第 29 卷第 5 期 安 徽 建 筑 大 学 学 报 Vol.29 No.5
2021 年 10 月 Journal of Anhui Jianzhu University Oct.2021

DOI：10.11921/j.issn.2095-8382.20210503

徽派建筑知识图谱的半自动化构建

张润梅[1]，杨超[2]，尹蕾[2]，张媛[2]

（1. 安徽建筑大学　机械与电气工程学院，安徽　合肥　30601；
2. 安徽建筑大学　电子与信息工程学院，安徽　合肥　230601）

摘　要：本文针对徽派建筑数据异构多源和非结构化的特点，提出一种 BiLSTM-CRF 模型与徽派建筑词典相结合的命名实体识别方法，利用先验知识的辅助作用，提升实体识别效果，完成对建筑实体进行的识别抽取。利用 Neo4j 图数据库存储知识，用属性图模型表示知识。最后使用 Neo4j 图数据库对构建的徽派建筑知识图谱进行了可视化展示。研究结果表明，此方法能够有效地构建徽派建筑领域知识图谱，为今后徽派建筑知识智能化推荐和搜索系统研究奠定基础。

关键词：徽派建筑；知识图谱；图数据库；BiLSTM-CRF 模型；Neo4j

中图分类号：TP391　　**文献标识码：**A　　**文章编号：**2095-8382（2021）05-013-07

Semi-automated Build of Huizhou Architectural Knowledge Graph

ZHANG Runmei[1]，YANG Chao[2]，YIN Lei[2]，ZHANG Yuan[2]

（1.School of Mechanical and Electrical Engineering，Anhui Jianzhu University，Hefei 230601，China；
2.School of Electronics and Information Engineering，Anhui Jianzhu University，Hefei 230601，China）

Abstract：Considering heterogeneous multiple sources and unstructured characteristics of Huizhou architecture data，a named entity recognition method combining BiLSTM-CRF model and Huizhou architecture lexicon is proposed in this paper to improve the entity recognition with prior knowledge. The Neo4j graph database is used to store knowledge and the attribute graph model to represent，hence the authors visualize Huizhou architectural knowledge graph. The results show that this method can build a knowledge graph in the field of Huizhou architecture effectively，which lays a foundation for further research on intelligent recommendation and search systems of Huizhou architecture knowledge.

Keywords：Huizhou architecture；knowledge graph；graph database；BiLSTM-CRF model；Neo4j

将传统建筑的特征元素融入到现代建筑设计中是实现传统建筑传承的必要手段，也是弘扬传统文化的有效途径。传统建筑及其构件本身具备独特的美感，且类型丰富，数量巨大[1]，通过传统的手段获取所需的传统建筑数据信息是一件费时费力的工作。2012 年，谷歌正式提出知识图谱的概念，旨在

收稿日期：2020-12-21

基金项目：安徽省自然科学基金（2008085MF218）；安徽省高校学科拔尖人才学术资助项目（gxbjZD26）；安徽省高校省级人文社会科学研究项目（SK2018A0578）；安徽省自然科学基金青年项目（1508085QF137）；安徽省教育厅自然科学研究项目（KJ2019JD12）；安徽建筑大学引进人才及博士启动基金项目（2019QDZ38）；安徽建筑大学校级科研项目（JZ192066）

作者简介：张润梅（1971-），女，教授，博士，研究方向：机器人技术、数据挖掘、虚拟现实及建筑数字化。

实现更加智能化的搜索引擎[2]。随着人工智能和大数据技术的不断发展和应用，知识图谱已在多个领域得到了广泛应用，如智能搜索[3]、智能问答[4]、个性化推荐[5]等。目前基于知识的智能问答和推荐系统有很多，如苹果手机智能语音助手 Siri、科大讯飞的讯飞开放平台等，但基于传统建筑知识库构建智能化推荐和搜索系统的研究尚不多见。因此，构建传统建筑知识图谱是实现大规模知识管理和应用的基础，具有重要的研究意义与应用价值。

近年来，特定领域知识图谱构建的研究受到研究者的广泛关注。祁志武[6]将知识图谱与地质标本相结合，通过七步法构建了地质标本知识本体，实现了地质标本知识图谱的构建。王良萸[7]针对碳交易领域的半结构化和非结构化数据，分别采用自定义的 Web 数据包装器，结合 BiLSTM-CRF 模型与依存句法分析实现了三元组抽取，构建了碳交易领域知识图谱。汤洁[8]提出了一种基于启发式规则的网页正文内容抽取算法，并提出基于最短路径算法和深度优先搜索算法来分析金融市场中各实体之间的关系。

目前，很多专业领域已完成了知识图谱构建，且基于知识图谱的各类应用开发也得到迅速发展。国内外很多大公司通过知识图谱来提高服务质量，如金融知识图谱[9]、医学知识图谱[10]、化学知识图谱[11]等。在建筑领域更多针对聚落基因图谱开展相关研究，如秦为径等人[12]对凉山彝族地区的乡土景观基因要素进行提取、分类和编码，完成了凉山彝族地区乡土景观基因图谱信息链的构建。聂聆[13]通过对徽州古村落景观特征进行研究识别，构建了徽州古村落景观基因图谱。翟洲燕等人[14]通过对陕西省 35 个传统村落的分析，识别并提取了传统村落文化遗产景观基因，绘制了陕西传统村落文化遗产景观基因组图谱。但以上均未形成完整的、专业的知识图谱。

徽派建筑形成于宋，成长于元，至明清达到鼎盛，是中国传统建筑的重要组成部分。徽派建筑种类繁多，建筑形式多样，时间跨度大。要实现数据有效整合，自动构建徽派建筑知识图谱存在诸多困难。本文从分析徽派建筑现存资料入手定义了徽派建筑知识图谱的概念层，通过对异构数据过滤、清洗、解析、进行实体、属性以及关系的抽取，并通过构建徽派建筑领域词典，结合先验知识提升了 BiLSTM-CRF 模型的实体识别效果，通过 Neo4j[15]图数据库实现知识的表示、存储并用 Cypher[16]实现知识查询。

1 知识图谱构建相关技术

1.1 命名实体识别

命名实体识别作为自然语言处理的一项基础技术，其主要任务是识别出文本数据中的专有名词和有特殊含义的词并将其归类到已定义的类型中。命名实体识别有基于规则的方法、基于大规模语料库的统计方法和基于机器学习的方法三种基本方法，本文采用的是基于机器学习的方法。

1.2 条件随机场模型

条件随机场模型（Conditional Random Field，CRF）由 Lafferty 等人[17]于 2001 年提出，结合了最大熵模型（MaxEnt）以及隐马尔科夫模型（HMM）的特点，是一种判别式概率无向图学习模型，本文采用线性条件随机场模型。若让 m 表示被观察的输入数据序列，n 表示得到的输出序列，$p(n|m)$ 定义为 n 的条件分布概率，则线性 CRF 中的输出序列 n 的联合概率定义为：

$$z=\sum_{n}\exp\left[\sum_{i,k}\lambda_k f_k\left(T_{i-1},T_i,m,i\right)\right] \tag{1}$$

$$p(n|m)=\frac{1}{z}\exp\left[\sum_{i,k}\lambda_k f_k\left(T_{i-1},T_i,m,i\right)\right] \tag{2}$$

其中：z 为规范因子，f_k 为特征函数，λ_k 是对应的权重。上式表示在输入数据序列 m 的条件下，得到输出序列 n 的概率。

1.3 长短期记忆网络模型

长短期记忆网络模型（Long Short-Term Memory，LSTM）是对循环神经网络模型（Recurrent Neural Network，RNN）改进后的特殊形式的模型，由 Hochreiter 等人[18]于 1997 年提出，主要思想是通过改变 RNN 中的隐藏层机构，采用门结构方式控制 RNN 中信息的传播方式，通过不同门结构来控制信息的输入、遗忘、变换、输出等过程。LSTM 的缺点是无法完整获取语句的上下文信息，因此，研究者们采用双向长短记忆网络（Bi-directional Long Short-Term Memory，BiLSTM）方法。

1.4　BiLSTM–CRF 模型结构

将 CRF 模块作为 BiLSTM 模块的输出层,解决了字向量经过 BiLSTM 层后可能得到无效标签序列的问题。CRF 层将 BiLSTM 层输出的标签数列进行集中解码,获得整个句子的序列标注,而不是仅对单一标签进行单独的解码。BiLSTM 模型加入 CRF 层后可以考虑到不同类型标签之间的关联性,使得输入的数据序列经过模型处理后可以得到一个最优的标签序列。BiLSTM–CRF 模型结构图如图 1 所示。

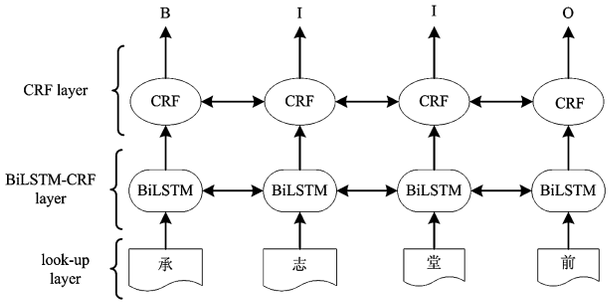

图 1　BiLSTM–CRF 模型结构图

2　徽派建筑知识图谱的半自动化构建

徽派建筑知识图谱的构建分为四个步骤,如图 2 所示。

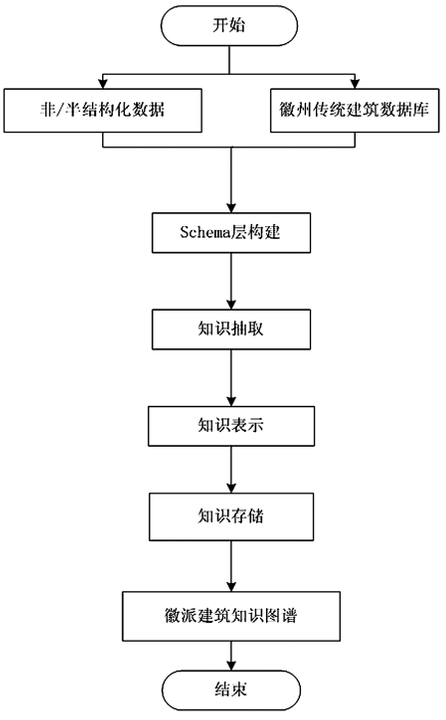

图 2　徽派建筑知识图谱构建流程图

(1)概念层的构建。本文采用传统的自顶向下的方法构建了徽派建筑知识图谱的基本概念层。

(2)利用结构化、半结构化以及非结构化的数据,包括网页数据,现有数据库等抽取实体、属性以及关系,然后进行命名实体识别。

(3)知识表示。徽派建筑知识图谱使用属性图为基本的表示形式。

(4)知识存储。使用 Neo4j 图数据库存储徽派建筑知识数据。

2.1　概念层构建

概念层构建是对徽派建筑知识图谱主体框架的构建,需要定义类及类之间的关系,即对知识图谱中的概念及概念之间的语义关系进行定义。

本文构建的是徽派建筑知识图谱,以民居、祠堂为主,设计并构建了徽派建筑领域知识图谱的概念层,主要从建筑基本信息、建筑平面信息、建筑立面、建筑空间分布、雕刻、文化特色六大类进行定义。

概念类通过相关属性进行详细描述,传统建筑基本信息属性包括建筑名称、类型、坐落位置、建造时期。建筑平面的属性包括建筑开间、布局、外观。立面属性包括马头墙、门楼。空间属性有檐高、屋脊高度、院落进数和拼接方式。雕刻属性包括石雕、砖雕、木雕。徽派建筑知识图谱模式层如图 3 所示。

2.2　基于 BiLSTM–CRF 模型的命名实体识别

从获取到的原始数据文本中获取实体是构建徽派建筑知识图谱的关键步骤,基于神经网络的命名实体识别模型已在通用领域中广泛使用[19]。神经网络模型方法的优势在于可以对数据特征进行自动提取,而且训练模型的过程是端到端的,生成的模型可以直接用于命名实体识别。因此,本文采用 BiLSTM–CRF 学习框架与徽派建筑词典相结合的方法,对徽州传统建筑的命名实体进行识别。图 4 为徽派建筑命名实体识别关键技术框架图。

2.2.1　分词

本文选择基于中文语料库的 CorpusWordParser 进行分词。CorpusWordParser 基于现代汉语通用平衡语料库开发,具有中文分词和词性标注等功能,用户可以自行添加词表来增强分词效果。分词结果如图 5 所示。

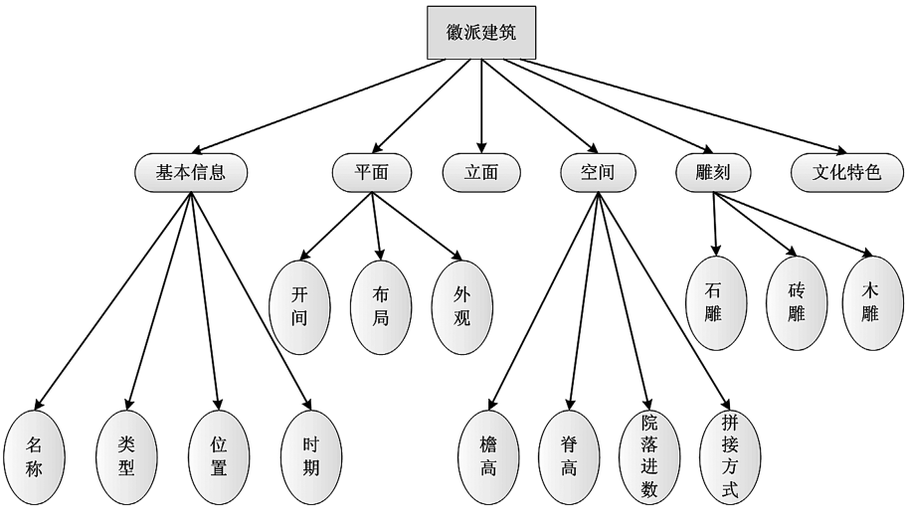

图 3　徽派建筑知识图谱模式层

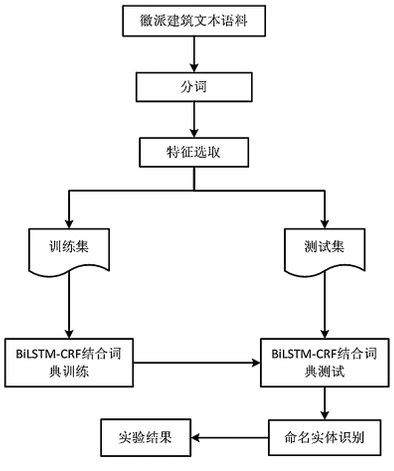

图 4　徽派建筑命名实体识别关键技术框架图

2.2.2　数据搜集与处理

由于目前缺乏用于徽派建筑命名实体识别的公开数据集，本文通过实验室已建成的数据库及百科词条构造了徽派建筑语料库，语料库涵盖了徽派建筑典型的建筑类型。另外，实验室已建成的国内唯一的徽州传统建筑特征元素数据库，收集了近百种建筑构件相关数据近万条，徽州地区 100 多个聚落、万幢建筑的相关信息。本文的实验数据来自经过整理分析的数据库数据和相关徽派建筑百度百科词条信息。

本文抽取了 168 个描述徽派建筑文本，将其中 80% 的样本数据作为训练集，20% 作为测试集。

图 5　分词结果图

当训练出的准确率达到设定的标准后,用训练好的模型从 168 条文本中抽取徽派建筑的实体,使用 BRAT 文本标注工具进行数据标注,对获取的语料进行数据格式转换。按照 BIO 格式对语料进行标记,标记为:B、I、O,分别表示实体的首字符、中间字符和非建筑名词。

2.2.3　BiLSTM-CRF 与徽派建筑词典相结合的命名实体识别

通过对《中国古代建筑辞典》的参考分析,构建本文所需要的徽派建筑词典,通过词典来获取非结构化文本中的语料类别信息,把获取的信息作为特征值传递给 BiLSTM-CRF 模型去识别数据中的徽派建筑实体,类别信息如表 1 所示。本文将描述徽派建筑的数据分为两类,一类是描述徽派建筑的术语,标记为 “HA” 。其他非建筑术语,标记为 “HO” 。

表 1　类别信息

类别	描述	例子	标记
建筑术语	承志堂整幢房屋开了九间天井	天井、马头墙	HA
其他	其他词汇	八仙椅、蟠龙	HO

2.2.4　实验与结果

实验抽取 168 条非结构化文本数据,任意选取其中的 130 条数据进行命名实体识别。将 100 条数据作为训练样本,30 条数据作为测试样本。为了缓解模型存在过拟合性,将 BiLSTM 模型网络输入与输出端的 Dropout rate 值设为 0.5,实验结果如表 2 所示。

表 2　识别结果统计

模型	Precision	Recall	F-measure
BiLSTM	0.7868	0.8360	0.8107
BiLSTM-CRF[20]	0.7865	0.8438	0.8141
BiLSTM-CRF+Dic	0.8214	0.8572	0.8382

为了判别 BiLSTM-CRF 模型结合徽派建筑词典特征的性能,分别进行了 BiLSTM 模型、BiLSTM-CRF 模型和 BiLSTM-CRF 模型结合徽派建筑词典特征的对比实验。根据表 2 的实验结果可以看出,结合词典特征的 BiLSTM-CRF 模型比其他两组实验,在准确率、召回率和 F1 值上都取得了最好的效果。BiLSTM-CRF 模型比 BiLSTM 模型效果好,是因为 BiLSTM-CRF 模型能够利用上下文的语义信息以及相邻标签间的关系,产生更优的标签序列。结合徽派建筑词典特征的 BiLSTM-CRF 模型比单独使用 BiLSTM-CRF 模型准确率提升了 3.49%,召回率上升了 1.34%,F1 值提高了 2.41%。分析实验结果发现,在徽派建筑训练数据集中没有明显特征的建筑名词被结合词典的 BiLSTM-CRF 模型准确地识别了出来,体现了作为先验知识的词典对实体识别起到了重要的辅助作用。例如佛塔,在徽州区,塔主要指的是村口的风水塔,如黟县柯村乡的旋溪塔。佛塔的相关数据不多,在本文的训练样本中没有描述佛塔的术语,但是徽派建筑词典能准确的识别出此类建筑术语,利用这些建筑术语的语料信息为 BiLSTM-CRF 模型提供支持,使得识别效果更好。

因为实验在准确率、召回率和 F 值上都取得了比较好的效果,因此,本文利用结合徽派建筑词典的 BiLSTM-CRF 模型,对 168 条非结构化数据进行徽派建筑实体的抽取,共抽取出 504 个徽派建筑实体。

2.3　知识表示

知识图谱是一种网络结构图,实体就是图里面的节点,实体之间的关系就是图的边。知识图谱有两种表示形式:三元组和属性图[21]。本文采用 Neo4j 图数据库来存储徽派建筑领域知识,用属性图模型表示知识。

属性图模型就是顶点、边、标签、关系类型和属性组成的有向图。实体可以表示成一个或多个键值对形式的属性:

(1)顶点。每个顶点具有一个唯一的 ID,每个顶点还有一个实体类,表示顶点所对应的概念类型,每个顶点属性的集合通过键值对来表示。

(2)边。每一条边都有一个唯一的 ID,每一条边都有一个头结点和尾结点。同时,每一条边有一个实体类 type,表示头节点和尾结点的关系,每条边也由键值对来定义边属性集合。

图 6 为 Neo4j 的一个实体属性图模型,实体大菩萨厅和空间布局串联之间的关系是拼接方式。其中,id 是实体的位置符号,是其唯一的标识符;type 表示实体类别;start 表示头结点 id;end 表示尾结点 id;name 表示对应节点属性描述。

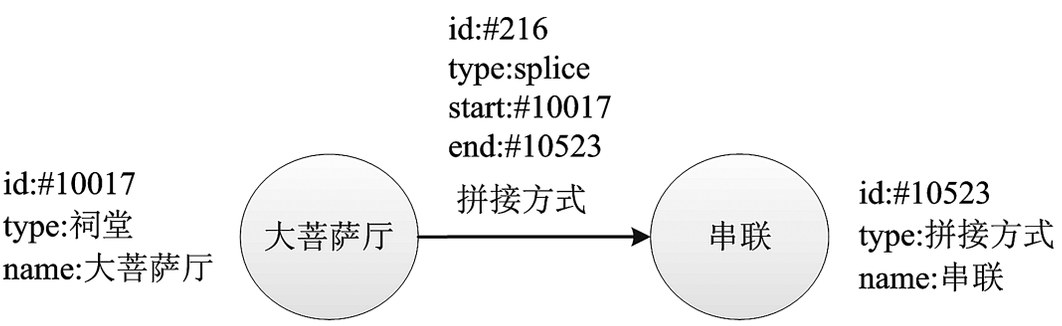

图 6 实体的属性图模型

2.4 知识存储与可视化展示

本文采用的是 Neo4j 图数据库来存储数据。Neo4j 是一个高性能的、基于 Java 语言开发并且开源的图形数据库，它将结构化数据存储在灵活的、面向对象的网络结构中而不是表格中，它还具备了完整的数据库特性。通过对知识图谱进行可视化展示，使得用户更加直观、清晰地了解实体之间的关联规则。

将提取出的徽派建筑的建筑实体以及实体与实体之间的关系等数据信息全部整理成结构化的数据，并设置相应的概念类存储在 CSV 格式的文件中，如表 3、4 所示。然后将所有的 CSV 文件通过 Cypher 导入到 Neo4j 图数据库中，即完成了徽派建筑知识图谱的数据存储。

表 3 徽派建筑实体在 CSV 文件中的录入格式

徽派建筑类型实体	标签
民居	类型
祠堂	类型
书院	类型
牌坊	类型
佛塔	类型

表 4 徽派建筑实体、类型实体及两者关系在 CSV 文件中的存储格式

徽派建筑实体	关系	类型
承志堂	是	民居
大菩萨厅	是	祠堂
大胜桥	是	桥梁
耕读园	是	书院

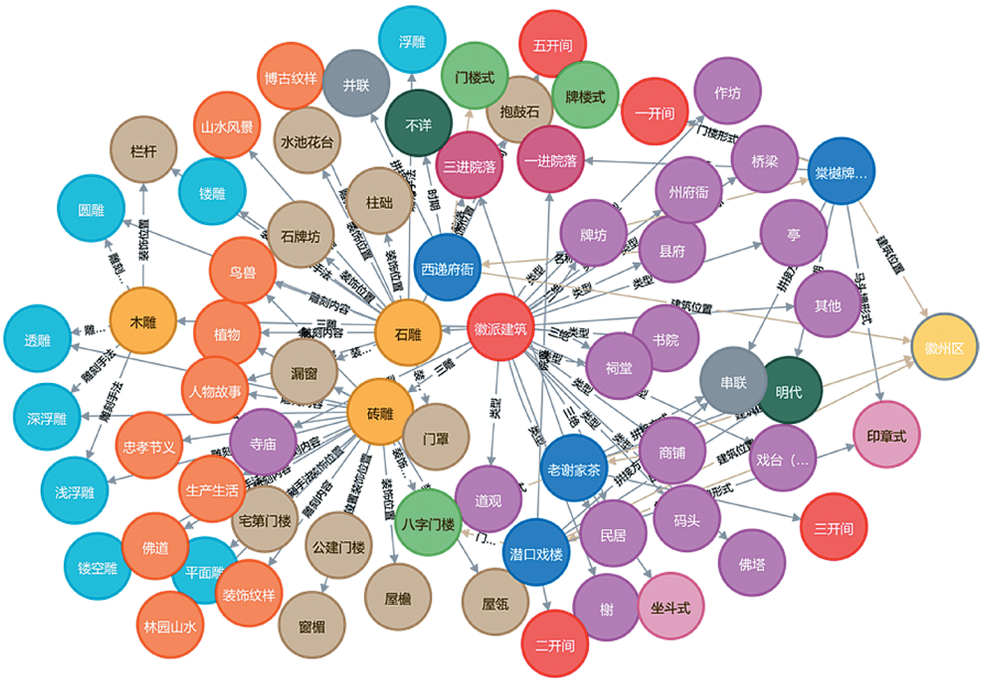

图 7 徽派建筑图谱（节选）

将徽派建筑知识数据存储到 Neo4j 图数据库后,图 7 徽派建筑知识图谱(节选)中紫色圆圈表示建筑类型实体,深蓝色圆圈表示建筑实体,连接建筑类型实体与建筑实体之间的线段表示这些实体之间相对应的关系。图中展示了包括民居、祠堂、牌坊等 16 种不同的建筑类型,每种建筑实体展示出了建筑位置,开间及门楼形式等信息,同时介绍了徽派建筑著名的三雕技术,包括 7 种雕刻手法,形式多样的雕刻内容和装饰位置等信息。Neo4j 图数据库使用 Cypher 语言对数据库进行增删改查操作,实现了对每一座建筑的检索、遍历等功能。

3　结论

本文详细描述了在传统建筑领域通过数据抽取来构建徽派建筑知识图谱的方法,并介绍了徽派建筑知识图谱的构建流程。针对徽派建筑数据异构多源和非结构化的特点,提出了 BiLSTM-CRF 模型结合徽派建筑词典的方法来对徽派建筑实体进行识别抽取。实验结果表明,在先验知识的辅助作用下,实体识别的效果更好。在获取到徽派建筑的知识之后,利用 Neo4j 数据库存储知识,用属性图模型表示知识。最后利用 Neo4j 图数据库可视化地展示了构建的徽派建筑知识图谱。本文所构建的徽派建筑知识图谱,为研究徽派建筑知识的智能化推荐和搜索系统奠定了基础。

参考文献:

[1] 季阳 . 关于传统元素在现代建筑装饰设计中的运用研究 [J]. 中国建筑装饰装修,2019(11):121.

[2] 漆桂林,高桓,吴天星 . 知识图谱研究进展 [J]. 情报工程,2017,3(1):4-25.

[3] 高龙,张涵初,杨亮 . 基于知识图谱与语义计算的智能信息搜索技术研究 [J]. 情报理论与实践,2018,41(7):42-47.

[4] 杜泽宇, 杨燕, 贺樑 . 基于中文知识图谱的电商领域问答系统 [J]. 计算机应用与软件,2017,34(5):153-159.

[5] 刘昱良 . 基于知识图谱的个性化学习资源推荐研究 [D]. 新乡:河南师范大学,2018.

[6] 祁志武 . 地质标本虚拟学习平台构建 [D]. 荆州:长江大学,2018.

[7] 王良萸 . 面向碳交易领域的知识图谱构建方法 [J]. 计算机与现代化,2018(8):114-119.

[8] 汤洁 . 构建金融知识图谱以及投资关系分析 [D]. 武汉:华中科技大学,2018.

[9] 赵高敏,马慧子,郭雨婷 . 基于知识图谱的我国互联网金融研究可视化分析 [J]. 商业经济研究,2019(2):154-156.

[10] 袁凯琦,邓扬,陈道源,等 . 医学知识图谱构建技术与研究进展 [J]. 计算机应用研究,2018,35(7):1929-1936.

[11] 钟亮 . 面向百度百科的化学知识图谱构建方法研究 [J]. 软件导刊,2017,16(8):168-170.

[12] 兰海峰,秦为径,成斌,等 . 凉山彝族地区乡土景观基因图谱构建及其保护传承研究 [J]. 安徽农业科学,2018,46(21):220-222.

[13] 聂聆 . 徽州古村落景观基因识别及图谱构建 [D]. 合肥:安徽农业大学,2015.

[14] 翟洲燕,常芳,李同昇,等 . 陕西省传统村落文化遗产景观基因组图谱研究 [J]. 地理与地理信息科学,2018,34(3):87-94,113.

[15] Sun Y H,Sarwat M.A spatially-pruned vertex expansion operator in the Neo4j graph database system[J]. GeoInformatica,2019,23(3):397-423.

[16] Francis N,Green A,Guagliardo P,et al.Cypher:an evolving query language for property graphs[C]// Proceedings of the 2018 International Conference on Management of Data. Houston TX USA.New York,NY,USA:ACM,2018:1433-1445.

[17] John L,Andrew M,Pereira Fernando C N.Conditional random fields:probabilistic models for segmenting and labeling sequence data[J].ICML ' 01:Proceedings of the Eighteenth International Conference on Machine Learning,2001,2001(June):282-289.

[18] Greff K,Srivastava R K,Koutn í k J,et al.LSTM:a search space odyssey[J].IEEE Transactions on Neural Networks and Learning Systems,2017,28(10):2222-2232.

[19] Luo L,Yang Z H,Yang P,et al.An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J].Bioinformatics,2018,34(8):1381-1388.

[20] Zhiheng Huang,Wei Xu,Kai Yu. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer Science,arXiv:1508.01991,2015.

[21] Arenas M,Cuenca Grau B,Kharlamov E,et al.Faceted search over RDF-based knowledge graphs[J].Journal of Web Semantics,2016,37/38:55-74.