关联数据和知识图谱的比较中可以看出,RDF 数据可以存放到图数据库或者三元组数据库中,但是考虑到数据的开放与发布,这里使用三元组数据库(OpenLink Virtuoso③)进行存储。

(6)知识推理:CBDB 中的数据大多数由人工整理产生,人物之间的关系(尤其是非直接关系)很多时候没有被添加完整,因此可通过制定推理规则进行关系的发现与挖掘。这也是本文重要的技术尝试。

(7)可视化:CBDB 中存在大量人物之间的亲属关系和社会关系,通过交互式的知识图谱设计呈现多维度的人际关系,如亲属圈、朋友圈、学术圈、政治圈等,有助于人文研究。

3 知识图谱在 CBDBLD 中的应用

3.1 人物关系模型设计

CBDB 中有近 500 种人物之间的关系,主要包括亲属关系和社会关系,这些人与人之间的关系可以用知识图谱形式来展示。这里以学术类关系中师生关系为例,阐述构建知识图谱的流程和方法。例如,图 3 中的 CBDB 社会关系分为“社会关系(笼统)”“政治关系类”“军事关系类”“朋友关系类”“著述关系类”“学术关系类”“医疗关系”“宗教关系类”“家庭关系”“财务关系类”十大类,其中对学术类关系进行展开,分为“学术交往”“师生关系”“学术主题相近”“同

① http://silkframework.org
② http://aksw.org/Projects/LIMES.html
③ https://virtuoso.openlinksw.com

为……之成员”“学术襄助”“文学艺术交往”和“学术攻讦”七大类,其中的“师生关系”又包含“为 Y 之门人”“为 Y 之学生”“为 Y 之弟子”“从 Y 游”和“为 Y 之考官”五类关系。这五类关系又有各自的反向对称关系,如“门人为 Y”和“为 Y 之门人”形成反向对称关系。对这些关系使用 RDF 进行组织,其中主要用到了 bf:relatedTo、shl:pairOf 和 dc:partOf 这三个对象属性。

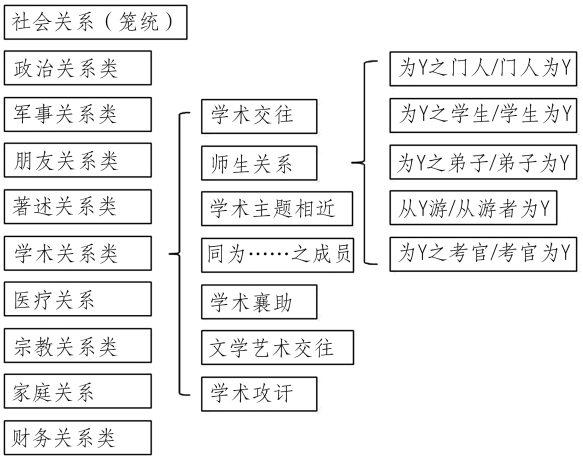

图 3 CBDB 社会关系(以学术关系类中师生关系展开)

1)bf:relatedTo(关系到……)用来表示具体的社会关系(asso_code)到所属类型(asso_type)的连接;

2)shl:pairOf(与……反向)用来连接两个互为反向的社会关系;

3)dc:partOf(……的部分)用来展示社会关系类型到其十大类社会关系之间的连接。

用知识图谱展示如图 4 所示,对应的 RDF 数据为:

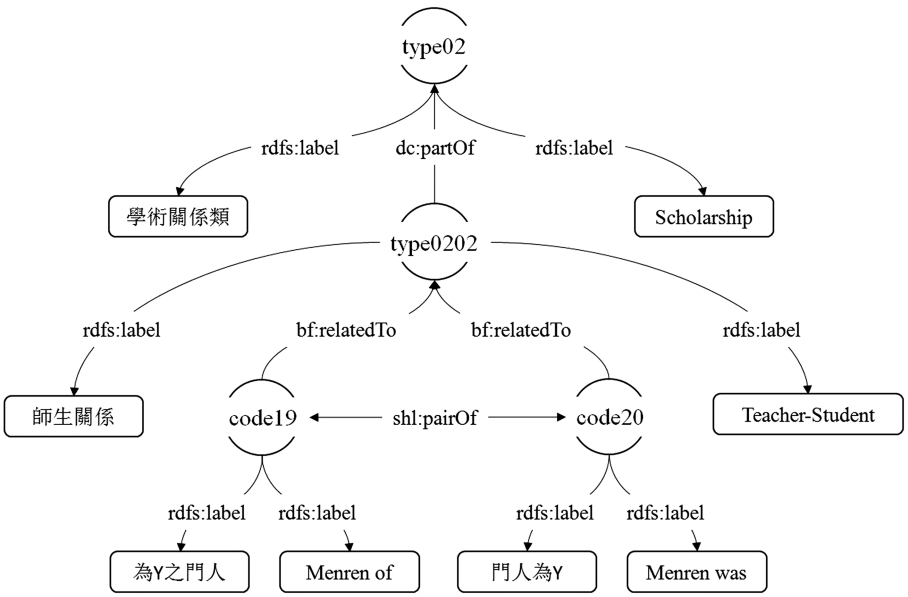

图 4 学术关系类中师生关系知识图谱

```
<http://cbdb.library.sh.cn/names/asso_code/20>
    rdfs:label '門人為 Y' ;
    rdfs:label 'Menren was'@ en;
    bf:relatedTo <http://cbdb.library.sh.cn/names/asso_type/0202> ;
    shl:pairOf <http://cbdb.library.sh.cn/names/asso_code/19> .
<http://cbdb.library.sh.cn/names/asso_code/19>
    rdfs:label '為 Y 之門人' ;
    rdfs:label 'Menren of'@ en;
    bf:relatedTo <http://cbdb.library.sh.cn/names/asso_type/0202> ;
    shl:pairOf <http://cbdb.library.sh.cn/names/asso_code/20> .
<http://cbdb.library.sh.cn/names/asso_type/0202>
    rdfs:label '師生關係' ;
    rdfs:label 'Teacher-Student'@ en ;
    dc:partOf <http://cbdb.library.sh.cn/names/asso_type/02> .
<http://cbdb.library.sh.cn/names/asso_type/02>
    rdfs:label '學術關係類' ;
    rdfs:label 'Scholarship'@ en .
```

3.2 社会关系知识图谱实现

以李清照的社会关系为例来说明知识图谱的构建过程。李清照在 CBDB 中存在学术类和著述类两大社会关系类,而王安石则存在学术类、政治类、朋友类、著述类和社会关系(笼统)五大类。在同一个图谱中显示某人的所有关系节点,将会非常混乱,因此在设计知识图谱时,可以根据其社会关系大类进行查看。如在李清照的学术类关系图谱(见图 5)中,这里仅显示她及其相关人物之间的学术类关系,并不夹杂其他的社会关系,因此可以清晰地看到她的单纯的学术图谱:图 5 中可以看出李清照的学生为韩玉父;结合李清照的亲属关系图谱,可以看出赵明诚是李清照的第一任丈夫,由赵明诚著、李清照整理的《金石录》可以看出两人之间的学术关系;此外,赵明诚仿欧阳修《集古录》作《金石录》30 卷,故从知识图谱中可以直接地看出其文风效法欧阳修。

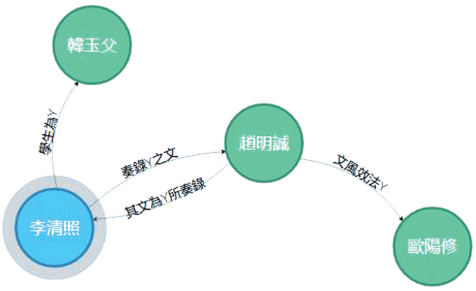

图 5 李清照学术类关系图谱

知识图谱除了可以直观地呈现人物之间的关系,还可以通过交互性增强体验感,图谱中所有的节点都可以通过单击进行下一层级关系的扩展显示,如图 6 所示,通过点击节点“欧阳修”则会展开欧阳修的学术类关系。这里可以看到欧阳修为晏殊的门人,同样的方法可以点击“晏殊”节点以查看晏殊的学术关系图谱。需要注意的关系为有向线段,当展开“欧阳修”节点时,会产生一条反向关系(文风为 Y 所效法)到“赵明诚”节点。

如何通过点击节点实现特定关系类型的提取是知识图谱呈现的关键,CBDB 中 42 万多人的数据以 RDF 格式存储在数据库中,下面通过 SPARQL 来分析人物特定关系类型提取的机制和原理。图 7 给出了学术关系类的知识图谱查询模式,当检索条件为欧阳修时,通过 SPARQL 将返回欧阳修学术关系类的关系人 URI(?rel_

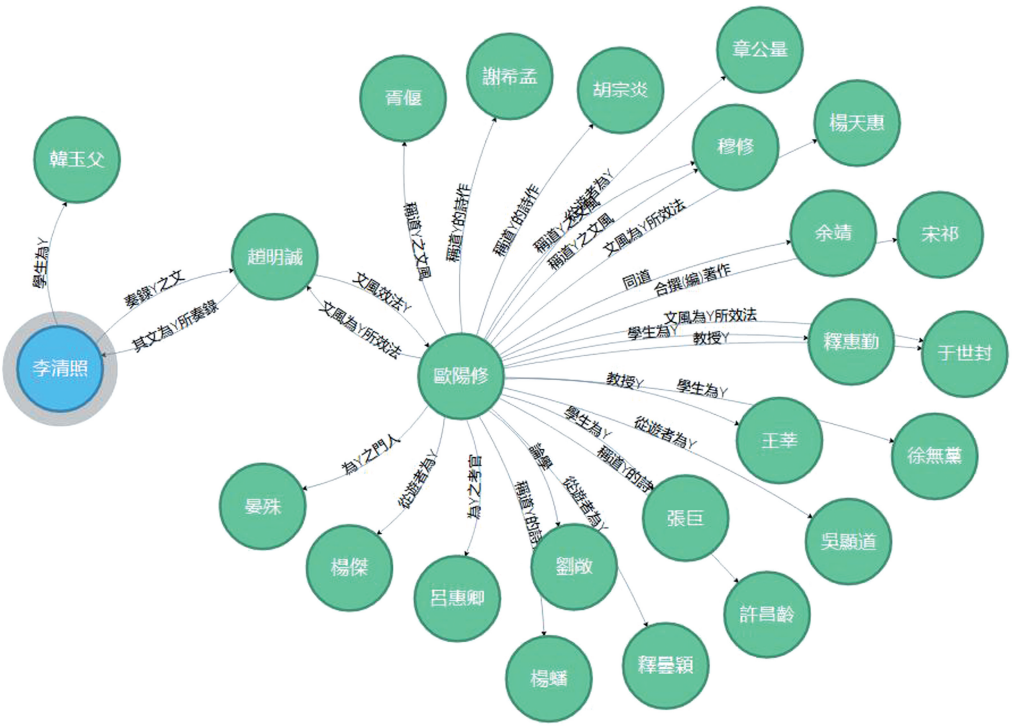

图 6　李清照著述类关系图谱(展开“欧阳修”节点)

```
1 SELECT DISTINCT ?rel ?rel_person_uri ?person ?rel_year
2 WHERE {
3 ?rel_uri shl:relationSubject <%1 $ s>;
4 shl:relationObject ?rel_person_uri;
5 shl:specialRelationType '社会关系';
6 shl:relationLabel ?rel;
7 shl:relationType '學術關係類' .
8 OPTIONAL {?rel_uri shl:temporal/rdfs:label ?rel_year .}
9 ?rel_person_uri foaf:name ?person .
10 FILTER (lang(?person) = 'cht')
11 } ORDER BY ASC(?rel_year)
```

person_uri)和姓名(?person)。①知识图谱中每个节点都有其唯一的人物 URI,查询时在 SPARQL 中传入变量%1 $ s。②用 SPARQL 模式查询该人物对应的关系人 URI(?rel_person_uri),这里的关系用 shl:relationObject 表示。③为 SPARQL 查询添加限制条件,属性 shl:specialRelationType 中有“亲属关系”和“社会关系”两类取值,这里以“社会关系”为例;属性 shl:relationType 来指明具体的社会关系,取值来源于前面划分的十大类社会关系,这里为“学术关系类”。④查询关系人与变量%1 $ s 之间具体的社会关系。⑤结合其他类,返回所需信息。这里返回了关系人姓名(foaf:name 对应的属性值)和关系年份(shl:temporal 对应的节点类的 rdfs:label 属性值)。

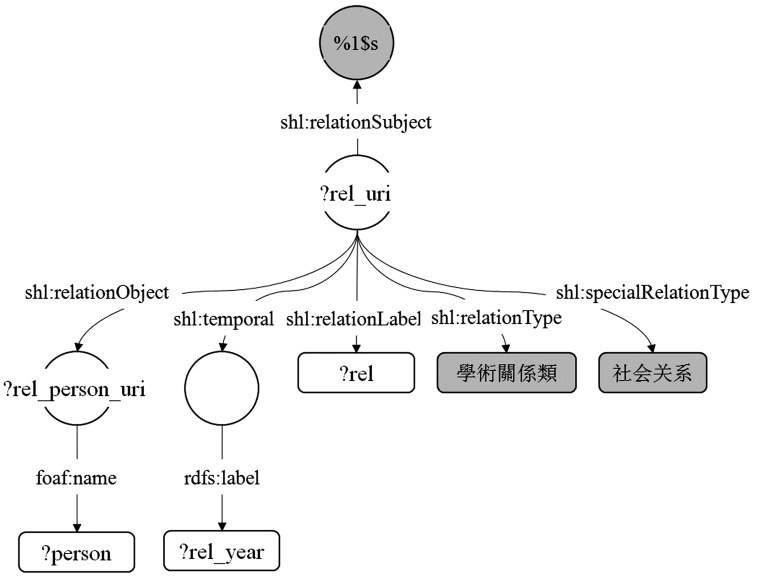

图 7 学术关系类图谱查询模式

3.3 知识推理

简单而言,知识推理就是指根据知识图谱中已有的知识,推断出新的、未知的知识。作为知识图谱应用中的核心功能之一,知识推理在数字人文项目中还很少有类似的成熟系统出现。CBDBLD 平台基于 Apache Jena 框架构建,可以非常方便地使用 Jena 中的推理引擎进行推理。在规则引擎中,通常将知识表达为规则(rules),把要分析的情况定义为事实(facts)。和人类的思维相对应,规则引擎中存在"正向推理(Forward-Chaining)"和"反向推理(Backward-Chaining)"两种推理方式。正向推理也叫演绎法,由事实驱动,从一个初始的事实出发,不断地应用规则得出新的结论。反向推理也叫归纳法,由目标驱动,首先提出某个假设,然后寻找支持该假设的证据,如能找到,说明原假设正确;反之,说明原假设不成立,此时需要建立新的假设[21]。现以正向推理为例,介绍如何在数字人文中使用推理规则和知识推理引擎。

Apache Jena 框架中包含许多预定义的推理器:传递推理器、RDFS 规则推理器、OWL 推理器、通用规则推理器[22-23]。在本研究中,我们使用通用规则(Rule-based)推理器[24-25]支持用户自定义的基于规则的推理,对应的部分规则如下所示。

```
@prefixshl: <http://www.library.sh.cn/ontology/>.
[r1: (?u shl:relationSubject ?a), (?u shl:relationObject ?b), (?u shl:relationLabel '父'),
    (?u1 shl:relationSubject ?a), (?u1 shl:relationObject ?c), (?u1 shl:relationLabel ?zf)
    regex(?zf, '^.+丈夫$') -> (?b shl:sonInLaw ?c), (shl:sonInLaw rdfs:label '女婿')]
[r2: (?b shl:sonInLaw ?c) -> (?c shl:fatherInLaw ?b), (shl:fatherInLaw rdfs:label '岳父')]
[r3: (?u1 shl:relationSubject ?a), (?u1 shl:relationObject ?c), (?u1 shl:relationLabel ?zf)
    regex(?zf, '^.+丈夫$') -> (?c shl:wife ?a), (shl:wife rdfs:label '妻子')]
```

这里包含了三条规则:①规则 1:如果 a 的父为 b,丈夫为 c,则推出 b 的女婿为 c;②规则 2:在规则 1 的基础上增加新的 RDF 断言,b 的女婿为 c,则 c 的岳父为 b;③规则 3:描述为 a 的丈夫为 c,则 a 为 c 的妻子。图 8 可用来解析这些推理规则,新的断言为图中虚线标记部分。

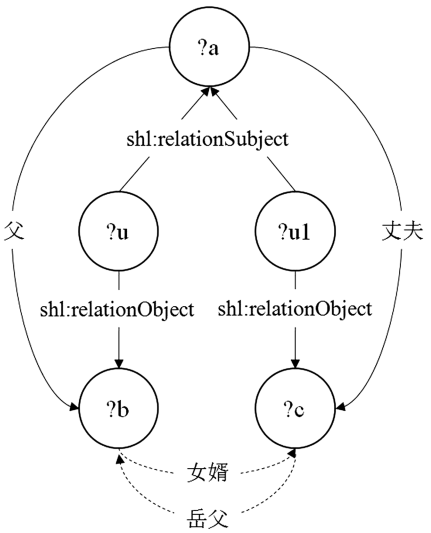

图 8　人物关系推理规则流程

结合规则文件,可以得出如下关系图谱:图 9 为未加载推理规则时的亲属关系图谱;图 10 为加载推理规则后的亲属关系图谱。在图 10 中用虚线标记出使用推理后得出的 RDF 断言,在原始的 CBDB 数据中并没有标记出张汝舟和李格非的关系,通过上述规则,可以得出张汝舟的岳父是李格非,李格非的女婿为张汝舟;同时 CBDB 中存在李清照到张汝舟的直接关系(第二任丈夫),却并没有张汝舟到李清照的关系,当检索张汝舟的妻子时,并不能直接得出结果。当使用推理规则后,将会生成张汝舟到李清照的关系断言,即张汝舟的妻子为李清照。

在使用规则进行推理时,要注意推理的效率。对于大数据集的推理可以考虑以下两种推理模式。①静态推理模式。在已知推理规则的前提下,可预先进行推理,并将推理的 RDF 断言存储到 Graph 中,在后续使用时直接调用 Graph 中的断言即可。②动态推理模式。可将大的数据集图谱分成多个子图,在子图中运行推理规则进行实时动态推理。两种推理模式各有优点,前者关注推理的精度,力求推理的全面和精

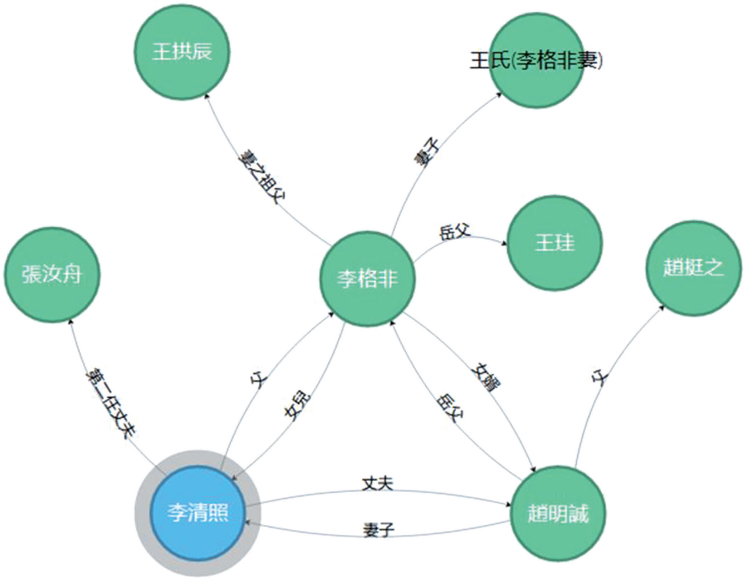

图 9　未使用推理规则时亲属关系图谱

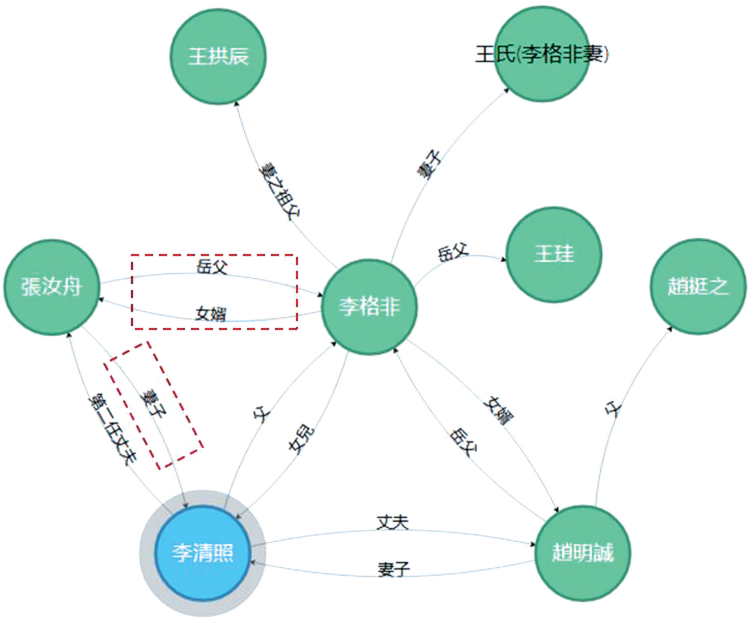

图 10 使用推理规则后亲属关系图谱

准;而后者则强调推理的速度,追求推理的效率和体验。本研究中采用了第二种推理模式,这里的子图选择并非根据某些类或者属性进行划分,而是视用户在关系图谱中的行为而定。当用户在关系图谱中展开 5 个人物节点时,则子图由这 5 个节点组成;当展开 10 个人物节点时,则子图由这 10 个节点组成。采用这种方法可在一定程度上限制推理范围,提高推理效率。

4 总结与展望

知识图谱按照存储方式分为语义知识图谱(关联数据)和广义知识图谱,本文从概念层面和技术层面对两类知识图谱进行了深入比较,广义知识图谱强调知识的挖掘与计算,关联数据则更侧重于知识的发布与链接。通过对比可以看出,关联数据作为原生态的知识图谱,对知识的组织有更高的要求,需要按照 W3C 的四原则进行数据的组织与发布。此外,本文提出的知识图谱在数字人文领域的研究框架已用于多个数字人文研究项目中,并成为众多关联数据应用平台的首选解决方案。论文以 CBDBLD 平台为例,通过知识图谱技术展示了人物之间丰富的社会关系,并借助推理规则对人物之间的隐性关系进行挖掘和揭示。尽管本文探讨的知识图谱方法和推理比较初步,但是将知识图谱的理念和算法与关联数据共同应用于数字人文将会是数字人文领域研究的重点,“共享、链接、智慧”将会带来数字人文研究的新时代。

当然,本研究仍有不少欠缺的地方,将来可从知识图谱研究的广度、深度和高度三方面开展进一步的研究和探索。首先,研究广度体现在规则数量上,将预置更多的公理和通用推理规则供用户使用。其次,研究深度体现在图谱算法上,逐渐实现更多基于图的计算,如图路径、图遍历、族群分析、权威点和相似点计算等。最后,研究高度体现在分析纬度中,本研究目前仅对人物之间的关系建立图谱,将来会从其他纬度考虑建立更多面的知识图谱,如人物官职图谱、著述图谱等。

参考文献

[1] 徐增林, 盛泳潘, 贺丽荣,等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(4):589-606. (Xu Zenglin, Sheng Yongpan, He Lirong, et al. Review on knowledge graph techniques[J]. Journal of University of Electronic Science and Technology of China, 2016, 45(4):589-606.)

[2] 秦长江, 侯汉清. 知识图谱——信息管理与知识管理的新领域[J]. 大学图书馆学报, 2009, 45(1):30-37,96.(Qin Changjiang, Hou Hanqing. Mapping knowledge domain: a new field of information management and knowledge management[J]. Journal of Academic Libraries, 2009, 45(1):30-37,96.)

[3] 刘峤, 李杨, 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. (Liu Qiao, Li Yang, Yang Duanhong, et al. Knowledge graph construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3):582-600.)

[4] Heath T, Bizer C. Linked data: evolving the web into a global data space (1st edition)[M]. Morgan & Claypool, 2011: 1-136.

[5] Moschitti A, Tymoshenko K, Alexopoulos P, et al. Exploiting linked data and knowledge graphs in large organizations[M]. Springer, Cham, 2017:181-212.

[6] Zheng Weiguo, Yu J X, Zou Lei, et al. Question answering over knowledge graphs: question understanding via template decomposition[C]//Proceedings of the VLDB Endowment, 2018, 11:1373-1386.

[7] Wang Ruijie, Yan Yuchen, Wang Jialu, et al. AceKG: a large-scale knowledge graph for academic data mining [C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM),2018: 1487-1490.

[8] Abbott A. The 'Time Machine' reconstructing ancient Venice's social networks[J]. Nature, 2017, 546(7658): 341-344.

[9] 曾蕾, 王晓光, 范炜. 图档博领域的智慧数据及其在数字人文研究中的角色[J]. 中国图书馆学报, 2018, 44(1):17-34. (Zeng Lei, Wang Xiaoguang, Fan Wei. Smart data from libraries, archives and museums and its role in the digital humanity researches[J]. Journal of Library Science in China, 2018, 44(1):17-34.)

[10] Huddleston S, Fennell C. Leveraging the benefits of IIIF in CONTENTdm[EB/OL]. [2019-04-07]. https://www.oclc.org/content/dam/support/contentdm/leveraging-benefits-IIIF.pdf.

[11] 夏翠娟, 林海青, 刘炜. 面向循证实践的中文古籍数据模型研究与设计[J]. 中国图书馆学报, 2017, 43(6):16-34. (Xia Cuijuan, Lin Haiqing, Liu Wei. Designing a data model of Chinese ancient books for evidence-based practice[J]. Journal of Library Science in China, 2017, 43(6):16-34.)

[12] 严承希, 王军. 数字人文视角:基于符号分析法的宋代政治网络可视化研究[J].中国图书馆学报,2018, 44(5):87-103. (Yan Chengxi, Wang Jun. Digital humanistic perspective: a study on the visualization of political network in Song Dynasty based on symbolic analysis[J]. Journal of Library Science of China, 2018, 44(5):87-103.)

[13] 曾子明, 周知, 蒋琳. 基于关联数据的数字人文视觉资源组织研究[J]. 情报资料工作, 2018(6):6-12. (Zeng Ziming, Zhou Zhi, Jiang Lin. Research on knowledge organization of linked data-based digital humanities visual resources[J]. Information and Documentation Services, 2018(6):6-12.)

[14] 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.(Hou Xilong, Tan Guoxin, Zhuang Wenjie, et al. Research on knowledge management of

intangible cultural heritage based on linked data[J]. Journal of Library Science of China, 2019, 45(2): 88-108.)

[15] 曹倩, 赵一鸣. 知识图谱的技术实现流程及相关应用[J]. 情报理论与实践, 2015, 12(38):127-132. (Cao Qian, Zhao Yiming. The realization process and related applications of knowledge graph[J]. Information Studies Theory & Application, 2015, 12(38):127-132.)

[16] 刘炜, 叶鹰. 数字人文的技术体系与理论结构探讨[J]. 中国图书馆学报, 2017, 43(5):32-41.(Liu K W, Ye F Y. Exploring technical system and theoretical structure of digital humanities[J]. Journal of Library Science in China, 2017, 43(5):32-41.)

[17] 陈涛, 张永娟, 刘炜,等. 关联数据发布的若干规范及建议[J]. 中国图书馆学报, 2019, 45(1):34-46. (Chen Tao, Zhang Yongjuan, Liu Wei, et al. Several specifications and recommendations for the publication of linked data[J]. Journal of Library Science in China, 2019, 45(1):34-46.)

[18] Best practices for publishing linked data[EB/OL]. [2019-04-03]. https://www.w3.org/TR/ld-bp.

[19] Büch H. Publishing linked data: different approaches and tools[EB/OL]. [2019-04-05]. https://hdms.bsz-bw.de/frontdoor/deliver/index/docId/1721/file/BuechHolger_PublishingLinkedOpenData_Kopie.pdf.

[20] Janowicz K, Hitzler P, Adams B, et al. Five stars of linked data vocabulary use[J]. Semantic Web, 2014, 5(3):173-176.

[21] Al-Ajlan A. The comparison between forward and backward chaining[J]. International Journal of Machine Learning and Computing, 2015, 5(2):106-113.

[22] Reasoners and rule engines: Jena inference support[EB/OL]. [2019-02-23]. http://jena.apache.org/documentation/inference/.

[23] Ali A. Semantic web based inference capabilities using Jena framework[J]. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2017, 6(2):135-138.

[24] Rattanasawad T, Buranarach M, Saikaew K R. A comparative study of rule-based inference engines for the semantic web[J]. IEICE Transactions, 2018, E101.D(1):82-89.

[25] Käfer T, Harth A. Rule-based programming of user agents for linked data[C]//The Web Conf Workshop: Linked Data on the Web (LDOW), Lyon, France, 2018.

陈 涛 上海图书馆/上海科学技术情报研究所,博士。上海 200031。
刘 炜 上海图书馆/上海科学技术情报研究所研究员。上海 200031。
单蓉蓉 上海大学图书情报档案系,博士研究生。上海 200444。
朱庆华 南京大学信息管理学院教授,博士生导师。江苏 南京 210023。

(收稿日期:2019-04-07)