doi:10.6041/j.issn.1000-1298.2023.03.029

基于知识图谱的花卉病虫害知识管理方法

陈　明[1,2]　朱珏樟[1]　席晓桃[1]

(1.上海海洋大学信息学院,上海 201306;2.农业农村部渔业信息重点实验室,上海 201306)

摘要:为解决花卉病虫害领域中病虫害防治因素关系复杂、知识冗余等问题,结合知识图谱对知识组织和管理的技术,提出一种基于知识图谱的花卉病虫害知识管理方法。首先,根据文献提取包括环境在内的花卉病虫害防治要素,构建花卉病虫害本体模型并存储在 RDF 图中,实现对知识规范性和完整性的控制;其次,对花卉病虫害领域文本进行分析,针对分析得到的文本特点,提出融合头尾实体分离"01"标注方法、轻量级双向转换编码表示模型(A lite BERT, ALBERT)和引入词性特征的级联标注模型(Cas_{POS}Rel)的抽取框架进行三元组抽取;之后利用自定义 RDF2PG 映射算法,按照 RDF 图中的本体模型将抽取到的三元组存入 Neo4j 数据库中,完成对花卉病虫害知识的存储及管理。实验结果证明提出的抽取框架中标注方法、预训练模型与抽取模型相比基线方法 F1 值分别提升 0.88、4.90、8.57 个百分点,同时得到抽取结果 F1 值为 95.07%。通过知识发现表明该知识管理方法能有效组织管理病虫害知识,帮助花卉种植人员进行更为有效的病虫害防治工作。

关键词:花卉;病虫害防治;知识图谱;知识抽取;知识管理

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1000-1298(2023)03-0291-10　　**OSID**:

Knowledge Management Method of Flower Diseases and Pests Based on Knowledge Graph

CHEN Ming[1,2]　ZHU Juezhang[1]　XI Xiaotao[1]

(1. *College of Information Technology*, *Shanghai Ocean University*, *Shanghai* 201306, *China*

2. *Key Laboratory of Fisheries Information*, *Ministry of Agriculture and Rural Affairs*, *Shanghai* 201306, *China*)

Abstract: In order to solve the problems of complex relationship of factors and mixed knowledge in the field of flower diseases and pests, combined with the knowledge organization and management technology of knowledge graph, a knowledge management method of flower diseases and pests based on knowledge graph was proposed. Firstly, according to the literatures, the flower diseases and pests control elements, including environment were extracted, the flower diseases and pests ontology model was constructed and stored in RDF to realize the control of knowledge standardization and integrity. Secondly, according to the text characteristics obtained from the analysis, the triple extraction framework was proposed which combined the "01" tagging method of head and tail entity separation, a lite bidirectional encoder representations from transformers (ALBERT) and cascade tagging model with part of speech features (Cas_{POS}Rel). Then using the custom RDF2PG mapping algorithm to complete the storage and management of flower diseases and pests knowledge. The experiments showed that the F1 value of the tagging methods, pretrained model and extraction model in proposed extraction framework was increased by 0.88, 4.90 and 8.57 percentage points compared with that of baseline methods, and the F1 value of the extraction result was 95.07%. The knowledge discovery showed that the knowledge management method effectively organized and managed the knowledge of flower diseases and pests, and helped the flower growers to carry out more effective pests control work.

Key words: flower; diseases and pests control; knowledge graph; knowledge extraction; knowledge management

收稿日期:2022-05-24　修回日期:2022-07-18

基金项目:上海市科技创新计划项目(20dz1203800)

作者简介:陈明(1966—),男,教授,博士,主要从事物联网技术、农业信息技术和知识图谱等研究,E-mail:mchen@shou.edu.cn

0 引言

花卉病虫害是导致花卉生产效率不高的重要原因,若不能及时处理花卉种植过程中发生的病虫害,则会严重影响花卉的种植收益[1]。随着信息技术的发展,各式各样的花卉病虫害防治知识大量分散在网络中,传统的关系型数据库知识管理方法无法对这些知识进行有效表示和存储,存在对异构数据无法融合、数据间关系无法高效表达、知识无法精炼等问题。知识图谱[2]是 Google 公司在 2012 年提出的一种通过语义网络来有效表达数据间关系的方法,利用知识图谱管理与传统知识管理方法不相适应的领域知识是当前各个领域的研究热点[3]。

在病虫害领域的知识图谱研究上,已有专家学者取得一些成果。一些学者集中在对领域本体的研究上,参考农业叙词表和病虫害相关文献,构建病虫害领域本体模型,通过本体模型解决实际问题[4-7];另一些学者集中在对领域数据的研究上,在对领域数据进行分析之后,采用知识抽取方法对数据中的实体和关系进行抽取,实现对知识的提炼[8-10];还有一些学者使用文献计量方法,通过关键词聚类构建知识图谱,对病虫害领域的研究进行可视化展示[11-13]。但是这些研究在病虫害知识的整理上都忽略了环境这一对病虫害防治至关重要的因素,在管理上还欠缺智能化和体系化,同时在知识图谱构建方法以及非结构化数据内容抽取方法的泛化性、准确性等方面还有很大提升。而在花卉病虫害领域的研究主要集中在对某一种、某一类花卉病虫害的防治策略、防治知识、发生原理等方面[14-15],缺少对花卉病虫害知识中实体关系的梳理、整合,遗漏了花卉病虫害实体间存在的关联关系导致花卉病虫害知识较为分散且存在冗余。

为了解决上述问题,本文以花卉病虫害为研究领域,提取包括环境在内的花卉病虫害防治要素,通过复用已有的病虫害知识体系构建花卉病虫害本体模型并使用 RDF 图进行存储。之后对花卉病虫害文献语料的分析,使用头尾实体分离“01”标注方法解决嵌套头尾实体的标注问题,利用 ALBERT 预训练模型[16]进行语义特征提取并提出将词性特征向量与级联标注模型(CasRel)[17]相结合的 $Cas_{POS}Rel$ 模型共同组成抽取框架,通过构建并训练关系标注器建立头尾实体映射的方法,完成对大量花卉病虫害文本中三元组的联合抽取。最后根据所构建的本体模型,采用自定义 RDF2PG 映射方法,按照 RDF 图中的本体结构在 Neo4j 图数据库中存储抽取到的三元组,完成对花卉病虫害知识的存储与管理,同时应用知识发现得出各种类花卉最易发病环境,从而起到预防病虫害的作用。

1 花卉病虫害知识管理框架

花卉病虫害知识管理框架如图 1 所示。将知识管理框架分 5 部分:① 本体建模。构建花卉病虫害领域本体,包括对概念类别定义、属性定义和关系定义。②数据获取。网络百科以及相关文献中的花卉病虫害文本是目前花卉病虫害知识的主要来源。③数据预处理。因为获取到的原始数据不能直接对其进行知识抽取,所以需要对数据进行初步处理、清洗、分类及内容标注。④知识抽取。根据历史经验及专家意见所构建的本体模型,对获取到的文本进行知识抽取,实现对知识的精炼和简化。⑤知识存储与管理。对抽取得到的知识进行组织、存储和管理,发掘知识之间存在的隐藏关系,从而实现更高效率的病虫害防治。

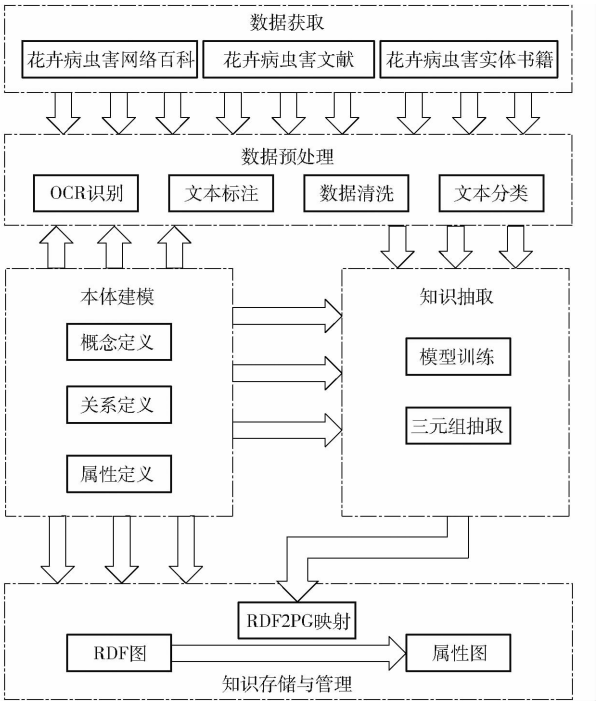

图 1 花卉病虫害知识管理框架

Fig. 1 Knowledge manangement framework of flower diseases and pests

1.1 花卉病虫害知识本体建模

知识体系又称为本体,其中包含了对数据的定义和描述。目前已有很多农业领域本体模型[18-20],但在花卉病虫害领域还尚无较为成熟的本体模型。考虑到建模的效率和质量,本文根据当前已有的病虫害知识体系[8,21]以及百度百科中关于病虫害的知识体系进行体系复用,并结合花卉病虫害防治先验知识进行自顶向下方式建模。按照与花卉病虫害关系的紧密程度,本文从文献[14-15,22-24]中提

取了表 1 中 10 类要素作为主要概念，包括花卉、花卉生长阶段、植物器官、地区、病虫害、化肥农药、防治方法、植物性状、环境、病原等。相比以往的病虫害本体额外考虑到了环境对花卉病虫害防治的重要影响，利用环境知识对病虫害进行主动防治是更为有效的防治措施。

表 1　花卉病虫害涉及要素

Tab. 1　Factors involved in flower diseases and pests

要素名称	相关性描述
花卉	病虫害所为害的花卉
花卉生长阶段	病虫害较易发生的时期
植物器官	病虫害所侵害的部位
地区	病虫害易发生地理位置
病虫害	花卉发生的病虫害
化肥农药	防治病虫害所需的化肥农药
防治方法	防治病虫害所使用的方法
植物性状	病虫害侵害花卉之后的表型特征
环境	利于病虫害发生的环境条件
病原	引起花卉病害的微生物

图 2 为本文构建的花卉病虫害领域内概念之间的关系，其中子类使用 subClassOf 表示，例如（病害，rdfs:subClassOf，病虫害）三元组表示病害为病虫害的子类。领域内属性有数据属性和对象属性两种属性，分别用 DatatypeProperty 和 ObjectProperty 表示，其中 DatatypeProperty 表示数值型属性的对象和类型，例如（病害，病害名称，xsd:string），表示病害的属性病害名称的定义域为“病害”类，值域为 string 类型；ObjectProperty 表示类与类之间的关系属性，例如（病虫害，为害部位，植物器官）表示关系“为害部位”的定义域和值域分别为“病虫害”类和“植物器官”类。上述定义的关系及属性均为对实例数据的约束，对实例数据起到规范作用。

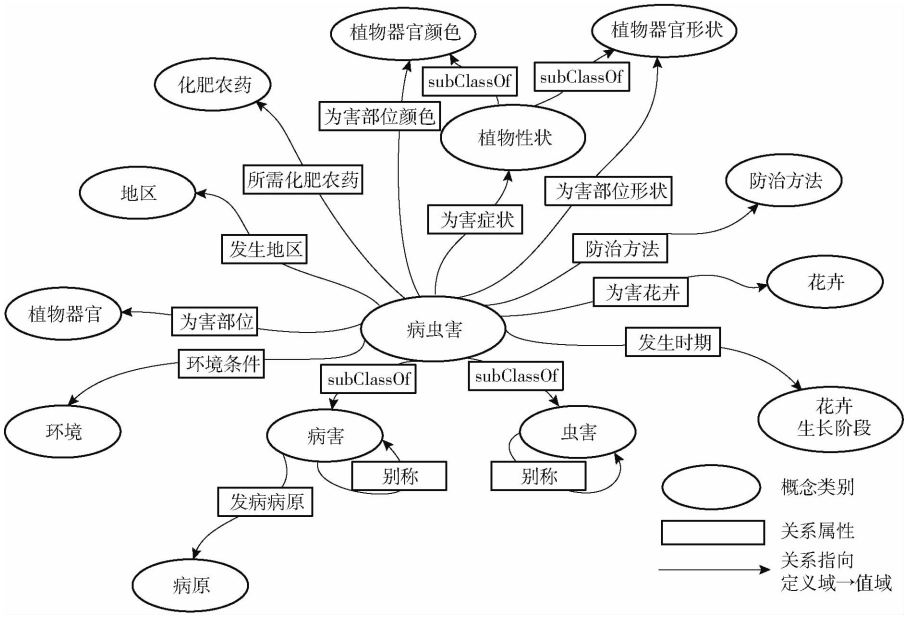

图 2　花卉病虫害概念关系

Fig. 2　Classes relation of flower diseases and pests

1.2　数据获取

当前花卉病虫害相关的知识大多以半结构化或非结构化的形式存在于网络百科文献中以及实体书中。根据查阅收集有关花卉病虫害的文献资料，本文从文献[22－24]及百度百科中整理了 745 个文档、160 余种花卉以及 170 余种害虫所导致的病害症状、病害发生环境条件以及病害防治方法等文档。以文献[22]为例，文献中内容如表 2 所示。

表 2　《花卉病虫害防治》中内容

Tab. 2　Contents in book “Flower Diseases and Pests Control”

病虫害名称	病虫害部分描述
香石竹枯萎病	枯萎病是香石竹较严重的病害之一，分布较广，在香石竹的栽培区都有发生。常引起植株枯萎死亡，降低切花产量
栀子花叶斑病	栀子花叶斑病由栀子叶点霉和栀子生叶点霉（真菌）侵染所致。病菌在病落叶或病叶上越冬。一般植株过密，通风、透光不良，盆花浇水不当，植株生长势弱时易发病

1.3　数据预处理

由于原始数据多为实体书籍或者 PDF 文档，无法进行进一步知识抽取，另外因为缺乏高质量的花卉病虫害语料库以及公开的标注模型，无法对原数据进行三元组标注，所以在进行模型训练时存在冷启动的问题。针对这个问题，JIANG 等[25]使用了符号主义方法，利用符号主义相比机器学习模型有可解释性强、细粒度可操控、不需要训练等优势，较好地解决了在大量未标注样本情况下的模型训练问题。本文先使用 OCR 文字识别的方式，利用 OCR

工具将原始内容转换为可读写的文档;之后根据定义好的本体结构,采用符号主义中的正则表达式辅助人工标注的方式,通过定义正则表达式对文本中格式较为规整的数据进行标注,例如“别称”可用“(又称|也称|还称|别名|别称|又叫|还叫|也叫|俗称|故有|又名)([\u4e00-\u9fa5]+)(等|之称|)(,|。)”进行表示,“发生地区”可用“在(.*?)等地”表示。但因为规则的覆盖率有限,不能完全覆盖所有句式,所以性状、环境等表述较为灵活的实体则用人工加以标注,同时将标注的实体加入实体字典来作为补充内容,以防规则以及人工标注时出现遗漏。考虑到模型对单句长度的限制和为了使模型能够得到更好的训练,通过分析文本,发现症状描述文本中常常出现多个植物器官,并且在植物器官后会伴随对该器官性状的描述。所以按照各个植物器官所在的文本中的位置,将复杂的症状描述文本进行细分来帮助模型更好地学习语义信息,其他文本内容则根据分号、句号进行分割。

1.4 知识抽取

面向文本的抽取是非结构化数据的知识抽取。通过对语料的分析,发现花卉病虫害的描述语料中存在大量头尾实体嵌套三元组以及重叠三元组的情况,特别对于高频出现的病害名称,例如三元组(栀子花叶斑病,为害花卉,栀子花)中头实体“栀子花叶斑病”包含尾实体“栀子花”,并且单个字的实体出现频率较高,例如叶、花等植物器官实体。

1.4.1 标注策略

根据上述对语料特征的分析,本文采用一种头尾实体分离标注加上“01”标注的方法。

(1)头尾实体分离标注方式。将标注序列数组分为头实体序列数组和尾实体序列数组,相较于传统的单序列数组标注头尾实体的方式[26],将头尾实体分为 2 个独立数组单独标注,解决了嵌套头尾实体、重叠头尾实体标注的问题。

(2)“01”标注方式。用两列数组,实体起始数组和实体结束数组分别来表示实体的起始位置和结束位置。首先对于输入的文本,创建 2 个数组长度与输入文本长度相同且元素全为“0”的数组进行初始化,再根据预先标注的实体内容,分别将实体的头尾位置在对应数组中分别标为“1”,当一个句子中有多个可能的实体时,依据就近原则,起始数组中的“1”到最近的结束数组的“1”之间的部分看成是一个实体,相比传统“BIO”标注方式[26],“01”标注方法只需进行二分类的标签预测,不需要预测多个标签类别,降低了预测的难度。另外“01”标注方法只对实体的头尾边界位置进行标注,降低了预测时出现实体错误或缺失的概率,同时对于单个字实体,可以进行较好的表示,不用引入额外的标注符号进行单独标注来进一步增加预测的标签类别,增加预测难度。以栀子花叶斑病为例,标注策略如图 3 所示。

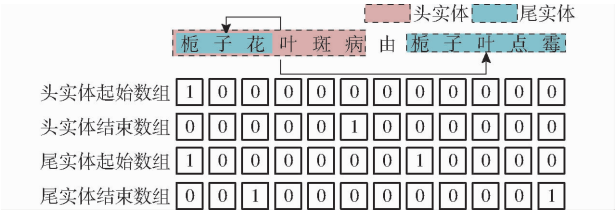

图 3 头尾实体分离“01”标注策略

Fig. 3 Separation of subject and object and “01” tagging strategy

1.4.2 $Cas_{POS}Rel$ 模型

在头尾实体分离“01”标注方法的基础上,本文参考同时进行命名实体识别和关系抽取任务的联合抽取方法[27-28]提出 $Cas_{POS}Rel$ 模型对三元组进行联合抽取,将常被忽视的词性特征向量融入级联标注模型,增强语义特征表示[29]。传统的抽取方法[30-31]大多都将关系看作是离散的标签,通过对抽取出的实体对进行关系分类从而实现三元组的抽取,这样的抽取方法在本文所用的花卉病虫害领域语料中可能会出现 3 个问题:①将命名实体识别任务和关系抽取任务独立进行会导致信息丢失、错误传递等问题。②类别分布不均匀的问题,小样本的类别分类效果较差。③多三元组重叠的问题,当一个实体有多种关系时,分类器可能会发生紊乱。

针对上述问题,$Cas_{POS}Rel$ 模型将关系看作是尾实体的标注器,对于一个头实体,使用所有预先定义好需要抽取的关系,通过标注器映射,找到和其对应的每个关系的尾实体,对于某一关系下未找到的尾实体,则认为头实体在该关系下的尾实体为“null”。通过这种映射方式,可以最大程度地找到所有与头实体有关的关系及其尾实体,解决了上述 3 个可能会出现的问题。依照上述本体模型中的关系三元组,发现不同关系的头实体大多词性相同,而尾实体词性具有明显区别,如表 3 所示,所以仅在尾实体标注时引入词性特征向量,加强语义表示。中文中的字和词都包含完整语义,所以本文采取字词融合的方式,将句子转换为 $\hat{\boldsymbol{v}}_c=(\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n)$ 和 $\hat{\boldsymbol{v}}_{POS}=(\boldsymbol{pos}_1,\boldsymbol{pos}_2,\cdots,\boldsymbol{pos}_m)$,其中 $\hat{\boldsymbol{v}}_c$ 表示字符向量,n 为字符在句子中的位置,$\hat{\boldsymbol{v}}_{POS}$ 表示词性向量,$\boldsymbol{pos}_m$ 为句子中的第 m 个词。

其中,对于输入的句子中每个字符向量 $\boldsymbol{c}_i$,可能为头实体起始和结束位置的可能性 $p_i^{\text{start}_{\text{sub}}}$ 和 $p_i^{\text{end}_{\text{sub}}}$ 计算式为

表 3　部分三元组尾实体词性

Tab. 3　Part of speech of partial triple tail entity

关系类型	尾实体词性
为害部位颜色	形容词(a)
为害部位形状	形容词(a)
发生时期	时间词(t)
所需化肥农药	专有名词(nz)

$$p_i^{\text{start}_{\text{sub}}} = \sigma(\boldsymbol{W}_{\text{start}}\boldsymbol{c}_i + \boldsymbol{b}_{\text{start}}) \tag{1}$$

$$p_i^{\text{end}_{\text{sub}}} = \sigma(\boldsymbol{W}_{\text{end}}\boldsymbol{c}_i + \boldsymbol{b}_{\text{end}}) \tag{2}$$

式中　σ——sigmoid 激活函数

$\boldsymbol{W}_{\text{start}}$、$\boldsymbol{W}_{\text{end}}$——可训练权重

$\boldsymbol{b}_{\text{start}}$、$\boldsymbol{b}_{\text{end}}$——可训练偏置

对每个头实体建立起和每一个关系特定标注器的映射，尾实体所在起始和结束位置的可能性 $p_{i,r}^{\text{start}_{\text{obj}}}$ 和 $p_{i,r}^{\text{end}_{\text{obj}}}$ 计算式为

$$p_{i,r}^{\text{start}_{\text{obj}}} = \sigma(\boldsymbol{W}_{\text{start}}(\boldsymbol{c}_i + \boldsymbol{sub}_k + \boldsymbol{pos}_i) + \boldsymbol{b}_{\text{start}}) \tag{3}$$

$$p_{i,r}^{\text{end}_{\text{obj}}} = \sigma(\boldsymbol{W}_{\text{end}}(\boldsymbol{c}_i + \boldsymbol{sub}_k + \boldsymbol{pos}_i) + \boldsymbol{b}_{\text{end}}) \tag{4}$$

式中　r——关系类型

$\boldsymbol{sub}_k$——第 k 个头实体的字符向量

$\boldsymbol{pos}_i$——第 i 个字符所在词的词性向量

通过以字符作为语义单元结合词性特征的方式，将字词特征相加来对字词特征进行融合，得到字、词性混合特征向量[29]。最后根据设置的激活阈值判定每个字符对应的标签。

1.4.3　知识抽取框架

本文使用头尾实体“01”标注的方式，先标记出单一句子中所有可能为头实体的实体，之后使用 ALBERT 预训练模型作为编码层，通过 ALBERT 预训练模型提取语料中的文本特征。在命名实体识别和关系抽取等自然语言处理(Natural language processing, NLP)的任务中，预训练模型对任务的结果有直接的影响。BERT 预训练模型是目前较为主流的预训练模型，在知识抽取方面表现较为良好[9,32]。ALBERT 预训练模型是在 BERT 预训练模型基础上，通过设计减少参数的方法来减小内存开销，加快训练速度的一种用于 NLP 预训练的无监督模型，相较于传统的预训练模型 Word2Vec[33] 等，通过使用掩码语言模型(Masked language model, MLM)和句子序列预测(Sentence order prediction, SOP)进行训练，使得 ALBERT 预训练模型不仅能够学习到词与词之间的词级上下文信息，还能掌握句子与句子之间的句子级别特征。之后采用由头实体标注器和各个关系特定的尾实体标注器构成的 Cas_{POS}Rel 模型完成提取。模型抽取框架如图 4 所示，通过 ALBERT 预训练模型对语句进行特征提取，得到具有丰富语义信息的序列向量，再将得到的序列向量作为输入，经过头实体标注器，计算最有可能为头实体的边界，起始和结束位置分别在起始数组和结束数组中用“1”表示，得到头实体“栀子花叶斑病”。之后将标注的头实体和包含句子信息的序列向量一同输入到各个关系特定的标注器中进行尾实体标注。根据输入头实体“栀子花叶斑病”，通过 jieba 分词工具进行词性标注并嵌入词性向量，将头实体字符向量和包含句子信息的字符向量进行融合，得到第 i 个字符的向量为

$$\bar{\boldsymbol{c}}_i = \boldsymbol{c}_i + \boldsymbol{pos}_i + \boldsymbol{sub}_1 \tag{5}$$

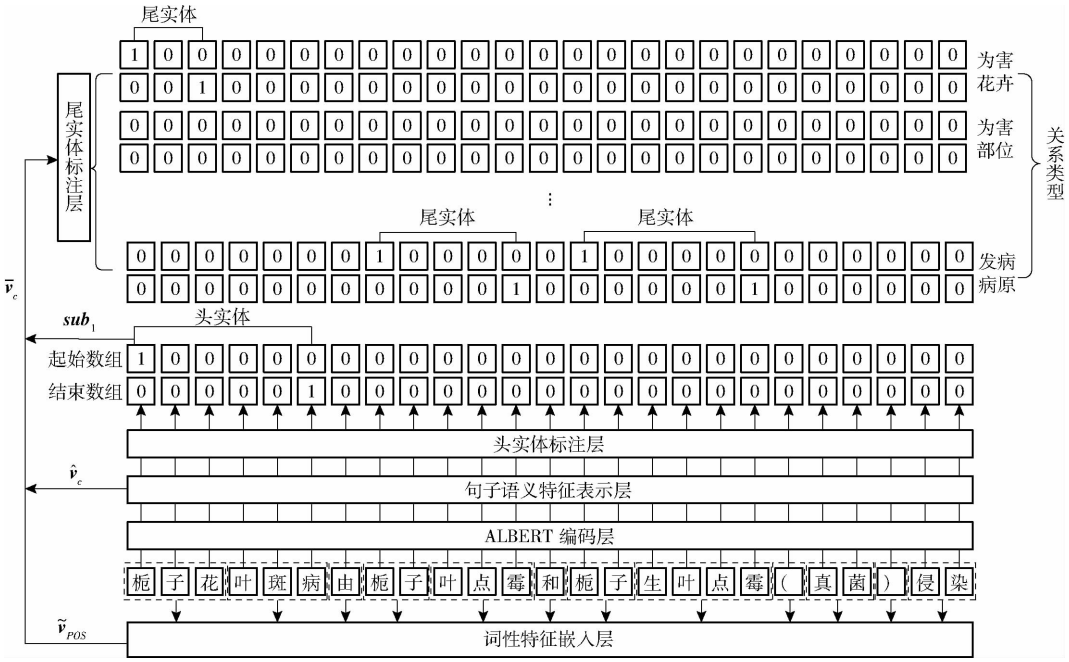

图 4　知识抽取框架结构

Fig. 4　Knowledge extraction framework

式中 $\boldsymbol{c}_i$——第 i 个字符的 ALBERT 编码字符向量

输入融合后的特征向量 $\overline{\boldsymbol{v}}_c=(\overline{\boldsymbol{c}}_1,\overline{\boldsymbol{c}}_2,\cdots,\overline{\boldsymbol{c}}_n)$ 到各个关系特定的标注器中进行尾实体标注。根据输入头实体“栀子花叶斑病”,在表示“为害花卉”关系的关系标注器得到尾实体“栀子花”,在表示“发病病原”关系的关系标注器得到尾实体“栀子叶点霉”和“栀子生叶点霉”,输出(栀子花叶斑病,为害花卉,栀子花)、(栀子花叶斑病,发病病原,栀子叶点霉)和(栀子花叶斑病,发病病原,栀子生叶点霉)3 个三元组,完成三元组抽取。

2 实验

2.1 实验评价指标

采用精确率、召回率和 F1 值来评估模型抽取效果,只有当抽取出的三元组与标注的三元组完全相同时才判别为正确。

2.2 实验环境与参数设置

本文使用实验设备为:Intel(R) Core(TM) i7-10875H CPU @ 2.30 GHz;内存 16 GB;GPU:NVIDIA GeForce RTX 2060 (6 GB)。运行环境为 Python 3.7,torch 1.9.1 + cu102。在设置模型参数时,考虑到运行环境,文本长度等因素以及多次实验,选择 Cas_{POS}Rel 模型核心参数设置如表 4 所示。

表 4 Cas_{POS}Rel 模型参数设置

Tab. 4 Cas_{POS}Rel parameters setting

参数	数值
批量大小	4
最大文本长度	256
迭代次数	150
学习率	10^{-4}
激活阈值	0.5

2.3 对比实验

本文经过整理以及预处理之后用于模型训练的文本总共 7 424 条,包含重叠实体语料 6 662 条,不重复三元组 15 933 个,按照比例 6∶2∶2 分别划分训练集、验证集以及测试集。依据本体结构对 10 类关系进行三元组抽取,各关系数量如图 5 所示。

为了证明本文抽取方法的有效性,本文分别使用 Word2Vec 预训练模型、头尾实体分离“BIO”标注方式、双向长短期记忆网络(BiLSTM)[34] 以及 ALBERT 流水线抽取模型作为基线方法,取在验证集最好效果,在测试集上进行对比实验。对比结果如表 5 所示。

(1)预训练模型对比。“BIO”标注方法下使用 Word2Vec 预训练模型和 ALBERT 预训练模型的 BiLSTM 模型的抽取结果表明,ALBERT 预训练模型

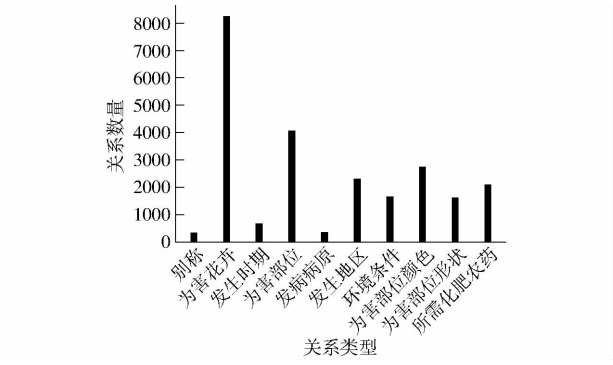

图 5 各个关系类别数量

Fig. 5 Quantity of each relationship category

表 5 不同抽取方法在测试集中的三元组抽取结果

Tab. 5 Performance of different extraction methods in test set %

模型	精确率	召回率	F1 值
BIO + Word2Vec + BiLSTM	83.82	88.08	85.90
BIO + ALBERT + BiLSTM	90.10	91.51	90.80
01 + ALBERT + BiLSTM	94.16	89.32	91.68
01 + ALBERT	89.60	83.62	86.50
01 + ALBERT + CasRel	94.62	92.91	93.76
01 + ALBERT + Cas_{POS}Rel	95.33	94.82	95.07

对于文本的语义特征提取和表示相比 Word2Vec 预训练模型要更加优越,F1 值提升 4.90 个百分点。

(2)标注方法对比。通过对比“BIO”标注方法下 BiLSTM 模型和“01”标注方法下 BiLSTM 模型的抽取结果可以发现,“01”标注方法只需进行二分类的标签预测,不需要预测多类别标签,对于实体的标注,相比传统的“BIO”标注方法在边界界定上更加准确,因为“01”标注方法只需要标注实体的头尾位置即可,而“BIO”方法则需要将实体的每个部分都标注出来,所以在进行标注时更容易导致实体内容的缺失和溢出,特别是对于由多个词组合而成的组合长实体,可能会出现预测实体拆分的问题。结果显示“01”标注的 BiLSTM 模型对三元组的抽取效果相比“BIO”标注方法的 BiLSTM 模型在测试集上 F1 值提高 0.88 个百分点,证明了“01”标注方法的优越性。

(3)抽取模型对比。通过对比“01”标注方法下 Cas_{POS}Rel 模型和“01”标注方法下 ALBERT 流水线模型的整体抽取结果如表 5 所示,各个三元组类别抽取效果如图 6 所示,可以发现 Cas_{POS}Rel 模型整体抽取结果在各个指标上均优于 ALBERT 流水线模型,F1 值提升 8.57 个百分点,同时在各个关系类别上的抽取效果较为均衡且均优于 ALBERT 流水线模型的抽取效果。由表 6 展现的对不同数量重叠三元组的抽取效果可以看出,与 Cas_{POS}Rel 模型采用的通过构建关系标注器来形成头尾实体映射的联合抽

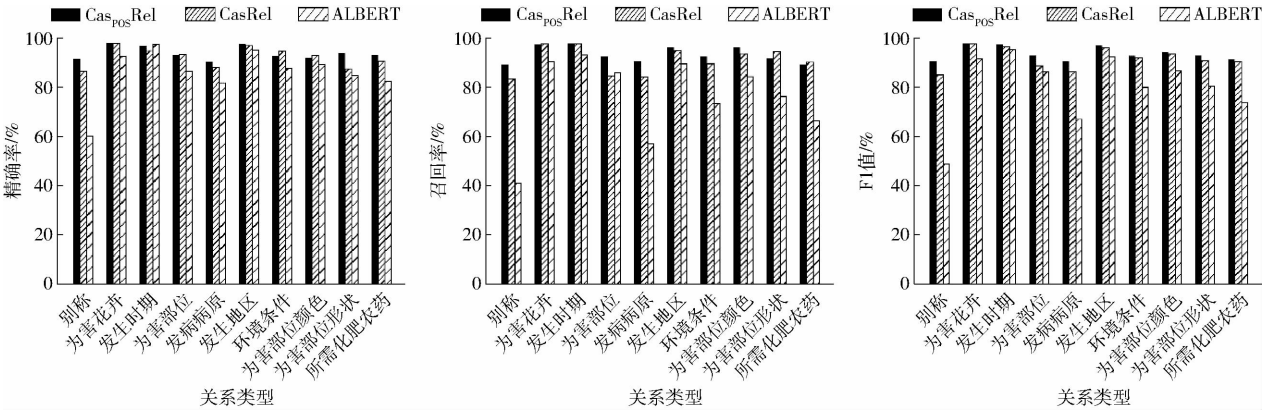

图 6　不同模型对各关系抽取效果对比

Fig. 6　Extraction performance of different models in different relationship categories

表 6　不同模型对不同数量重叠三元组的抽取效果

Tab. 6　Extraction performance of different models in different numbers of overlapping triples　%

重叠三元组数量	模型	精确率	召回率	F1 值
2	ALBERT	90.11	89.14	89.62
	Cas$_{POS}$Rel	94.45	96.02	95.23
3	ALBERT	91.45	85.05	88.13
	Cas$_{POS}$Rel	96.04	93.98	95.00
4	ALBERT	90.67	83.06	86.70
	Cas$_{POS}$Rel	95.43	94.35	94.89
≥5	ALBERT	89.56	79.99	84.50
	Cas$_{POS}$Rel	96.35	94.84	95.59

取方法相比 ALBERT 流水线模型中采用的关系分类方法在对多重叠三元组的抽取上更具优势。这种方法会对所有头实体，每个关系标注器进行尾实体的映射，尽可能地找出语句中可能存在的所有三元组，更不容易造成三元组缺失和遗漏的问题。另外 Cas$_{POS}$Rel 在 CasRel 模型的基础上引入词性特征向量之后，相比原模型在 3 个评价指标上分别提升 0.71、1.91、1.31 个百分点，同时在各关系类别的抽取效果都有不同程度的提升，但在“环境条件”关系类别中抽取效果提升的幅度较小。经过对抽取结果和原标注语料进行分析，发现有关环境条件的描述较为灵活多变，词性特征较不明显，从而导致抽取效果提升不大。

3　知识存储、管理及发现

RDF 图和属性图（Property graph，PG）是目前用于知识存储与管理的主要方式。本文将 RDF 图和属性图相结合，使用 RDF 图存储本体结构，属性图存储实例数据，通过将两种方式进行互补，来保证知识本体结构的完备和实例数据存储更新的要求。

RDF 图与属性图的映射是将 RDF 图中的概念映射成为属性图中的节点类型，将 RDF 图中的属性定义转换为属性图中节点的键值对属性，将 RDF 图中的概念间关系转换为节点之间的边，映射关系如图 7 所示。

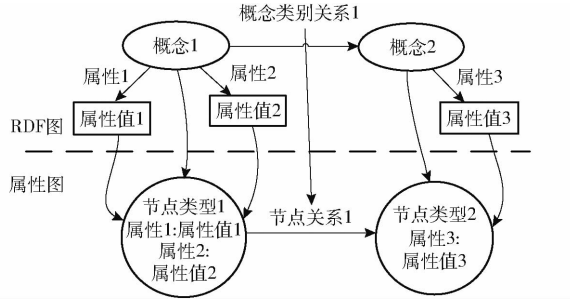

图 7　RDF 图与属性图的映射关系

Fig. 7　Mapping relationship between RDF and property graph

为了保证知识的时效性和基于知识图谱进行的知识发现的有效性，需要对知识图谱进行及时的更新以及存储细粒度上的控制，所以本文自定义了将抽取到的三元组根据 RDF 图中存储的本体结构直接存储到属性图中的 RDF2PG 映射方法，使用 Jena API 对本体文件进行读取和推理，使用 Neo4j 作为属性图存储工具，具体算法流程如图 8 所示。

最终使用自定义映射算法在 Neo4j 中存储三元组 15 855 个。根据上述对花卉病虫害知识的管理，可以实现花卉病虫害知识可视化以及知识发现，例如可以通过管理的知识发现某类花卉易病环境及泛用化肥农药。对某类花卉病虫害所共同连接的环境节点和化肥农药节点入度进行排序以及显示，得到较易发病环境和泛用化肥农药从而进行更为集中有效的病虫害防治。以月季为例，可以发现月季容易在多雨、潮湿、通风不良等环境下发生病害，所以避免此类环境的发生可以预防大多数月季病害，多菌灵、波尔多液、百菌清等化肥农药可以防治大多数月季病虫害，图 9 展示了月季病虫害知识的部分可视化内容。

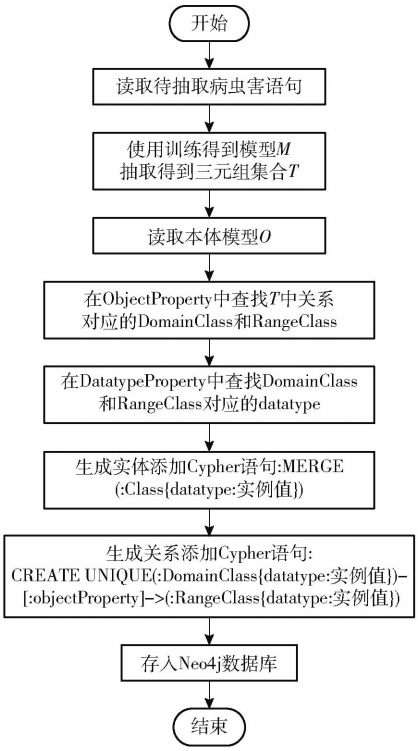

图 8 自定义 RDF2PG 映射算法流程图

Fig. 8 Custom RDF2PG mapping algorithm progress

4 结论

(1)提出了一种基于知识图谱的花卉病虫害知识管理方法。通过研究领域相关文献,构建了能够满足基本花卉病虫害防治要求的本体模型,相比其他目前已有的本体,额外考虑到了环境对花卉病虫害防治的影响,环境因素对于花卉病虫害的防治上不仅着重于治疗,更注重于防范,及时防范病虫害能够进一步减小花卉的损害。

(2)在非结构化数据的抽取中,面对未标注的低质量数据采用基于符号主义的正则表达式进行标注,根据对语料的分析,采用了头尾实体分离“01”标注方法解决了传统单数组序列标注无法应对的嵌套头尾实体标注问题,并用 ALBERT 预训练模型和 Cas_{POS}Rel 抽取模型在引入词性特征加强了语义表示的同时解决了重叠三元组的抽取问题,通过实验得到 F1 值为 95.07%,证明了该方法在本领域中抽取三元组的有效性和实用性。

(3)采用 RDF 图存储本体结构,根据自定义 RDF2PG 映射方法,将抽取出的三元组不经过其他存储方式,直接按照本体模型的结构存储进 Neo4j 数据库,规范了所管理的知识,提高了存储效率及自动构建图谱能力。

图 9 月季病虫害知识部分可视化内容

Fig. 9 Visual content of knowledge graph of Chinese rose diseases and pests

(4)使用知识抽取技术以及结合 RDF 图和属性图的存储方式,虽然能够对花卉病虫害知识进行较为有效的管理,但是病虫害防治是花卉种植流程中的一部分,不把花卉种植与病虫害防治相融合会与实际情况不相适应。所以后续可将本文知识管理方法同样用于花卉种植流程,将种植管理与病虫害防治相结合,能够更好地为花卉生产者提供帮助,增加生产效率。

参 考 文 献

[1] 冯艳霞. 园林花卉常见病虫害综合防治技术探究[J]. 广东蚕业, 2022, 56(2):88 - 90.

[2] AMIT S. Introducing the knowledge graph: things, not strings[EB/OL]. [2012 - 05 - 16]. https://blog.google/products/search/introducing-knowledge-graph-things-not/.

[3] 王萌,王昊奋,李博涵,等. 新一代知识图谱关键技术综述[J]. 计算机研究与发展,2022, 59(9): 1947 - 1965.
WANG Meng, WANG Haofen, LI Bohan, et al. Survey on key technologies of new generation knowledge graph[J]. Journal of Computer Research and Development, 2022, 59(9): 1947 - 1965. (in Chinese)

[4] 李悦. 农作物病虫害知识图谱构建研究[D]. 北京:中国农业科学院, 2021.
LI Yue. Research on the construction of knowledge graph of crop diseases and pests[D]. Beijing: Chinese Academy of Agricultural Sciences, 2021. (in Chinese)

[5] CHAUDHARY S, BHISE M, BANERJEE A, et al. Agro advisorty system for cotton crop[C]//7th International Conference on Data Engineering and Communication Technology, 2017:665 - 674.

[6] 郑颖, 金松林, 张自阳, 等. 基于本体的小麦病虫害问答系统构建与实现[J]. 河南农业科学, 2016, 45(6):143 - 146.
ZHENG Ying, JIN Songlin, ZHANG Ziyang, et al. Construction of question answering system related to wheat diseases and insect pests based on ontology[J]. Journal of Henan Agricultural Sciences, 2016, 45(6):143 - 146. (in Chinese)

[7] 巩如悦. 基于本体的苹果病虫害垂直搜索引擎研发[D]. 杨凌:西北农林科技大学, 2017.
GONG Ruyue. Research of vertical search engine for apple pests and diseases based on ontology[D]. Yangling: Northwest A&F University, 2017. (in Chinese)

[8] 于合龙, 沈金梦, 毕春光, 等. 基于知识图谱的水稻病虫害智能诊断系统[J]. 华南农业大学学报, 2021, 42(5):105 - 116.
YU Helong, SHEN Mengjin, BI Chunguang, et al. Intelligent diagnostic system for rice diseases and pests based on knowledge graph[J]. Journal of South China Agricultural University, 2021, 42(5):105 - 116. (in Chinese)

[9] 袁培森, 李润隆, 王翀, 等. 基于 BERT 的水稻表型知识图谱实体关系抽取研究[J]. 农业机械学报,2021,52(5):151 - 158.
YUAN Peisen, LI Runlong, WANG Chong, et al. Entity relationship extraction from rice phenotype knowledge graph based on BERT[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(5):151 - 158. (in Chinese)

[10] 李林, 周晗, 郭旭超, 等. 基于多源信息融合的中文农作物病虫害命名实体识别[J]. 农业机械学报, 2021, 52(12):253 - 263.
LI Lin, ZHOU Han, GUO Xuchao, et al. Named entity recognition of diseases and insect pests based on multi source information fusion[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(12):253 - 263. (in Chinese)

[11] 陈洁, 吕凯, 刘桂民. 基于文献计量的我国棉花病虫害研究进展[J]. 安徽农业科学, 2021, 49(23):231 - 238.
CHEN Jie, LÜ Kai, LIU Guimin. Research progress of cotton diseases and insect pests based on bibliometrics[J]. Journal of Anhui Agricultural Sciences, 2021, 49(23):231 - 238. (in Chinese)

[12] 汪敏. 茶树病虫害检测及防治信息挖掘与可视化分析[D]. 合肥:安徽农业大学, 2020.
WANG Min. Tea tree diseases and insect pests detection and control information mining and visual analysis[D]. Hefei: Anhui Agricultural University, 2020. (in Chinese)

[13] 武丽丽, 孙爻, 张礼生, 等. 基于 CNKI 文献计量的我国生物防治学科研究进展与发展态势分析[J]. 中国生物防治学报, 2019, 35(6):958 - 965.
WU Lili, SUN Yao, ZHANG Lisheng, et al. Research and analysis of Chinese biological control based on bibliometrics[J]. Chinese Journal of Biological Control,2019,35(6):958 - 965. (in Chinese)

[14] 袁丹. 花卉苗木冬季病虫害防治[J]. 农业与技术, 2016, 36(22):201.

[15] 芩情. 园艺花卉常见病虫害的治理原则与防治措施[J]. 江西农业, 2020(10):16,18.

[16] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT forself-supervised learning of language representations[C]// Proceedings of the 8th International Conference on Learning Representations, 2020: 1 - 17.

[17] WEI Z, SU J, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[C]// Proceedings of the 58th Annual Meeting of the Asscociation for Computational Linguistics, 2020:1476 - 1488.

[18] 吴茜. 基于知识图谱的农业智能问答系统设计与实现[D]. 厦门:厦门大学, 2019.
WU Qian. Design and implementation of agricultural intelligent Q&A system based on knowledge graph[D]. Xiamen: Xiamen University, 2019. (in Chinese)

[19] 夏迎春. 基于知识图谱的农业知识服务系统研究[D]. 合肥:安徽农业大学, 2018.
XIA Yingchun. Agriculture knowledge service system based on knowledge graph[D]. Hefei: Anhui Agricultural University, 2018. (in Chinese)

[20] 黄思思. 基于领域本体的中文信息抽取研究——以植物物种多样性领域为例[D]. 上海:华东师范大学, 2016.
HUANG Sisi. A research on Chinese information extraction based on construction of domain ontology—take plant species diversity for example[D]. Shanghai: East China Normal University, 2016. (in Chinese)

[21] 吴赛赛, 周爱莲, 谢能付, 等. 基于深度学习的作物病虫害可视化知识图谱构建[J]. 农业工程学报, 2020, 36(24):177 - 185.
WU Saisai, ZHOU Ailian, XIE Nengfu, et al. Construction of visualization domain-specific knowledge graph of crop diseases and pests based on deep learning[J]. Transactions of the CSAE, 2020, 36(24):177 - 185. (in Chinese)

[22] 徐明慧. 花卉病虫害防治[M]. 北京:金盾出版社, 1993.

[23] 蔡祝南. 花卉病虫害防治大全:彩图[M]. 北京:中国农业出版社, 2002.

[24] 陶降文, 陶泽文, 许丽娟. 花木病虫害防治图册[M]. 长沙:湖南科学技术出版社, 2010.

[25] JIANG C Y, ZHAO Y G, CHU S B, et al. Cold-start and interpretability: turning regular expressions into trainable recurrent neural networks[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: 2020: 3193 - 3207.

[26] RAMSHAW L A, MARCUS M P. Text chunking using transformation-based learning[M]// Natural language processing using

very large corpora. Springer, Dordrecht, 1999: 157 – 176.
[27] DAI D, XIAO X Y, LÜ Y J, et al. Joint extraction of entities and overlapping relations using position-attentive sequence labeling[C]// Thirty-third AAAI Conference on Artificial Intelligence, 2019: 6300 – 6308.
[28] 胡滨,汤保虎,姜海燕,等. 家禽诊疗文本多实体关系联合抽取模型研究[J]. 农业机械学报, 2021, 52(6): 268 – 276.
HU Bin, TANG Baohu, JIANG Haiyan, et al. Joint extraction model of multi-entity relations for poultry diagnosis and treatment text[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(6): 268 – 276. (in Chinese)
[29] 李伟康, 李炜, 吴云芳. 深度学习中汉语字向量和词向量结合方式探究[J]. 中文信息学报, 2017, 31(6):140 – 146.
LI Weikang, LI Wei, WU Yunfang. Combination methods of Chinese character and word embeddings in deep learning[J]. Journal of Chinese Information Processing, 2017, 31(6):140 – 146. (in Chinese)
[30] 李新琴, 史天运, 李平, 等. 基于文本的高速铁路信号设备故障知识抽取方法研究[J]. 铁道学报, 2021, 43(3):92 – 100.
LI Xinqin, SHI Tianyun, LI Ping, et al. Research on knowledge extraction method for high-speed railway signal equipment fault based on text[J]. Journal of the China Railway Society, 2021, 43(3):92 – 100. (in Chinese)
[31] 陈晓玲, 唐丽玉, 胡颖, 等. 基于 ALBERT 模型的园林植物知识实体与关系抽取方法[J]. 地球信息科学学报, 2021, 23(7):1208 – 1220.
CHEN Xiaoling, TANG Liyu, HU Ying, et al. Extracting entity and relation of landscape plant's knowledge based on ALBERT model[J]. Journal of Geo-information Science, 2021, 23(7):1208 – 1220. (in Chinese)
[32] 赵鹏飞, 赵春江, 吴华瑞, 等. 基于 BERT 的多特征融合农业命名实体识别[J]. 农业工程学报, 2022, 38(3):112 – 118.
ZHAO Pengfei, ZHAO Chunjiang, WU Huarui, et al. Recognition of the agricultural named entities with multi-feature fusion based on BERT[J]. Transactions of the CASE, 2022, 38(3):112 – 118. (in Chinese)
[33] SIENČNIK S K. Adapting word2vec to named entity recognition[C]// Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015), 2015: 239 – 243.
[34] GRAVES A, FERNÁNDEZ S, SCHMIDHUBER J. Bidirectional LSTM networks for improved phoneme classification and recognition[C]// International Conference on Artificial Neural Networks, 2005.

(上接第 189 页)

[15] 何杰,朱金光,罗锡文,等. 电动方向盘插秧机转向控制系统设计[J]. 农业工程学报,2019,35(6):10 – 17.
HE Jie, ZHU Jinguang, LUO Xiwen, et al. Design of steering control system for rice transplanter equipped with steering wheel-like motor[J]. Transactions of the CSAE,2019,35(6):10 – 17. (in Chinese)
[16] 张智刚,罗锡文,李俊岭. 轮式农业机械自动转向控制系统研究[J]. 农业工程学报,2005,21(11):77 – 80.
ZHANG Zhigang, LUO Xiwen, LI Junling. Automatic steering control system of wheeled model farming machinery[J]. Transactions of the CSAE,2005,21(11):77 – 80. (in Chinese)
[17] ZHUA Qingyuan, WEI Chena, HU Huosheng, et al. Multi-sensor based attitude prediction for agricultural vehicles[J]. Computers and Electronics in Agriculture, 2019,156:24 – 32.
[18] QIU Quan, FAN Zhengqiang, MENG Zhijun, et al. Extended-Ackerman steering principle for the coordinated movement control of a four-wheel drive agricultural mobile robot[J]. Computers and Electronics in Agriculture, 2018,152: 40 – 50.
[19] 张瑶,李佩娟,贾茜,等. 基于自适应粒子群的农机自动转向系统分数阶控制[J]. 甘肃农业大学学报,2019,54(4):199 – 204.
[20] 沈跃,何思伟,刘慧,等. 高地隙喷雾机自转向电动底盘控制系统设计与试验[J]. 农业机械学报,2020,51(11):385 – 392,402.
SHEN Yue, HE Siwei, LIU Hui, et al. Modeling and control of self-steering electric chassis structure of high clearance sprayer[J]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(11):385 – 392,402. (in Chinese)
[21] 刘国海,李持衡,沈跃,等. 同步转向高地隙喷雾机模糊自适应轨迹跟踪预测控制[J]. 农业机械学报,2021,52(9):389 – 399.
LIU Guohai, LI Chiheng, SHEN Yue, et al. Trajectory tracking and fuzzy adaptive model predictive control of high clearance synchronous-steering sprayer[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021,52(9):389 – 399. (in Chinese)
[22] 高志强. 自抗扰控制思想探究[J]. 控制理论与应用,2013,30(12):1498 – 1510.
GAO Zhiqiang. On the foundation of active disturbance rejection control[J]. Control Theory & Applications, 2013,30(12): 1498 – 1510. (in Chinese)
[23] GAO Zhiqiang. Scaling and bandwidth-parameterization based controller-tuning[C]// Proceedings of the 2003 American Control Conference. Denver: IEEE,2003: 4989 – 4996.
[24] HAN Jingqing. From PID to active disturbance rejection control[J]. IEEE Transactions on Industrial Electronics, 2009, 56(3):900 – 906.
[25] 李向阳,高志强. 抗扰控制中的不变性原理[J]. 控制理论与应用,2020,37(2):236 – 244.
LI Xiangyang, GAO Zhiqiang. The invariance principle in disturbance rejection control[J]. Control Theory & Applications, 2020, 37(2): 236 – 244. (in Chinese)