文章编号：2095-2163(2020)10-0047-07 中图分类号：TP391 文献标志码：A

基于百科数据的林业知识图谱的构建与应用

胡宸恺，魏 鑫，姜国强，李发强，金玉舜
（东北林业大学 信息与计算机工程学院，哈尔滨 150040）

摘 要：本文提出了基于百科数据的林业知识图谱构建方式。首先，通过 Scrapy 爬虫框架爬取 Wikidata 和互动百科上与林业相关的数据，构建了林业知识图谱的概念层、数据层；其次，在命名实体识别和分类的步骤中提出了构建林业实体类别分类表的方法来进行实体分类；最后，将构建好的林业知识图谱存储到 Neo4j 图数据库中，并搭建了一个简易的知识图谱网站，使用 Django 框架和 ECharts 实现知识图谱的网页展示。

关键词：林业知识图谱；Scrapy 爬虫框架；Neo4j 图数据库

Construction and Application of Forestry Knowledge Graph Based on Encyclopedia Data

HU Chenkai, WEI Xin, JIANG Guoqiang, LI Faqiang, JIN Yushun
(College of Information and Computer Engineer, Northeast Forestry University, Harbin 150040, China)

【Abstract】 This paper proposes a method of constructing forestry knowledge graph based on encyclopedia data. First, through the Scrapy crawler framework, crawling Wikidata and HuDong Baike on forestry-related data, the concept layer and data layer of the forestry knowledge graph are constructed. Secondly, the method of constructing the forestry entity category classification table is proposed in the step of named entity recognition and classification to classify entities. Finally, the constructed forestry knowledge graph is stored in the Neo4j graph database, and a simple knowledge graph website is built. The Django framework and ECharts are used to realize the web display of the knowledge graph.

【Key words】 Forestry knowledge graph; Scrapy crawler framework; Neo4j graph database

0 引 言

目前，林业领域的用户大多仍都通过搜索引擎查询林业领域知识，会花费大量时间在庞大的互联网冗余数据里查找对自己有价值的信息，获取林业领域信息存在检索精准度和效率较低的问题。为了解决林业领域用户查找相应信息费时费力的问题，提升林业大数据信息化水平，构建林业知识图谱显得尤为重要。林业知识图谱是将互联网上稀疏、缺乏关联的林业数据组合起来，提高用户检索林业知识的效率。由于目前互联网上与林业领域相关的网站提供的林业数据陈旧、分散、不直观，构建林业知识图谱可以帮助林业领域专业人员更好地创建、完善并更新林业相关知识，同时也能更好地推进林业信息化发展。将林业知识图谱改造成林业知识科普网站还能更好地向大众宣传和普及林业相关知识。

目前国内外科研机构、大型互联网公司同时也在推出针对不同限定领域的知识图谱产品来提高对应领域的信息检索效率。其中针对林业领域，数据库资源最为专业的是由中国林科院科信所建成的中国林业信息网以及中国工程院建成的林业专业知识服务系统，但都是数据库资源形式，而没有增加知识图谱展现的形式。即在林业领域还尚未有专业性强、成熟度高的林业知识图谱，构建林业知识图谱在大数据时代对林业数据资源整合有着重要的作用。

1 相关工作

知识图谱可以比作是一个大型结构化的语义网络，由概念实体和语义关系组成。知识图谱最早由谷歌在 2012 年 5 月 17 日通过其官方博客正式提出。早在 2010 年，谷歌收购 Metaweb 公司用于知识图谱的信息收集和构建工作，其关键技术是将不同文字的表述与同一个实体链接起来，并找出实体之间属性的联系。其实在谷歌提出知识图谱的概念之前便已经有了如 DBPedia、Freebase、开放政府数据以及苹果公司推出的智能语音助手 Siri 等知识图谱技术的相关应用。在国内，学术界的知识图谱研究有上海交通大学的 zhishi.me 中文知识图谱平台，工

基金项目：2018 年度东北林业大学国家级大学生创新项目(201810225172)。

作者简介：胡宸恺(1998-)，男，本科生，主要研究方向：数据科学与大数据技术。

收稿日期：2020-05-11

业界已经应用了知识图谱技术的有百度的下一代搜索引擎雏形“百度知心”和搜狗知识库搜索引擎“知立方”等[1]。

2013 年段庆峰收集了近 24 年林业经济领域的发展趋势,从核心作者、核心机构和关键词、利用知识图谱的分析方法展现了林业经济研究中心(如东北林业大学经管学院)对于中国林业经济研究的动态[2];同年还有张哲等从 Web of Science、知网收集了有关森林碳汇研究的文献总计 5000 余篇,运用知识图谱的方式梳理了对该研究方向有重要贡献的机构、相关任务和研究热点等[3];在 2015 年,陈丽荣等使用 CiteSpace,收集知网中 1997—2015 年的3 081 篇天然林保护工程文献为研究对象,使用知识图谱分析方法揭示中国天然林保护工程研究的整体脉络[4];同年还有苏松锦等以黄山松为研究对象,分析两百余篇相关文献,揭示了围绕黄山松研究有重要影响力的作者和科研机构[5];2016 年苗润清以林业项目、林业专利数据,从年份、申请人、关键词 3 个方面使用知识图谱展现林业专利的申请趋势[6]。这些研究都只是针对所提及的研究对象,以知识图谱的数据可视化形式展现出来,还没有发现利用林业百科数据来构建知识图谱的这类研究。但是国内有许多学者对限定领域的知识图谱构建进行了研究,如针对中医药的知识图谱构建、微博的知识图谱构建、双语影视知识图谱的构建、宠物医学知识图谱的半自动构建方法等等。这些知识图谱的构建都是基于相关领域专业数据库资源还有爬取相关领域的百科内容,互联网上公开的海量领域数据和领域专业数据库资源给相应领域知识图谱的构建给予了很大的帮助。但对于数据资源较为稀缺(如新技术、基础科学突破发展方向)或者数据保密性强(如军工)的领域则不太容易构建限定领域知识图谱。

2 林业知识图谱的构建

2.1 知识图谱的定义

知识图谱有二种基本组成单元形式,一种为<“实体”-“关系”-“实体”>,另一种为<“实体”-“属性名称”-“属性值”>。知识图谱通过这两种形式来描述不同现实事物之间的关系与属性。构建知识图谱的数据分为 3 类:

(1)结构化数据,如关系数据库中存储的表;

(2)半结构化数据,如百科数据、JSON 文档、XML 文档等;

(3)非结构化数据,如图片、音频和视频等。

一般有二种方式存储知识图谱的数据,一种是通过 RDF(资源描述框架)存储,另一种方法就是通过数据库存储[7]。本文采用第二种方法并使用图数据库存储。虽然简单的知识图谱是可以通过关系数据库存储的,但是当一个知识图谱涉及关系和属性越复杂、实体节点数量越大的时候,图数据库在关联查询的效率会比传统的关系数据存储方式有显著地提高。图数据库中实体之间能通过关系相互连接,构建出知识图谱的网状结构。知识图谱的架构分为逻辑架构和技术架构。在逻辑上,知识图谱分为数据层和模式层。数据层存储真实的数据,模式层是知识图谱的核心,在数据层之上通过本体库管理知识图谱。

图 1 展示了一个以水曲柳为实例的简单知识图谱,可以解析出这样一个简单的知识图谱的模式层为:实体-关系-实体;实体-属性-属性值。而数据层的表示为:水曲柳-界-植物界;水曲柳-产地-黑龙江。水曲柳节点代表单个实体,周围蓝色圆点如黑龙江、植物界、翅果表示与水曲柳相关联的实体或属性值,不同实体之间的连线上有描述实体与实体间的关系与属性。因此,一个高质量、成熟的领域知识图谱是由大量实体节点互相链接,且实体之间包含多个关系与属性所组成的,每个实体通过设置唯一 ID 来与其它实体区分。

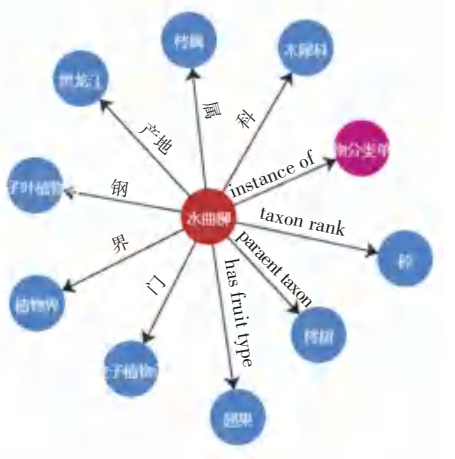

图 1 水曲柳的实体查询结果

Fig. 1 Entity query results of Fraxinus mandshurica

2.2 知识图谱构建技术

知识图谱的技术架构如图 2 所示。首先从网络上爬取数据,数据可能是结构化的、半结构化的以及非结构化的;其次,基于这些数据来构建知识图谱,通过一系列自动化或半自动化的技术从原始数据中提取知识,即提取实体关系;最后,存入所构建的知识库的模式层和数据层。

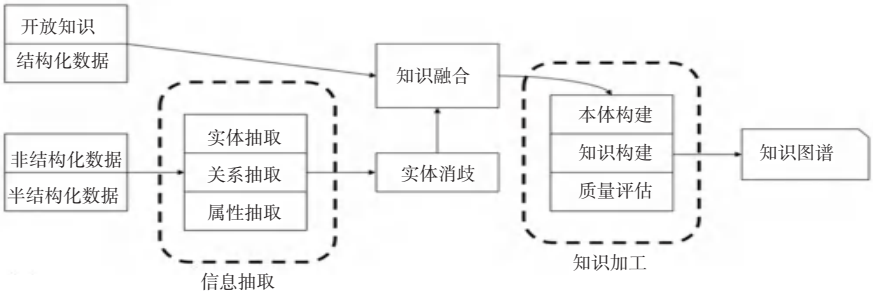

图 2　知识图谱技术架构

Fig. 2　Knowledge Graph Technology Architecture

以目前知识图谱构建技术的发展而言，构建知识图谱可以通过自顶向下或者自底向上的方法[8]。在知识图谱发展初期，多数的科研机构与企业对于知识图谱及其应用主要采用自顶向下的方法构建。自顶向下的构建方法是指爬取百科类网站（如本文中用到的互动百科和维基百科）的半结构化数据或使用开放的专业领域的结构化数据，从这些高质量的数据中提取本体和模式信息，加入到知识库中。自底向上的构建方法则是借助公开采集的数据，从中提取出资源模式，再选取其中置信度较高的模式，经过人工审核加入到知识库中。采用自底向上的方法必须要有 3 个步骤：

（1）信息抽取。从半结构化或非结构化的数据中抽取三元组的基本信息，如实体、关系以及实体属性等结构化信息，包括实体抽取、关系抽取和属性抽取 3 个关键技术。实体抽取是指从网页语料或具有某种特殊性质的文本语料中选取迭代多次后置信度高的实体作为结果输出，其是最为基础和关键，而且是最能影响知识获取效率和质量的部分；

（2）知识融合。在半结构化或非结构化数据中取得实体、实体间关系还有实体属性的信息，这些信息可能包含错误或者冗余的信息，经过初步筛选的数据也缺少逻辑和层次结构，最后清理无用的信息并整合有用的信息。这个部分包含二个内容，实体链接和知识合并，基本思想是消除实体间歧义后将知识库中对应的实体链接到一起，并验证和评估实体含义；

（3）知识加工。包含 3 个步骤，本体构建、知识推理和质量评估。其基本思想是将实体对象间的关系结构化处理。

随着知识抽取和知识加工技术的不断发展，知识图谱及其应用将逐渐采用自顶向下构建概念层、自底向上的方法构建数据层的方式。因为专业领域的知识更注重知识的深度、准确性和具体的层次结构，所以需要领域内丰富且具有准确性的数据来构建知识库。因此本文构建的林业知识图谱是将两种构建方法综合起来，使用爬取百科里的林业相关的数据，辅以 Wikidata 上实体和实体说明的开放数据，最后结合自定义林业实体分类表的模式去匹配采集的数据。

2.3　抽取林业领域百科数据

数据爬取采用开源的 Scrapy 分布式爬虫框架从 Wikidata、互动百科的网页中爬取林业相关的数据。

首先，从 Wikidata 爬取数据。Wikidata 是维基百科推出的一个大型知识库，由志愿者自主发布或导入数据，经过管理员的严格审核保障了提交的数据质量和可信度，该网站包含了大量的实体和实体间的关系。每一个实体为一个网页，使用“实体+说明”的数据模型展现数据融合结果。因此，通过 Scrapy 爬虫框架爬取 Wikidata 上面实体及其实体的说明，过程需要完成：

（1）爬取 Wikidata 定义的所有关系，如图 3 所示的生物学包含的实体关系定义；

（2）爬取 Wikidata 上的实体，如图 4 所示的“海棠”实体数据；

（3）以第一步爬取的实体数据为基础，爬取实体和实体间的三元组关系，爬取与实体相联系的其他实体与关系，如图 5 所示的是实体与实体间三元组关系。

还要在互动百科上爬取林业相关的数据，这构建林业知识图谱概念层的重要一步。首先要爬取互动百科的“植物”分类树，百科的分类树可作为林业知识图谱的概念层，如图 6 左下角所示。通过爬取的分类树列表再爬取“植物”分类树列表下的所有子词条。对于单个词条要分为主要信息和次要信

息,如图 7 展示的是“黄杨冬青”词条的主要信息:标题、开放分类和基本信息,主要信息用来实现信息抽取的步骤,如词条内容提供关于“黄杨冬青”实体的属性有:界门纲目科属种以及分布区域。在浏览林业相关词条的网页时可以发现,一部分数据在互动百科里拥有开放分类,而一部分数据没有开放分类,开放分类能帮助实体更准确地被识别。如图 8 展示的爬取“黄杨冬青”词条的次要信息,词条的次要信息作为冗余数据的补充,如形态特征、生长环境、分布范围,可以通过模板匹配的方式查询这些信息,提高信息检索的效率。

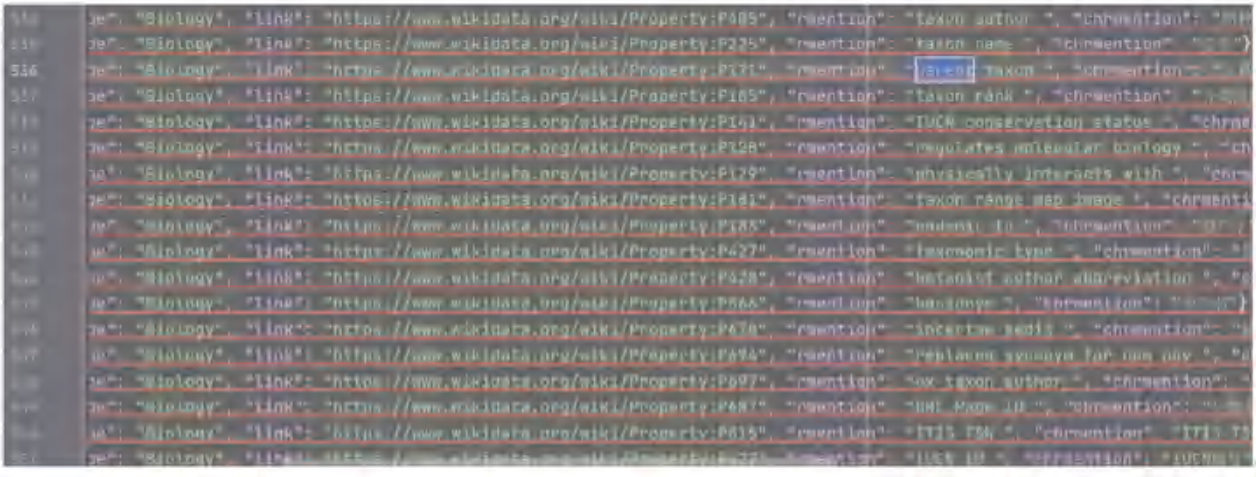

图 3 爬取 Wikidata 上关于生物学的实体关系定义

Fig. 3 Crawl the definition of entity relationship on biology onWikidata

图 4 爬取 Wikidata 上的海棠实体数据

Fig. 4 Crawling Begonia entity data onWikidata

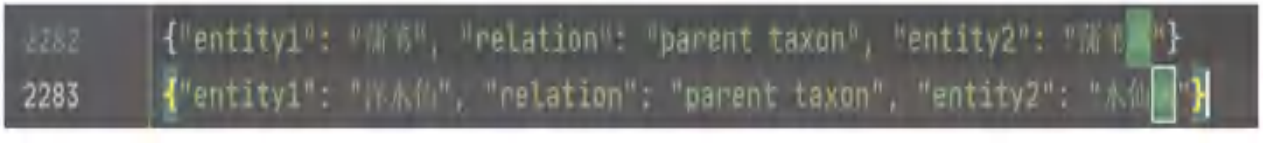

图 5 爬取 Wikidata 上的实体与实体间三元组关系

Fig. 5 Crawling the entity-to-entity triplet relationship onWikidata

2.4 命名实体识别和分类

本文采用了基于规则和词典的命名实体识别方式,首先要构建林业实体类别分类表,见表 1,然后通过下面 3 个步骤识别并分类实体。

图 6　互动百科“植物”分类树详情页面

Fig. 6　Details page of the interactive tree "Plants" classification tree

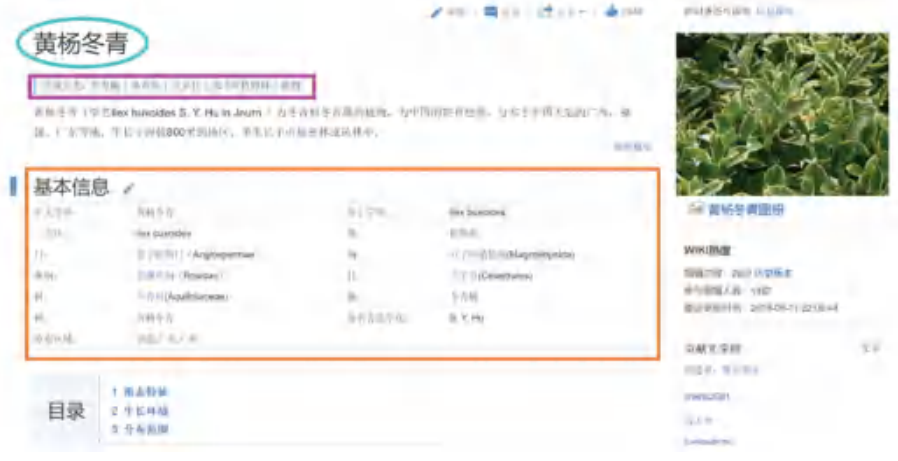

图 7　互动百科“黄杨冬青”词条页面主要抓取的信息

Fig. 7　Crawl the main information of theHuDong Baike "Ilex buxoides" entry page

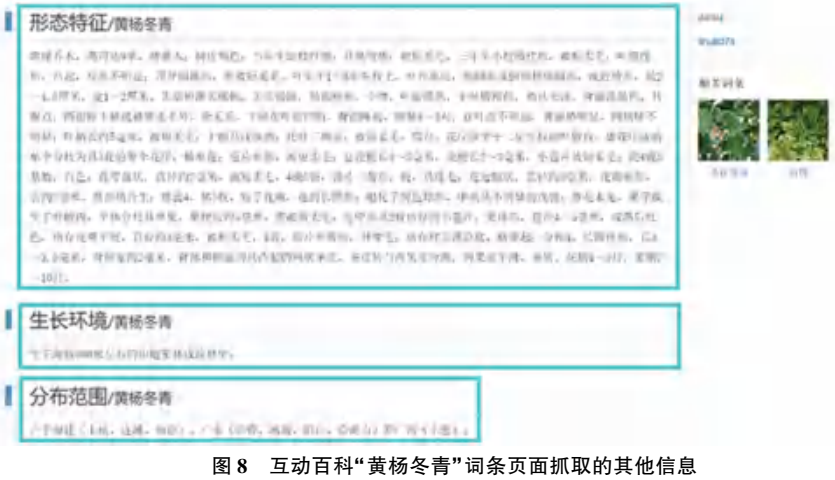

图 8　互动百科“黄杨冬青”词条页面抓取的其他信息

Fig. 8　Crawl the other information of theHuDong Baike "Ilex buxoides" entry page

表 1 林业实体类别分类表

Tab. 1 Classification table of forestry entities

序号	命名实体
1	人
2	地名
3	机构名
4	植物学名词
5	气候名词
6	植物疾病名词
7	自然灾害名词
8	与林业无关的名词实体
9	无关数据

(1)首先将语料分词,使用 THULAC 工具将爬取的林业百科数据进行中文词法分析,进行分词操作,见表 2。

表 2 THULAC 词性解释表

Tab. 2 THULAC part of speech interpretation table

THULAC 词性标记集(格式:词性标记/词性标记解释)				
n/名词	np/人名	ns/地名	ni/机构名	nz/其它专名
m/数词	q/量词	mq/数量词	t/时间词	f/方位词
s/处所词	v/动词	vm/能愿动词	vd/趋向动词	a/形容词
d/副词	h/前接成分	k/后接成分	i/习语	j/简称
r/代词	c/连词	p/介词	u/助词	y/语气助词
e/叹词	o/拟声词	g/语素	w/标点	x/其它

(2)通过林业实体类别表识别命名实体(人、地名、机构名),再从非命名实体中过滤掉与林业无关的名词实体;

(3)剩下的词或词组合,匹配知识库中已经分好类的实体,如果没有匹配到则自动划分到无关数据。

2.5 知识图谱的存储与表示

在完成知识图谱的构建工作后,所有的数据都被处理成结构化的数据存储。对于传统的关系型数据库而言,大规模的结点与边的存储和查询效率较低,而 Neo4j 作为图数据库的代表很好地解决了上述的问题。Neo4j 图数据库中的节点、关系和属性 3 种元素与知识图谱中知识的图模型相对应;支持分布式存储,更好地应对了大规模数据增长的问题;通过其使用的 Cypher 语句也能直观地看出图数据的操作和不同实体之间的关系;具有较高的扩展性和可靠性、支持完整的 ACID 事务。由此,本文所构建的林业知识图谱用 Neo4j 图数据库进行存储,方便用户获取林业知识结构。对知识图谱的查询可以转换为 Neo4j 图数据库上的查询,以查询实体“蓖麻子”的知识图谱查询实体流为例,如图 9 所示。

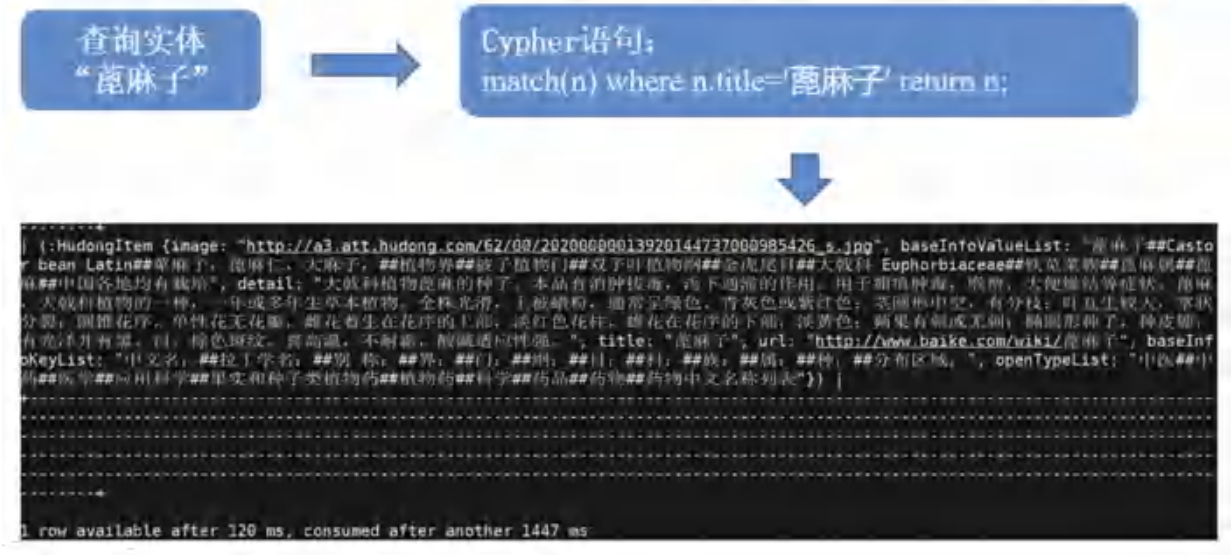

图 9 以“蓖麻子”为实例查询知识图谱的过程图

Fig. 9 The process diagram of querying the knowledge graph with "Castor Seed" as an example

3 林业知识图谱的应用

3.1 林业知识问答系统

知识问答系统主要分为二类:

(1)基于信息检索的问答系统。先将问题转化为对图数据库中知识库的查询,再从知识库中搜索与问题中实体相关信息相似度比较高的几个实体作为问题的备选答案,最后再按相似度降序排序的备选答案中选出问题的确切答案;

(2)基于语义分析的问答系统。对问题进行语义上的分析,其中包括词汇映射和语法树的构建,将问题转换为知识库的语义表示,再通过对知识库的推理查询从而找到正确答案。

本文搭建了一个简易的林业知识图谱网站,通过匹配问题中的实体和问题模板的形式转换成对 Neo4j 图数据库的 Cypher 查询语句进行答案检索。

3.2　林业知识问答系统架构设计

该问答系统解答一个提问需要经过 3 个模块的流程，如图 10 所示。

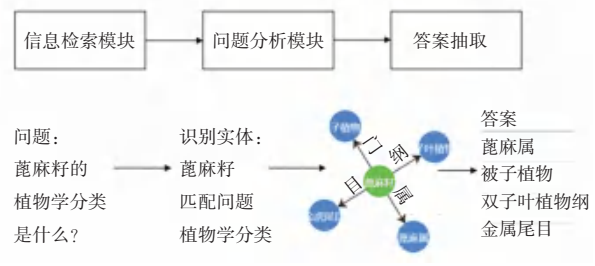

图 10　林业知识问答流程图

Fig. 10　Flowchart of forestry knowledge quiz

（1）问题分析模块负责对用户输入的问题进行分析，通过命名实体识别技术检测出问题中的实体词，将实体词反馈给知识查询模块；

（2）知识查询模块以实体词作为关键词在知识库中查询，得到的查询结果可能会有多个，再通过实体链接找到问题中实体词表示的实体对应哪一个之后，再重新查询知识库中该候选实体相关联属性的信息生成候选答案；

（3）答案抽取模块分析问题和候选答案的语义相似性，选取相似度最高的答案输出。

4　结束语

本文提出了一种基于百科数据的林业知识图谱的构建方式，并详细介绍了构建林业知识图谱需要的步骤，最后展示了林业知识图谱的实例，简单介绍和展示了林业知识图谱的应用：林业知识问答系统。首先，通过爬取互动百科分类树设计构建了林业知识图谱的概念层，然后通过 Scrapy 爬虫框架爬取 Wikidata 提供的“实体+说明”开放数据，分离出实体、实体和实体间的关系等与上一步爬取的互动百科数据一起构建林业知识图谱的数据层；其次，命名实体识别和分类步骤，利用基于规则和构建林业实体类别分类表的方式来对数据中的实体进行识别；最后，将整合好的知识存储在 Neo4j 图数据库里，并通过 Neo4j 内置的数据可视化方法显示林业知识图谱。

知识图谱相关的技术飞速发展，在搜索、推荐系统、报告生成、风险监控等业务都有较多的应用。未来每行每业都会产生更多的数据，越来越多的公司也会收集相关数据构建领域知识库。或许会发展以知识图谱管理信息的方式，为以后的“智能大脑”奠定基础。虽然目前处于知识图谱的探索阶段，但对于林业领域来说，创建、获取高质量的林业数据仍为构建林业知识图谱的重要基础环节。为了建立高质量的林业知识图谱，本文提出的构建方式仍有许多可以优化的地方，如数据来源单一，只能对大部分实体进行实体对齐的操作，少数可能无法进行实体对齐；对于数据清洗环节需要专业人员的介入，方便后期林业知识库的维护与扩展；优化已有的林业知识库建立检索效率更高的知识问答系统等。

参考文献

[1] 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.
[2] 段庆锋. 我国林业经济研究知识图谱分析[J]. 林业经济，2013(4)：115-119.
[3] 张哲，沈月琴，龙飞，等. 森林碳汇研究的知识图谱分析[J]. 浙江农林大学学报，2013，30(4)：567-577.
[4] 陈丽荣，曹玉昆，朱震锋，等. 中国天然林保护工程研究的知识图谱分析[J]. 农林经济管理学报，2015，14(6)：622-629.
[5] 苏松锦，刘金福，兰思仁，等. 黄山松研究综述(1960-2014)及其知识图谱分析[J]. 福建农林大学学报(自然科学版)，2015，44(5)：478-486.
[6] 苗润清. 林业科技项目管理系统研发及知识图谱分析[D]. 北京林业大学，2016.
[7] 郭琴芳. 基于知识图谱的初中数学在线学习系统及应用[D]. 西安理工大学，2019.
[8] 朱木易洁，鲍秉坤，徐常胜. 知识图谱发展与构建的研究进展[J]. 南京信息工程大学学报(自然科学版)，2017，9(6)：576-577.

（上接第 46 页）

4　结束语

本文借助大数据平台对大学生社交网络信息进行聚类分析，利用 k-means 算法对具有相似特质和兴趣爱好的大学生进行了分类，获得五类大学生群体，并对每一个群体所代表的兴趣爱好做了特征分析，实现了数据挖掘的效果。相关学校可以通过聚类结果，探究大学生群体用户的兴趣关注点，分析兴趣爱好信息，有效地支持大学生的个性分析、行为分析和心理分析。

参考文献

[1] KLASSEN S，KLASSEN C F. The Role of Interest in Learning Science through Stories[J]. Interchange，2014，45(3-4)：133-151.
[2] Do C X，Tsukai M. Exploring potential use of mobile phone data resource to analyze inter-regional travel patterns in Japan[M]// Data mining and big data. Cham：Springer，2017.
[3] 孙红卫，吕春燕，祁爱琴，等. 综合评价中数据标准化的原理研究[J]. 中国卫生统计，2015，32(2)：342-344，349.