张涛, 周卫. 广西非遗文化知识图谱构建与数据处理研究[J]. 智能计算机与应用,2025,15(3):72-78. DOI:10.20169/j.issn.2095-2163.250310

广西非遗文化知识图谱构建与数据处理研究

张 涛[1], 周 卫[2]
(1 广西民族大学 电子信息学院, 南宁 530006; 2 广西民族大学 人工智能学院, 南宁 530006)

摘 要: 非物质文化遗产代表着地区文化历史的沉淀,是中华优秀传统文化的重要组成部分,也是人类文明的宝贵财富,具有无可替代的历史文化价值。对于维护文化多样性来说,保护和传承非物质文化遗产至关重要。然而,在当前网络环境下,广西的非物质文化遗产信息存在着杂乱无章、结构不清晰的问题。针对此问题,通过采用 Python 爬虫技术,对广西非物质文化遗产信息进行了系统采集,通过应用自然语言处理模型、特别是命名实体识别和关系抽取技术,能够将其中的非结构化信息转化为结构化数据,随后对这些数据进行了全面整理和清洗。最终,运用知识图谱技术的强大信息整合和表示能力,成功构建出一个结构清晰的广西非物质文化遗产知识图谱。

关键词: 知识图谱; Python 爬虫; 命名实体识别; 关系抽取; Neo4j 图数据库; RoBERTa
中图分类号: TP393 **文献标志码**: A **文章编号**: 2095-2163(2025)03-0072-07

Research on the construction and data processing of knowledge graph for Guangxi intangible cultural heritage

ZHANG Tao[1], ZHOU Wei[2]
(1 College of Electronic Information, Guangxi Minzu University, Nanning 530006, China;
2 College of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China)

Abstract: Intangible cultural heritage represents the cultural history of a region and is an important component of traditional Chinese culture. Meanwhile, it is also a precious asset of human civilization, and has irreplaceable historical and cultural value. For the preservation of cultural diversity, the protection and inheritance of intangible cultural heritage are of paramount importance. However, in the current online environment, there are issues with the disorderly and unclear structure of Guangxi's intangible cultural heritage information. To address this problem, systematic collection of Guangxi's intangible cultural heritage information is conducted using Python Web scraping technology. By applying natural language processing models, especially named entity recognition and relationship extraction techniques, the unstructured information within intangible cultural heritage is transformed into structured data, followed by careful organization and cleansing of this data. Ultimately, leveraging the powerful information integration and representation capabilities of knowledge graph technology, a well-structured knowledge graph for Guangxi's intangible cultural heritage is successfully constructed.

Key words: knowledge graph; Python Web crawler; named entity recognition; relationship extraction; Neo4j graph database; RoBERTa

0 引 言

广西是一个多民族共存的地方,生活着壮、侗、瑶等 12 个世居民族,在长期的历史发展过程中,民族相互依存、共同发展,创造了丰富多彩、特色鲜明的非物质文化,为丰富和发展中华民族文化做出了贡献[1]。壮族的刘三姐歌谣、侗族的侗戏、瑶族的密洛陀等都是广西独有的文化瑰宝。

非物质文化遗产是民族文化的重要组成部分,是民族精神和情感的重要载体,也是维系国家统一、民族团结的思想基础,是联系世界的桥梁,也是人民智慧的结晶。非遗文化对于维护文化多样性、促进文化交流与对话,以及实现可持续发展都具有至关重要的作用。因此保护传承和传播广西非物质文化遗产,不仅是文化造福于广西人民经济发展的独特动力源泉,更是广西走向世界的一扇门户[2]。

作者简介: 张 涛(1998—),男,硕士研究生,主要研究方向:自然语言处理,问答系统。
通信作者: 周 卫(1980—),男,博士,副教授,主要研究方向:大型软件系统,计算机视觉,自然语言处理等。Email: zhouwei@gxmzu.edu.cn。

收稿日期: 2023-09-08

知识图谱的概念是 Google 于 2012 年正式提出[3]，自从谷歌首次引入这个概念以来，知识图谱已经成为信息科学领域的一个热门话题，引起了学术界、工业界和广大用户的广泛兴趣。知识图谱可细分为通用知识图谱与领域知识图谱，两者主要存在覆盖范围和使用方式上的差异。其中，通用知识图谱面向通用领域，包含了大量常识性知识，覆盖面广[4]。领域知识图谱又称为行业知识图谱或垂直知识图谱，是面向某一特定领域的，是由该领域的专业数据构成的行业知识库[5]，因其基于行业数据构建，有着严格而丰富的数据模式，所以对该领域知识的深度、知识准确性有着更高的要求。

知识图谱由三元组组成，三元组的一般结构为“实体-关系-实体”或“实体-属性-值”[6]。这种三元组结构的设计使得知识图谱能够以一种机器可理解的方式表示丰富的信息，不仅包括实体之间的关系，还包括实体的属性信息。这使得知识图谱成为了用于知识表示[7]、推理和检索的有力工具，能够帮助计算机更好地理解和处理语义信息。知识图谱的应用领域广泛，包括自然语言处理[8]、搜索引擎改进、智能问答系统、数据集成和语义网等。

通过研究使用了多种方法，包括手动下载和 Python 爬虫，以全面收集网络中广西非物质文化遗产的信息。首先，对这些非物质文化遗产信息进行了分类。对于那些非结构化的数据，采用了 RoBERTa+BiLSTM+CRF 模型进行命名实体识别和关系抽取，以提取有用的信息并将其组织成结构化数据。然后，将结构化数据整合在一起进行数据去重和错误修复的工作。最后，通过 Python 的 py2neo 库，将这些数据转化为节点和关系，从而构建一个广西非物质文化遗产的领域知识图谱。

广西非物质文化遗产图谱的构建不仅有助于保护和传承广西当地宝贵的文化遗产，还为学术研究、旅游推广、非遗政策制定等方面提供了有力的支持，促进了文化的多层面发展。

1　数据来源

鉴于非物质文化遗产的独特性，网络上目前尚无公开的可供使用的数据集。因此，采取了主要依赖 Python 爬虫的方式来收集非物质文化遗产相关的数据，以构建非遗数据集。

通过爬取广西非物质文化遗产网、广西壮族自治区文化旅游厅官网以及广西各城市非物质文化遗产信息网等多个网站，成功地搜集到了超过 1 100 项广西非物质文化遗产项目的信息，同时也获取了来自 900 多位非物质文化遗产传承人的相关信息。这一全面的数据收集工作为后续的研究提供了坚实的数据基础，有助于深入探讨和记录广西非物质文化遗产的珍贵内容，以促进其保护和传承。

1.1　HTML

HTML（HyperText Markup Language）是一种标记语言，用于创建网页。由一系列的标签组成，每个标签都描述了网页上的不同元素，如文本、图像、链接等。

HTML 标签通常由尖括号括起来，如<html>、<head>、<body>等。这些标签可以包含属性，属性提供了有关元素的额外信息。

HTML 语言是万维网的核心标准之一。万维网是一个由许多相互链接的超文本文档构成的系统，在这个系统中，每个资源都有一个唯一的 URL 来标识，并通过超文本传输协议（HTTP）向用户提供。用户可以通过点击链接来访问不同的 URL，获取 HTML 网页中包含的各种信息。通过使用各种技术、如数据抓取和信息提取，来从这些网页中筛选出所需的数据，以满足各种需求。这使得互联网成为一个信息丰富的资源池，为人们提供了广泛的信息和服务。

1.2　requests

1.2.1　requests 介绍

requests 是 Python 中一个常用的库，用于发送 HTTP 请求和处理 HTTP 响应。

requests 可以让使用者与 Web 服务进行通信，获取和发送数据。

1.2.2　requests 使用

（1）*reps = requests. get*（*url*）：使用 *get* 函数来向指定的 URL 发送一个 HTTP GET 请求，并将获取到的响应存储在名为 *reps* 的变量中。

（2）*resp. status_code*：*resp* 对象将包含服务器对请求的响应信息，其中 *status_code* 属性存储了服务器返回的 HTTP 状态码。

HTTP 状态码是一个三位数的数字，表示了请求的处理结果。常见的 HTTP 状态码见表 1。

表 1　常见 HTTP 状态码

Table 1　Common HTTP status codes

HTTP 状态码	返回	表示含义
200	OK	请求成功
404	NotFound	请求的资源不存在
500	InternalServerError	服务器内部发生错误

(3) *resp. text*: 发送 HTTP GET 请求后, 可以使用 *resp. text* 来获取响应的文本内容。

1.3 BeautifulSoup

1.3.1 BeautifulSoup 介绍

BeautifulSoup 是一个用于解析 HTML 和 XML 文档的 Python 库,可以帮助使用者从网页或其他文档中提取数据,非常适合用于网页抓取和数据挖掘任务。

1.3.2 BeautifulSoup 使用

(1) *soup* = BeautifulSoup (html _ doc, ´html. parser´): BeautifulSoup 库将 HTML 文档解析为一个 BeautifulSoup 对象,创建的 BeautifulSoup 对象 *soup*, 可以借此来浏览、搜索和操作 HTML 文档的各个部分。

(2) *soup. find*("div") : *find*( ) 方法可以查找特定的 HTML 元素,如 div。

(3) *soup. find_all*("div") : *find_all*( ) 方法可以查找多个某一类 HTML 元素。

1.4 爬取数据展示

爬取的数据包含了结构化和非结构化信息,如图 1~图 4 所示。

图 1 结构化非遗信息

Fig. 1 Structured intangible cultural heritage information

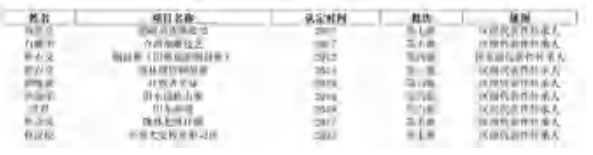

图 2 结构化传承人信息

Fig. 2 Structured inheritor information

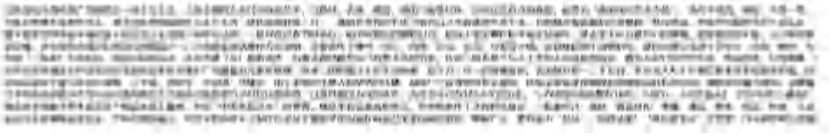

图 3 非结构化非遗信息

Fig. 3 Unstructured intangible cultural heritage information

图 4 非结构化传承人信息

Fig. 4 Unstructured inheritor information

2 知识抽取

2.1 RoBERTa 模型

2.1.1 RoBERTa 及其特点

RoBERTa (A Robustly Optimized BERT Pretraining Approach)[9] 是一种基于 Transformer[10] 架构的自然语言处理模型,是由 FacebookAI 于 2019 年提出的。RoBERTa 模型基于 BERT(Bidirectional Encoder Representations from Transformers)[11] 的思想,但通过一系列优化策略,取得了更好的性能。

RoBERTa 模型的特点主要有:

(1)更大的训练数据。RoBERTa 使用了远比 BERT 更大的文本语料库进行预训练,这使得模型能够更好地理解自然语言。

(2)更长的训练时间。RoBERTa 进行了更长时间的预训练,这有助于模型更好地捕捉语言的复杂性。

(3)动态掩码。RoBERTa 使用了一种动态掩码策略,对输入文本中的词汇进行了更多的掩码,从而使模型能够更好地学习上下文。

(4)更长的序列长度。RoBERTa 支持更长的序列长度,这在 NLP 任务中非常有用。

2.1.2 RoBERTa 模型变种 RoBERTa-wwm-ext-base

鉴于非物质文化遗产中文文本的特殊性,采用 RoBERTa 模型变种—中文 RoBERTa-wwm-ext-base[11-12] 预训练模型进行实验。

RoBERTa-wwm-ext-base 是在大规模的中文文本数据上进行预训练的。这些数据通常包括互联网上的中文文本,例如维基百科、新闻文章、社交媒体帖子。预训练之后,RoBERTa-wwm-ext-base 模型可以进行微调,以适应特定的自然语言处理任务,例

如文本分类、命名实体识别、关系抽取、情感分析等。

2.2　BiLTSM 模型

长短时记忆网络 LSTM（LongShort－TermMemory）[13]是一种循环神经网络（RNN）[14]，专门设计用来处理序列数据，具有捕捉长期依赖性的能力，避免了传统 RNN 中的梯度消失问题。LSTM 通过细胞状态（Cell State）和门控机制（Gate Mechanism）来管理信息的流动，使其适用于各种序列任务。

BiLSTM（Bidirectional LongShort－TermMemory，BiLSTM）就是双向 LSTM[15]，单向的 LSTM 模型只能捕捉到从前向后传递的信息，而双向的网络可以同时捕捉正向信息和反向信息，使得模型能够同时考虑过去和未来的信息，更好地捕捉到序列数据中的上下文关系。

2.3　CRF 模型

条件随机场（Conditional Random Field，CRF）[16-17]是一种概率图模型，通常用于序列标记任务，例如命名实体识别（NER）、词性标注（POStagging）、语法分析等。

CRF 属于概率图模型的一种，其中建模了输入序列和输出标签序列之间的条件概率分布。具体而言，CRF 用于对给定输入序列条件下的输出序列进行建模，从而在序列标记任务中捕捉标签之间的依赖关系。在 NER 任务中，CRF 的输入为文本序列，输出则是识别出的命名实体的标签序列。每个标签对应输入文本中的一个位置。

2.4　命名实体识别

命名实体识别（Named Entity Recognition，NER），也称为实体识别，其主要目标是从文本中识别并分类出具有特定意义的实体，如人名、地名、组织机构名、日期等。NER 的目的是将文本中的词语或短语与预定义的实体类别相匹配，以便更好地理解文本的语义结构。

通过采用 RoBERTa－wwm－ext－base 结合 BiLSTM 和 CRF 的模型，成功地对广西非物质文化遗产信息的非结构化部分进行了实体提取，准确地识别了非遗的各个关键信息，包括非遗名称、非遗分类、非遗级别、录用批次、收录时间、申报地区以及所属地域等。在地域信息方面，具体到市级，例如，提取出布洛陀属于百色市；而在申报地区方面，具体到了县级，比如，确定布洛陀的申报地区为百色市田阳区。此外，还对非结构化的广西非遗传承人信息进行了实体提取，成功地识别出了传承人的姓名、民族、出生年份以及代表性的非遗项目等重要信息。命名实体识别如图 5 所示。

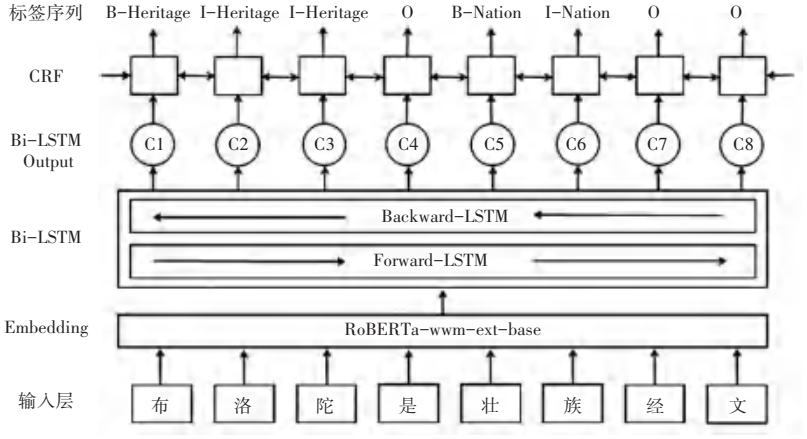

图 5　RoBERTa+BiLSTM+CRF 模型

Fig. 5　RoBERta+BiLSTM+CRF model

2.5　关系抽取

关系抽取（Relation Extraction，RE）是自然语言处理中的一个关键任务，其目标是从文本中识别出不同实体之间的语义关系。关系抽取通常涉及以下几个方面：

（1）实体识别。从文本中识别出与关系抽取任务相关的实体，例如人名、地点、组织等。

（2）关系分类。对抽取的实体之间的关系进行分类。关系可以是预定义的，也可以是动态确定的。

例如，“洪琪于 2009 年被荣誉命名为第三批国家级项目的邕剧自治区级代表性传承人”这句话，通过关系抽取，研究成功提取了实体“洪琪”和“邕剧”，并确定了两者之间的关系，即“邕剧的传承人是洪琪”。

2.6 数据整理

经过命名实体识别和关系抽取的结构化非遗数据还存在以下问题：

(1)数据不完整。某些信息因为收录年份过于久远,可能导致无法从网络上获取。例如,在非物质文化遗产传承人中,可能存在一些已经去世的情况,因此与其相关的民族和出生年份等信息可能无法获得。

(2)数据不一致。同一个非物质文化遗产项目可能会在不同的网站上具有不同的信息,例如收录时间。这意味着需要进行手动搜索和比对,以确保数据的准确性和一致性。

(3)数据重复。同一个非物质文化遗产项目可能会有多个不同的名称,例如藤县狮舞和舞狮(藤县),实际上指的是同一个非物质文化遗产名录项目。因此,需要对这些信息进行修正和统一,以确保数据没有存在冗余的现象。

(4)数据更新。非物质文化遗产信息可能随时间而变化,例如,以“壮族百衣鸟故事”为例,最初的记录显示其录用时的保护单位为横县文化馆。然而,截至 2021 年,横县已撤销,建立了县级横州市,同时横县文化馆也更名为横州市非物质文化遗产保护中心。

3 知识图谱构建

3.1 Neo4J 图数据库

3.1.1 Neo4J 图数据介绍

Neo4j[18]是一个广泛用于存储、查询和可视化图数据的图数据库管理系统。能以图形数据库的方式存储数据,图数据库与传统的关系型数据库有所不同,适用于需要处理复杂关系和连接的数据集。Neo4j 图数据库的一些关键特点和概念如下。

(1)节点(Nodes):节点是图数据库中的基本单元,用于表示实体。每个节点可以包含属性(Properties),这些属性是键值对,用于描述节点的特征。

(2)关系(Relationships):关系用于连接节点,表示节点之间的关联性。关系也可以包含属性,用于描述关系的特征。

(3)属性(Properties):节点和关系都可以包含属性,这些属性用于存储额外的数据信息。属性可以是文本、数字、日期等类型的数据。

(4)查询语言(Cypher)[19]:Neo4j 使用 Cypher 查询语言进行数据查询和操作。Cypher 是一种类似 SQL 的查询语言,专门用于处理图数据。

3.1.2 主流图数据库对比

目前,主流图数据库[20]有很多,每个都有其自己的特点和适用场景。主流的图数据库特点和对比见表 2。

表 2 主流图数据库对比

Table 2 Comparison of mainstream image databases

图数据库	特点	适用场景
Neo4j	具有强大的图算法库和易于使用的查询语言 Cypher。具有 ACID 事务支持	社交网络分析、推荐系统、网络和 IT 运营管理等需要复杂图结构的应用
AmazonNeptune	完全托管的服务,兼容 Gremlin 和 SPARQL 两种查询语言,适用于 AWS 环境	需要高度可扩展性和容灾能力的应用
OrientDB	支持多模型数据库,可以存储图形、文档、键值等数据。具有 ACID 事务支持	适用于混合数据模型的应用,如内容管理系统和物联网
TigerGraph	高度并行的图数据库,适用于大规模图分析任务,具有内置的图算法	需要高性能的图分析、推荐和欺诈检测等应用
JanusGraph	基于 ApacheTinkerPop 框架的分布式图数据库,支持 Gremlin 查询语言	复杂图分析和关系建模的应用

3.2 py2neo

py2neo 是一个用于与 Neo4j 图数据库进行交互的 Python 库。主要提供了一种便捷的方式来连接、查询和操作图数据库。使用 py2neo,开发者可以轻松地创建、修改和查询节点、关系以及其属性,从而实现了对图数据库的全面控制。py2neo 主要有以下方法：

(1)graph。连接到 Neo4j 数据库。

(2)run。执行 cypher 语言。

(3)create(node)。创建节点。

(4)create(relationship)。创建关系。

(5)connecter. create_node_matcher()。查找节点。

(6)connecter. create_relationship_matcher()。查找关系。

4 知识图谱可视化

最终构建的广西非遗知识图谱共包含 2 787 个

节点，14 308 条关系。实体节点类型见表 3。

表 3　实体标签和实体类型

Table 3　Entity labels and entity types

实体标签	实体类型
Batch	非遗或传承人收录年份
Category	非遗类别
Heritage	非遗项目
Inheritor	非遗传承人
Nation	民族
Region	地区

实体节点间关系见表 4。

表 4　实体间关系及含义

Table 4　Relationship and meaning between entities

关系名	含义
belongs_to	地区从属关系
declaration_region	非遗项目申报地区
heritage_category	非遗所属类别
representative_inheritor	非遗代表性传承人
heritage_level	非遗所属级别
inheritor_level	非遗传承人所属级别
protection_institution	非遗保护机构

研究得到的知识图谱展示如图 6 所示。

图 6　实体节点及实体间关系

Fig. 6　Entity nodes and relationships between entities

5　结束语

本文致力于构建广西非物质文化遗产及传承人知识图谱，旨在为这一宝贵的文化遗产提供更好的保护、传承和推广手段。研究中采用多种技术手段，包括 Python 爬虫技术、自然语言处理模型和知识图谱构建工具，从而功地将广西非物质文化遗产的信息成功地从网络混沌中提取出来，并整合成了一个完整的知识图谱。非遗知识图谱的构建不仅为学者、研究人员和文化工作者提供了宝贵的研究资源，还有助于在更大范围内大力推广广西非物质文化遗产。通过本文研究工作，能够更好地记录、传承和传播这些珍贵的非遗文化，使非遗文化在当代社会生活、文化实践和文化建设中发挥应有的作用。

然而，知识图谱的建立只是一个开端，未来仍需进行大量工作。在未来的研究中，需要不断探索更为先进的技术手段，以丰富知识图谱的内容，使其更加全面。此外，还需要深入思考如何更好地将这些非物质文化遗产融入当代社会，促进其传承与创新。最后，本次研究能够为广西非物质文化遗产及传承人知识图谱的构建提供积极助力，也为将来的研究工作打下坚实的基础。

参考文献

[1] 覃健. 广西非物质文化遗产保护研究[J]. 知识文库，2016(11)：13-14.

[2] 梁松，何成战. 广西非物质文化遗产的数字化传承与传播研究[J]. 大众科技，2018，20(4)：96-98.

[3] Google. Introducing the Knowledge Graph：Things，notStrings. GoogleBlog[EB/OL]. [2019-11-04]. https://blog. google/products/search/introdcing-knowledge-graph-things-not/.

[4] 王晓剑. 基于领域知识图谱的服务社区管理技术研究[D]. 北京：北方工业大学，2022.

[5] 唐照翔. 基于知识图谱的图神经网络推荐系统研究及其应用[D]. 重庆：重庆大学，2021.

[6] 胡杰，李源洁，耿號，等. 基于深度学习的汽车故障知识图谱构建[J]. 汽车工程，2023，45(1)：52-60.

[7] CHANG C C，LIN C J. LIBSVM：A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST)，2011，2(3)：27.

[8] SUTSKEVER I，VINYALS O，LE Q V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems，2014，27：3104-3112.

[9] LIU Zhuang，LIN W，SHI Yawei，et al. A robustly optimized BERT pre-training approach with post-training[C]//China National Conference on Chinese Computational Linguistics. Cham：Springer，2021：471-484.

[10] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. Advances in Neural Information Processing Systems，2017，30：5998-6008.

[11] DEVLIN J，CHANG Mingwei，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv，1810.04805，2018.

[12] CUI Yiming，CHE Wanxiang，LIU Ting，et al. Pre-training with whole word masking for Chinese bert[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2021，29：3504-3514.

[13] KARPATHY A, JOHNSON J, LI Feifei. Visualizing and understanding recurrent networks[J]. arXiv preprint arXiv, 1506.02078, 2015.

[14] LIPTON Z C, BERKOWITZ J, ELKAN C. A critical review of recurrent neural networks for sequence learning[J]. arXiv preprint arXiv, 1506.00019, 2015.

[15] ZHANG Shu, ZHENG Dequan, HU Xinchen, et al. Bidirectional long short-term memory networks for relation classification[C]// Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. Bali, Indonesia: dblp, 2015: 73-78.

[16] MA Xuezhe, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF[J]. arXiv preprint arXiv, 1603.01354, 2016.

[17] HUANG Zhiheng, XU Wei, YU Kai. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv, 1508.01991, 2015.

[18] GUIA J, SOARES V G, BERNARDINO J. Graph databases: Neo4j analysis [C]// Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017). Portugal: Scitepress, 2017: 351-356.

[19] 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. New York: ACM, 2018: 1433-1445.

[20] FERNANDES D, BERNARDINO J. Graph databases comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB[C]//Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018). New York: ACM, 2018: 373-380.