DOI: 10.12398/j.issn.2096-7217.2024.04.003

基于知识图谱的农作物病虫害问答系统研究

赵泽行[1]，吴晓鹏[1]，王怡馨[2]，闫小丽[1]，黄玉祥[1]，高筱钧[3, 4*]

(1. 西北农林科技大学机械与电子工程学院，陕西杨凌，712100；2. 西北农林科技大学信息工程学院，陕西杨凌，712100；3. 南京农业大学工学院，江苏南京，210031；4. 农业装备技术全国重点实验室，北京市，100083)

摘要：在大数据时代的推动下，智能问答系统在各个领域得到广泛应用，为用户提供高效的答案，相比传统的文本知识收集和网络搜索引擎检索，具有明显优势。随着知识图谱技术的快速发展，智能问答系统迎来了新的发展阶段。本研究响应了智慧农业的需求，构建了一个基于知识图谱的农作物病虫害智能问答系统，旨在为用户提供与农作物病虫害相关的问答服务。主要工作包括：(1) 农作物病虫害数据获取：利用分布式爬虫框架爬取农作物病虫害相关网页数据，并进行数据清理、分析、结构化等预处理操作。(2) 知识图谱构建：对数据进行分析后，定义知识本体的实体类别与关系类别，完成知识图谱的模式层构建。利用基于规则的三元组模板对半结构化文本进行实体提取，构建数据层，并将三元组存储至Neo4j图数据库。(3) 问答算法设计：利用BERT-BiLSTM-CRF模型进行问句实体识别，BERT-RNN模型进行问题分类。匹配模板后通过Cypher语句进行查询，将答案处理为自然语言形式并返回。(4) 问答系统实现与可视化：结合农作物病虫害知识图谱与问答算法，使用Flask框架和多种web技术实现用户提问、实体识别、知识检索和答案返回等功能。试验结果表明，实体识别与问题分类模型的准确率(Precision)、召回率(Recall)和$F1$分别达到了93.22%、92.69%、92.21%和94.37%、93.92%、92.66%。与其他搜索途径相比，问答系统展现出了较高的准确性和稳定性。这项研究为农业信息化提供了一种智能化的解决方案，为用户获取农作物病虫害知识提供了新途径。

关键词：农业信息化；农作物病虫害；知识图谱；问答系统

中图分类号：S24 **文献标识码：**A **文章编号：**2096-7217 (2024) 04-0039-12

赵泽行，吴晓鹏，王怡馨，闫小丽，黄玉祥，高筱钧．基于知识图谱的农作物病虫害问答系统研究[J]．智能化农业装备学报(中英文)，2024，5(4)：39－50

ZHAO Zexing, WU Xiaopeng, WANG Yixin, YAN Xiaoli, HUANG Yuxiang, GAO Xiaojun. Research on question answering system for crop diseases and pests based on knowledge graph [J]. Journal of Intelligent Agricultural Mechanization, 2024, 5(4): 39－50

0 引言

随着农业领域与信息技术的深度融合，以及大数据技术在农业生产中的广泛应用，农业生产过程中的数据量不断攀升，庞大的农业数据源从土壤传感器[1]、农业无人机[2]、气象站[3]等多种监测源涌出，目前，信息化与数据化已经成为推动农业现代化的关键组成部分[4]。

在众多农业问题中，作物病虫害是限制农业可持续发展的主要因素之一，因而其相关知识与数据是灾前预防和灾中缓解的重要信息。目前，绝大多数农业从业者选择通过网络获取病虫害相关信息，然而，互联网上的农业知识呈现出多源异构、混乱无序的特点，使得准确获取所需信息变得相当困难[5]。

当前，传统搜索引擎在搜索农作物病虫害信息时普遍存在以下问题：(1)搜索引擎容易出现对用户意图的偏差理解、遗漏理解和错误理解。(2)搜索结果之间相互独立，难以形成连贯性的信息体系。(3)搜索结果的质量参差不齐，网页内容鱼龙混杂，用户体验差。

随着农业信息化的发展，农作物病虫害知识受到地域性差异的影响，表现出翻译、称谓和别名不一致的问题[6]。因此，在庞大而分散的农业数据中快速准确获取农作物病虫害信息成为当前急需解决的难题。基

收稿日期：2023-12-27 修回日期：2024-03-21

基金项目：农业装备技术全国重点实验室开放课题(NKL-2023-008)；陕西省重点研发计划项目(2023-YBNY-204)；中国博士后科学基金项目(2022M722609)；陕西省博士后科研项目(2023BSHEDZZ149)

第一作者：赵泽行，男，2001年生，河南许昌人；研究方向为智慧农业与深度学习。E-mail：2019015213zzx@nwafu.edu.cn

*通信作者：高筱钧，男，1990年生，黑龙江齐齐哈尔人，博士，副教授；研究方向为高速精播技术与装备等。E-mail：gxj-1234@126.com

于知识图谱的智能问答系统(KBQAS)为解决这一问题提供了高效的解决方案。

知识图谱(knowledge graph,KG)是一种结构化的语义知识库,采用了“节点—边—节点”的形式描述客观世界中的概念及其关系[7]。在这个模型中,节点对应实体,即客观世界中存在的概念,而边代表关系,即实体之间的联系。实体属性则表示实体所具备的可观测性质。这种高度结构化的数据库可以清晰地存储庞大且复杂的病虫害数据网络。节点可以涵盖多个维度的实体,例如农作物、具体病害和虫害种类,而边则代表作物与特定病害的关联性或不同病害之间的相互影响,实体属性则详细记录各实体的参数、性质等可观测特征。

当前,越来越多从事农业信息化研究的学者致力于农业知识图谱的建设。扬州大学构建了多模态农业知识图谱及其应用系统[8];中国农业大学实现了基于知识图谱的作物病虫害关联挖掘及智能诊断[9]。此外,学者们还专注于某一作物领域的知识图谱构建,如王丹丹[10]采用自顶向下和自底向上相结合的方法,成功构建了水稻知识图谱。张永亮[11]、闫丽华[12]则分别致力于苹果和葡萄领域相关病虫害知识图谱的建立。

为解决传统搜索引擎在搜索农作物病虫害相关知识时效性低下的问题,借助知识图谱将知识结构化存储可以有效减缓知识同源异构的问题,同时提高知识的准确性与关联性。然而,为实现高效的搜索与问题回答,仍需要建立能够精准理解用户意图,并能够准确定位相关知识最终返回答案的智能问答系统。

基于以上问题,本研究以农业病虫害知识信息化为出发点,旨在提出高效农作物病虫害知识搜索方案,构建知识图谱,并在知识图谱的基础上构建智能问答系统。

1 问答系统模块设计

面对传统搜索引擎在搜索农作物病虫害相关知识时存在的多方面问题,本研究拟利用分布式网络爬虫获取农作物病虫害知识,构建基于图数据库的农作物病虫害知识图谱,以实现知识的结构化存储。同时,运用基于语义分析和模板匹配相结合的方法设计以知识图谱为基础的问答算法。最终,结合上述工作与后端框架,构建一种具备准确性和稳定性的农作物病虫害知识问答系统,并实现系统的可视化与实用性。

对于上述任务与目标,本研究的主要工作分为两个方面,即农作物病虫害知识图谱的构建与农作物病虫害智能问答系统的实现。其中,农作物病虫害知识图谱的构建旨在将从网络上获取的知识文本以结构化存储在图数据库中。这一过程包括构建前的数据获取与数据预处理,以及构建过程中模式层与数据层的建立,最终完成构建后的数据存储。而农作物病虫害智能问答系统的实现则将构建好的知识图谱作为基础知识库,结合前后端框架为用户提供问答服务。因此,问答系统的实现不仅包括智能问答算法的开发,还需完成系统的可视化与测试工作。本研究的具体工作与系统模块框架如图1所示。

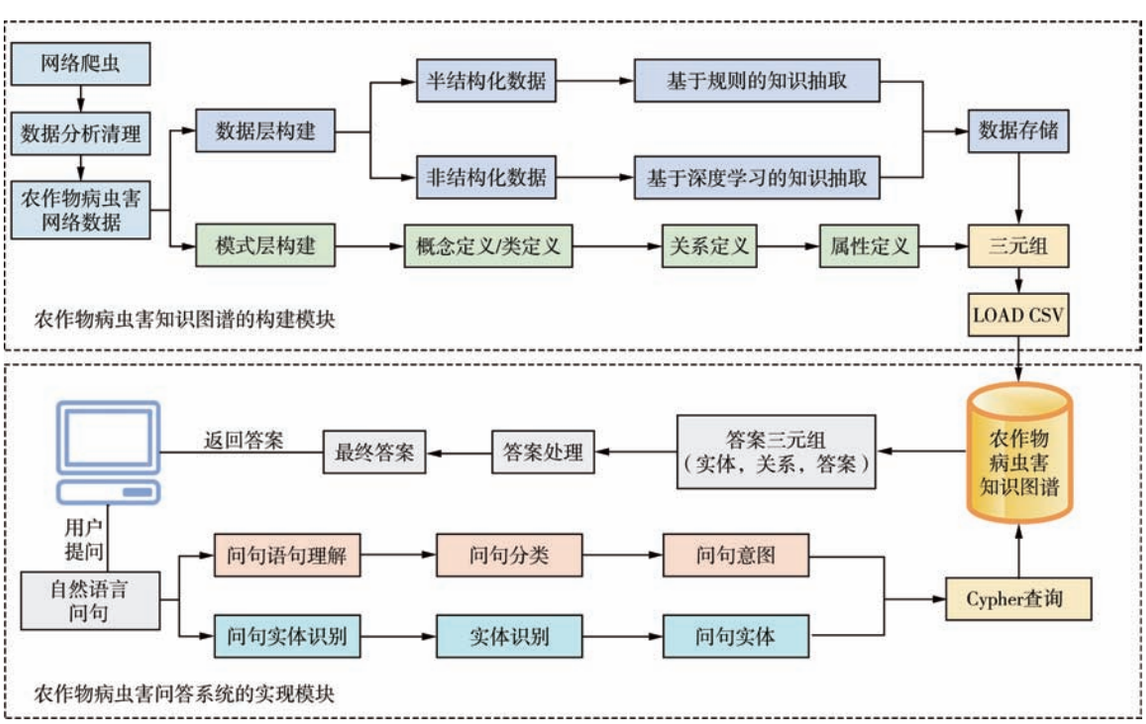


图1 基于知识图谱的农作物病虫害问答系统的模块设计

Figure 1 Module design of Q&A system for crop diseases and pests based on knowledge graph

2　农作物病虫害知识图谱的构建

农作物病虫害知识作为特定领域的深度数据，适合采用垂直领域知识图谱的构建方法，即自顶向下的构建方式。这种方法首先进行模式层的构建，然后再进行数据层的构建，以应对领域内概念数量庞大、概念层次关系复杂的情况[13]。结合网络文本知识数据，通过数据清洗、知识抽取等步骤，可以完成农作物虫害知识图谱的构建。

2.1　知识获取与预处理

本研究选取中国作物种质信息网(Chinese Crop Germplasm Resources Information System)—农业知识—植物病虫害知识专栏(https://www.cgris.net/disease/default.html)中16类农作物的病虫害知识作为主要知识源，以维基百科(Wikipedia)与国家农业科学数据中心(https://www.agridata.cn/#/home)的对应农作物病虫害数据作为补充知识源[14]。

数据获取的具体操作包括以下步骤：首先，通过定位目标网站的统一资源定位系统(URL)，找到包含农作物病害和虫害信息的网页。然后，通过解析网页信息，提取农作物病害和虫害的各种属性信息，包括但不限于学名、寄主、危害特点、形态特征、生活习性、防治方法、症状、病原、传播途径和发病条件等属性。为了高效地进行数据获取，可以利用Python中的分布式爬虫框架，从网页上爬取病虫害知识，并将其保存为纯文本文件，通常使用json格式进行保存。最终，通过这样的数据获取方式，可以获得约18 000条病虫害知识文本数据。

为了确保知识图谱构建过程中的数据抽取工作高效准确，我们需要对爬取的知识文本进行数据预处理。这些文本保留了源网页的结构分布，但存在一些问题，如干扰字符、字段冗余、计量单位不一致、错行跳行等情况，导致结构混乱、内容重复。因此，我们需要进行以下预处理工作：

1)数据清理去噪声：去除文本中的重复和无关内容，例如资料来源和数据声明等其他信息，以保证数据的纯净性。

2)数据分析：分析知识内容之间的从属关系，帮助确定每种病虫害的属性及其之间的关联关系，为后续的数据结构化提供依据。

3)数据结构化：根据前期的数据分析结果，将文本数据按照统一的结构进行布置，使每一种病虫害的知识内容能够以清晰、有序的方式呈现，便于后续的知识图谱构建工作。

经过预处理工作，我们将得到纯净的半结构化农作物病虫害文本数据，每一种病虫害的知识内容将以分段分布的形式呈现，属性内容完整且条理清晰。这样的数据将有助于后续的知识图谱构建工作，使其更加准确和高效。

2.2　知识图谱本体构建

农作物病虫害知识图谱的构建过程可以分为两个主要步骤：模式层和数据层的建构。模式层作为知识图谱的核心组成部分，旨在明确定义农作物病虫害领域知识本体的边界与范围。而数据层则包含了具体的知识库信息，与模式层相互关联，共同构成完整的知识图谱。本研究首先着重于模式层的定义，设计了农作物病虫害知识图谱的数据模式。随后，结合经过预处理的知识数据，系统地构建了数据层，以确保知识图谱的完备性与准确性。具体的模式层与数据层的构建框架如图2所示。

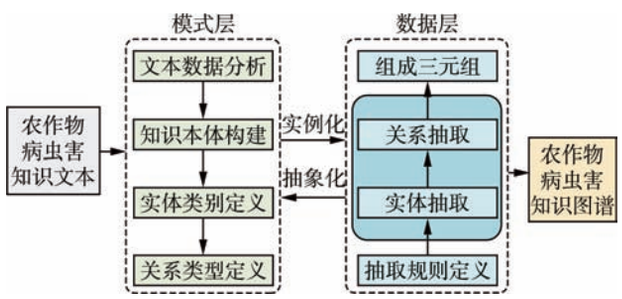


图2　基于知识图谱的农作物病虫害问答系统的模块设计

Figure 2　Construction framework of knowledge graph schema layer and data layer

2.2.1　模式层构建

通过对农作物病虫害领域知识实体的边界确定与数据结构分析，最终确定了农作物病虫害知识图谱的模式概念，包括12个类别，涵盖了1 619种农作物病虫害实体及其下属子实体。考虑数据结构特点以及自动问答系统的实际应用场景需求，本研究对某一农作物病虫害进行了细分，将其划分为病害和虫害两个部分。这两个部分分别包含若干个病害或虫害的相关概念实体，其中病害部分包括44个相关概念，虫害部分包括5个相关概念。

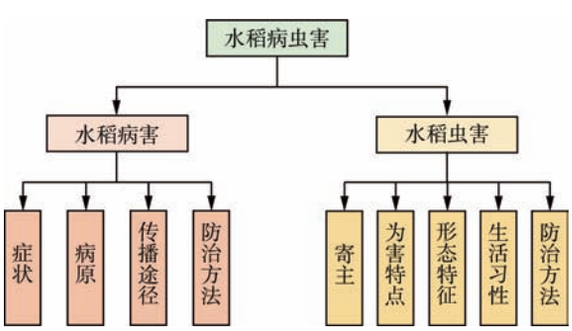


图3　农作物病虫害知识图谱概念层次

Figure 3　Conceptual hierarchy of knowledge graph

图3以水稻病虫害为例,详细展示了具体的概念层次结构。根据已构建的农作物病虫害概念层次,筛选高频概念,过滤部分模糊、低频概念,合并相同或相似概念,最终可以确定农作物病虫害知识图谱中实体的类别包括10类。具体类别信息如表1所示。

表1 农作物病虫害知识领域实体类别

Table 1 Knowledge domain entity categories

实体类别	含义描述	实体定义
农作物	常见农作物	CRO
病害	农作物常见病害	DIS
虫害	农作物常见虫害/害虫	PES
症状	农作物病害的为害症状	SYM
病原	引发农作物病害的根源	PAT
传播途径	病害病原的传播方法	TRA
寄主	农作物害虫的为害对象	HOS
为害特点	农作物虫害的为害症状	CHA
形态特征	农作物害虫的形态特征	FIG
生活习性	农作物害虫的生活习性	HAB
防治方法	农作物病虫害的防治途径	MET

在得到农作物病虫害知识本体内所有实体概念后,分析定义概念之间的关系,得到知识图谱中的所有关系共9种,如表2所示。

表2 农作物病虫害知识领域实体类别

Table 2 Knowledge domain relationship categories

关系类别	含义描述
Disease Pest	农作物——病/虫害
Disease Symptom	病害——症状
Disease Pathogeny	病害——病原
Disease Transmission	病害——传播途径
Pest Host	虫害——寄主
Pest Character	虫害——为害特点
Pest Figure	虫害——形态特征
Pest Habit	虫害——生活习性
Prevent Method	病/虫害——防治方法

根据以上概念层次和设计的实体与关系类别,本文又分别设计了涵盖各类关系的病害知识图谱概念模式和虫害知识图谱概念模式。病害和虫害分别包含4类实体,4类关系和5类实体,5类关系。如图4所示。

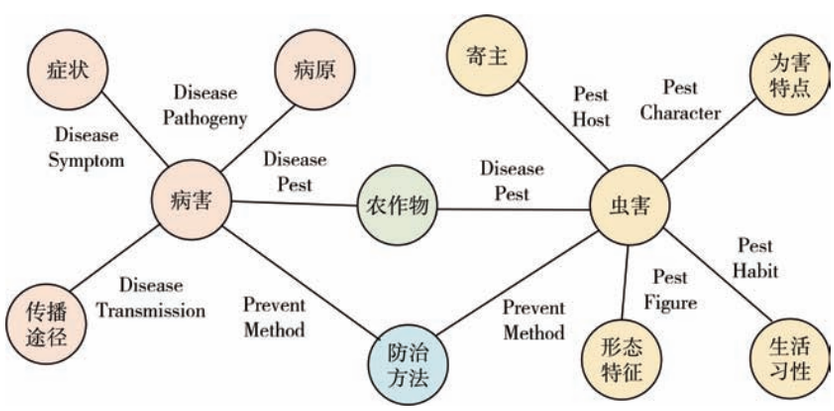


图4 农作物病虫害知识图谱模式层

Figure 4 Schema layer of knowledge graph

2.2.2 数据层构建

经过先前的数据清理和一系列数据预处理操作之后,获取了大量农作物病虫害的半结构化知识文本。这些文本的标题即为作物病虫害实体的名称,段落层级表示了某一病虫害实体下属实体的分布,而段落的小标题则对应下属实体的名称。例如,水稻稻瘟病实体下分布有症状、病原、传播途径、防治方法等下属实体,小标题后的文本则包含属性内容等信息。

因此,本研究可以利用这些半结构化特征,制定相应的规则进行三元组的提取。具体的规则形式为(首实体1,关系1,尾实体1)、(首实体2,关系2,尾实体2)……(首实体n,关系n,尾实体n)。在这些三元组中,关系属于之前定义的9类关系之一;首实体和尾实体分别代表关系连接的两个实体,它们都属于之前定义的10类实体,且不属于同一类别。由于本文未对被动式关系进行定义,因此在三元组中,首实体与尾实体的位置不能互换。例如,对于文本“水稻显纹纵卷叶螟的寄主为水稻、陆稻”,三元组的提取规则应为(虫害实体,PestHost,寄主实体),因此(水稻显纹纵卷叶螟,PestHost,水稻/陆稻)是合规三元组,而(水稻/陆稻,PestHost,水稻显纹纵卷叶螟)是不合规三元组。

对于部分非结构化数据,置信度较高的做法是人

工标注，虽然可靠准确，但是工作量较大[15]。本研究做法是通过自然语言处理手段对文本进行分段、分句、分词、去除停用词后再进行文本的结构化，识别文本中的实体与关系后，利用定义好的规则进行提取。

通过规则提取出的实体数据和三元组数据保存在CSV格式的文件中，每类关系单独储存，以病害实体和病原实体之间的关系——DiseasePathogeny为例，具体形式如表3所示。

表3　三元组数据储存形式

Table 3　Storage of triple

首实体	关系	尾实体
水稻恶苗病	DiseasePathogeny	Fusarium moniliforme Sheld
水稻烂秧病	DiseasePathogeny	Fusarium graminearum Schw
水稻白绢病	DiseasePathogeny	Sclerotium rolfsii Sacc
水稻白叶枯病	DiseasePathogeny	Xanthomonas oryzae
…	DiseasePathogeny	…

2.3　知识存储

目前常见的存储知识的方式有3种，三元组形式的资源描述框架[16]（resource description framework，RDF）存储，传统关系型数据库存储以及图数据库存储。相较于前两种方法，图数据库将实体概念抽象为图中的节点，并将实体属性和关系表示为节点之间的边。这种存储方式使得知识图谱以图的形式存储数据，能够最大程度地保留各实体与概念之间的位置结构和属性关系。由于图数据库直观地反映了知识图谱的内部结构，因此能够为图查询、实体链接以及知识推理等操作提供便利。此外，图数据库具有较强的可扩展性，能够有效地处理大规模的数据，从而满足知识图谱在实际应用中对于高效存储和管理大量数据的需求[17]。鉴于图数据库具有多种优势，本研究所构建的农作物病虫害知识图谱使用Neo4j图数据库来存储，并利用Load CSV方法，批量导入已写好CSV格式文件创建节点与关系。

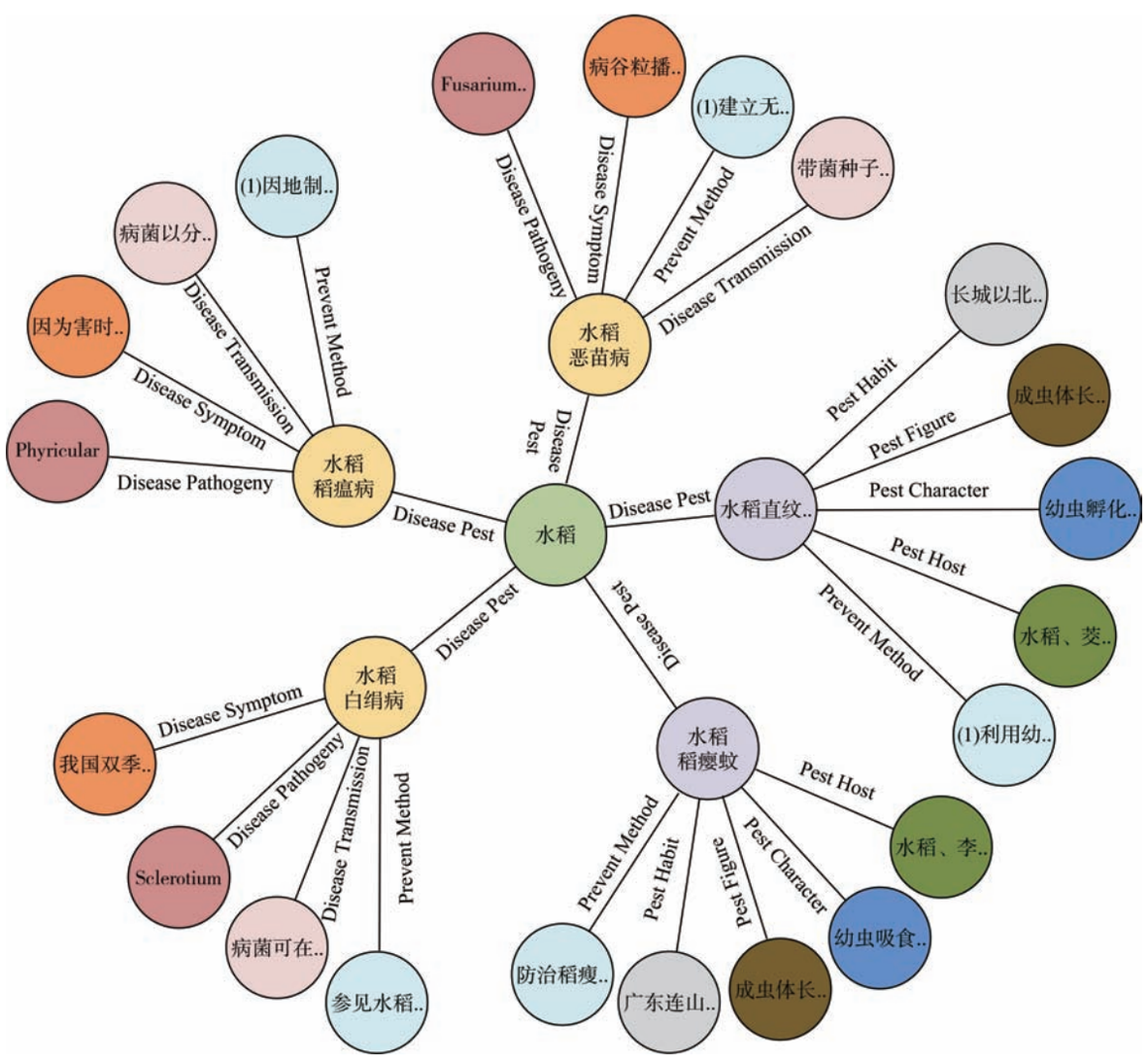


图5　农作物病虫害知识图谱部分实例

Figure 5　Partial examples of knowledge graph of crop pests and diseases

将所有三元组CSV文件导入到Neo4j图数据库中之后,可以通过Neo4j提供的浏览器将农作物病虫害知识图谱的可视化。Neo4j中圆形节点代表实体,连线代表关系。以水稻实体的部分知识图谱为例,如图5所示。

3 农作物病虫害智能问答系统的实现

在农作物病虫害知识图谱构建完成后,该图谱将作为知识基础,为智能问答系统提供丰富的知识储备。智能问答系统充当用户与知识图谱之间的桥梁,实现二者之间的有机交互。一个完整的问答系统应具备接受用户问题,并根据问题向用户返回相关答案的功能。因此,问答算法的设计、系统的搭建以及可视化是问答系统最终实现过程中不可或缺的重要组成部分。

3.1 问答系统工作流程

在系统实现问答功能的过程中,系统首先对用户输入的中文问题进行实体与关系识别,以获得问题语句的关键词。然后通过问题分类器,将用户问题近似看成分类问题,并根据分类结果选择概率最大的确定问题类别。随后,针对分类问题,系统利用Cypher语句在建立的知识图谱中进行搜索和查找,并返回最佳匹配答案。最终,系统通过可视化界面将答案呈现给用户。具体工作流程如图6所示。

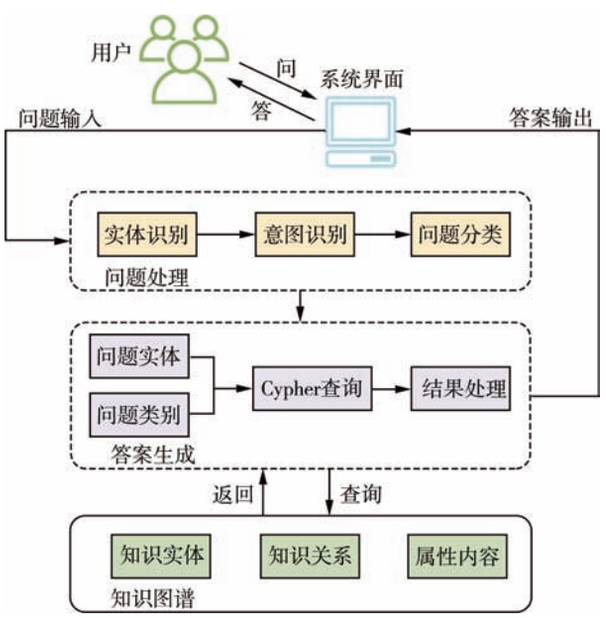


图6 问答系统工作流程

Figure 6 Workflow of Q&A system

3.2 问答算法的设计

3.2.1 基于BERT-BiLSTM-CRF的问题实体识别

由于中文属于高语境语言,语义较为含蓄,语言结构复杂,语句中没有明显的词语分隔符[18],农业知识领域中名词复杂,容易引起歧义[19],经典的实体识别模型不能很好地应对这类问题。因此本研究采用BERT-BiLSTM-CRF实体识别模型[20],即将基于语料库预训练的BERT模型与BiLSTM-CRF实体识别模型结合,通过对字的混合语义表示进行建模,利用词向量技术来解决多义性问题,并分析目标词的上下文信息,进一步实体识别的准确性。

BERT-BiLSTM-CRF作为有监督学习模型,在进行训练时需要带有标签信息的文本数据作为数据集,本研究抽取部分先前得到的经数据预处理后农作物病虫害文本语料,并采用BIOES标注模式进行标注。BIOES标注模式中,B为Begin代表实体的首字符,I为Inside代表实体的中间字符,E为End代表实体的尾字符,S为Single,O为Other分别代表单字实体与其他内容[21]。

在确定标注规则后,BERT-BiLSTM-CRF实体识别模型的具体工作流程如下,BERT模型作为预训练模型,率先对用户输入的中文问句进行向量化表示,并将字符向量作为输出直接输入到BiLSTM模型中,随后字符向量进入由两个LSTM模块组成的编码层,字符向量序列先被输入正向模块学习前向的隐特征向量,然后被输入反向模块学习反向的隐特征向量,最后被拼接组成最终的向量。最后输出的向量经过CRF层,被加上标签概率约束,输出整体概率最大的实体标注序列。详细框架如图7所示。

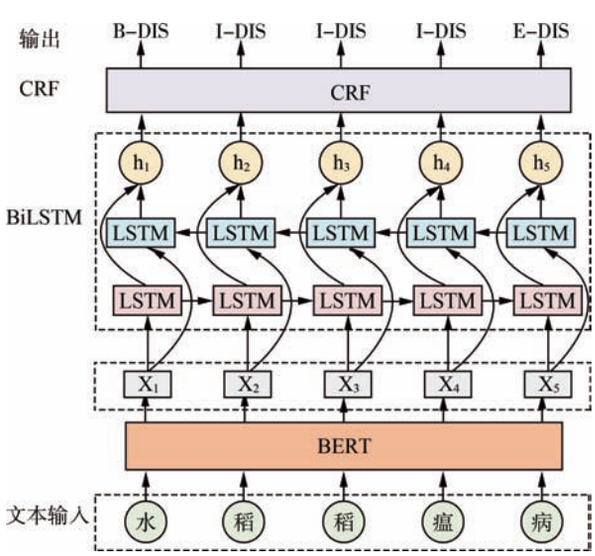


图7 BERT-BiLSTM-CRF工作框架

Figure 7 Framework of BERT-BiLSTM-CRF

3.2.2 基于BERT-RNN的问题分类

对用户输入的问题进行实体识别后,可以得到目标实体,即用户咨询的对象,之后需要对问题进行分类,进一步确定用户关于此对象的问题意图。以实体“水稻稻瘟病”为例,问题“什么是水稻稻瘟病”的意图

则是了解此种病害，而问题“水稻稻瘟病的防治方法是什么”的意图则是进一步了解此种病害的具体防治方法。

本研究采用BERT-RNN模型[22]分类，与上文出现的BERT-BiLSTM-CRF模型相似，BERT模型同样作为预训练模型，在用户输入问题后，同样由BERT模型投影至向量空间，进行向量化表示，随后输入RNN（recurrent neural network）中进行序列建模。RNN中包含循环神经网络单元，可以接收输入数据以及来自前一个时间步的隐藏状态，并生成新的隐藏状态。该单元通过将当前时间步的输入与前一个时间步的隐藏状态结合起来，来捕捉输入语句的时序信息。随后通过全连接层，将RNN输出的语义表示映射到预定义的类别空间。详细框架如图8所示。

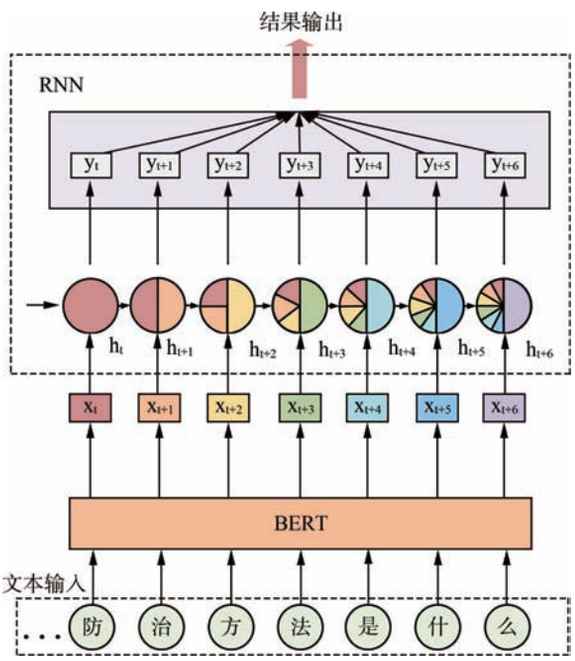


图8　BERT-RNN工作框架

Figure 8　Framework of BERT-RNN

3.2.3　试验结果与分析

本文抽取先前获得的数据中共计10 000条农作物病虫害文本，进行标注后作为试验数据。以7∶3的比例随机抽取数据作为训练集和测试集，分别用于实体识别与问题分类试验。

本研究基于Ubuntu 22.04操作系统，使用版本3.7的Python作为脚本语言，利用Pytorch 1.7.1版本作为深度学习框架进行试验，具体试验环境配置如表4所示。

对于模型训练过程中神经网络的超参数本文参考前者的试验工作后确定，具体超参数设置如表5所示。

表4　试验环境配置

Table 4　Experimental environment configuration

工具/试验环境	版本参数
Python	3.7
Pytorch	1.7.1 + cu110
Sklearn	0.24.1
Numpy	1.20.1
GPU	RTX 3090
操作系统	Ubnutu 22.04

表5　模型超参数配置

Table 5　Model hyperparameter configuration

参数名称	参数值
Batch size	64
Optimizer	Adam
Transformer layer	12
Learning rate	0.000 1
Dropout ratio	0.3/0.2
Embedding Size	512/256
LSTM_emb	200/—
Epoch	100/200

本研究的试验属于自然语言处理的任务，采用3项试验指标对结果进行评价，分别是精确率P（Precision），召回率R（Recall）和$F1$值，以实体识别任务为例，精确率P即在所有被识别出的实体中，确实是正确实体的比例。其中TP表示正确预测，FN表示实际为正确，但预测为假，FP反之。计算公式如式（1）。

$$P=\frac{TP}{TP+FP}\times 100\% \tag{1}$$

召回率R即正确识别实体数量与实体总数量的比例，即问句中的实体是否能完全识别出，反映实体识别的查全率，计算公式如式（2）。

$$R=\frac{TP}{TP+FN}\times 100\% \tag{2}$$

$F1$是通过对精确度和召回率的计算得到的，是一个较为综合的参考指数，公式如式（3）。

$$F1=\frac{2PR}{P+R}\times 100\% \tag{3}$$

实体识别与问题分类的试验结果如表6所示。

根据试验结果，相较于其他同类型模型，BERT-BiLSTM-CRF和BERT-RNN在各项评估指标上表现更为优秀，这两个模型在实体识别和问题分类任务中均展现出更出色的性能。此外，结果还表明了BERT预训练模型在生成词向量方面的卓越能力，使其能够

更有效地表征不同语境下的语义信息。这种向量表的方式有助于提升模型的泛化能力,使其能够更好地适应多领域的文本,从而提高在实际应用中的性能。

表 6 试验结果

Table 6 Experimental result

模型对比		P/%	R/%	$F1$/%
实体识别	BiLSTM	88.68	87.36	86.91
	BiLSTM-CRF	91.85	90.33	90.54
	BERT-BiLSTM-CRF	93.22	92.69	92.21
问题分类	BERT	94.04	93.26	92.21
	RNN	93.55	93.36	91.95
	BERT-RNN	94.37	93.92	92.66

3.3 基于Cypher语句的答案检索

在对用户输入的问题进行实体识别与分类后,就可以将问题转化为模板形式,并计算其与预定义模板之间的相似度。若相似度大于已设定的阈值则将相似度最高的模板作为用户输入问题,随后执行相对应的Cypher查询语句,返回用户输入问题答案。具体流程如图9所示。

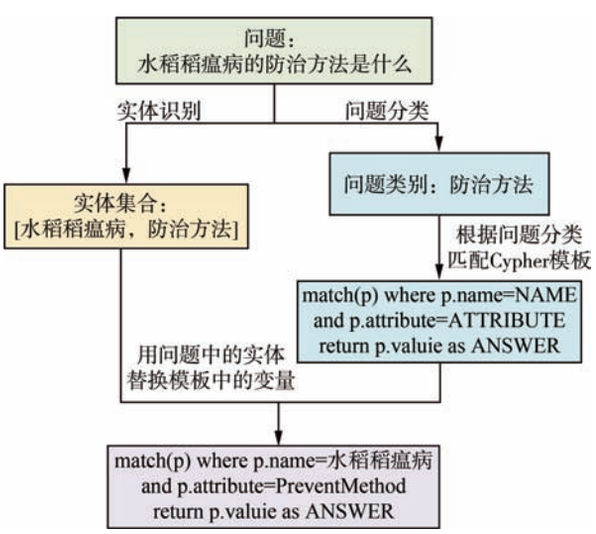


图 9 Cypher答案检索流程

Figure 9 Cypher answer retrieval process

3.4 问答系统的实现与可视化

3.4.1 系统构架设计

作为一个能为用户回答农作物病虫害知识的问答系统,本系统需完成问答界面的可视化,以满足用户的交互需求,同时还需链接构建好的知识图谱与结构化知识库,通过检索答案后返还,以满足用户的知识需求。为实现问答系统各项功能,满足用户的各类需求,经分析本系统的采用垂直的5层框架,由4层向上堆叠的主体层和1层作为基础的基础技术层组成,如图10所示。

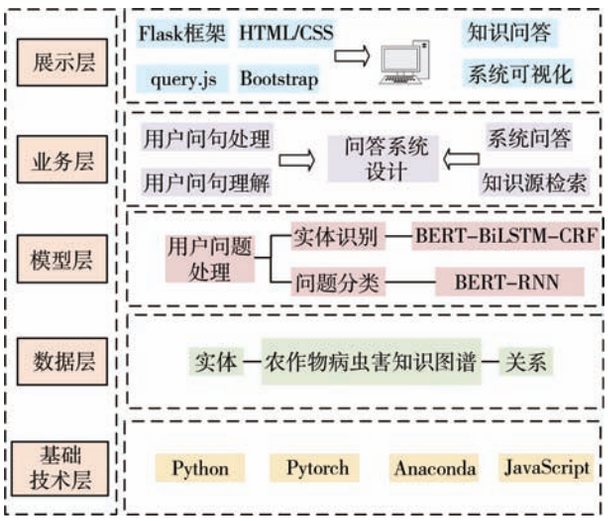


图 10 问答系统构架

Figure 10 Q&A system architecture

3.4.2 系统可视化

本研究的问答系统所采用的是基于Python轻量级web框架——Flask框架,通过运行本地Python脚本后,向web端发送网页请求,系统即可在浏览器中访问。作为一个完整的系统,本问答系统主要包含两个界面,主界面与问答界面。

主界面用于系统的信息展示,如图11所示。界面设计简洁,嵌入了一个“开始问答”按钮,用于进入问答界面;问答界面主要分为两个模块,问答交互模块与用户输入模块,如图12所示。问答交互模块类似于常规的聊天软。



图 11 问答系统主界面

Figure 11 Main interface of Q&A system

为了应对一些无法作答的或问句实体无法识别的常用问句,本系统采用基于关键词相似度匹配的快速回答机制,即针对某些特殊关键词,用户输入后会直接进行相似度匹配,如相似度达标,则快速返回既定的模板答案。如系统预先设定“你好/hi/hello”等关键词,用户向系统打招呼,输入“你好”,系统识别到关键词,快速回复“你好,我是农作物病虫害知识问答系统…”。如果问句中未出现关键词,则照常进行问句实体识别,根据问句实体返还答案,具体流程如图13所示。



图 12　问答系统问答界面

Figure 12　Q&A interface of Q&A system

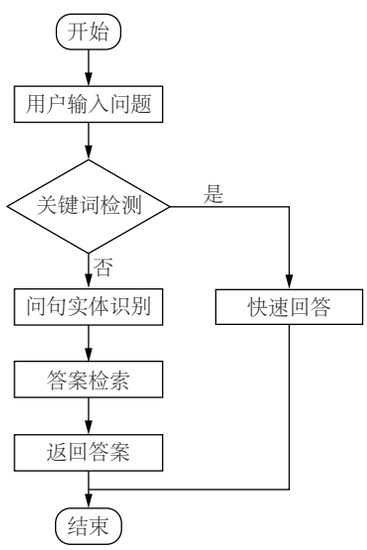


图 13　快速回答机制的工作流程

Figure 13　Workflow of the quick answer mechanism

为检测问答系统是否可以在一定程度上满足用户对于农作物病虫害知识问答的需求，本研究采用多系统问答结果对比来验证本系统的优越性[23-24]。具体工作为，选取某一符合提问规则且具有确切答案的农作物病虫害知识问题，分别向传统搜索引擎、本文的农作物病虫害知识问答系统、大规模语言模型提问，并收集回答与正确答案对比。本文选取的传统搜索引擎为百度搜索，大规模语言模型为基于 ChatGPT 3.5；所选取的问题为："水稻稻苗疫病的防治方法是什么"，结果对比如图 14 所示，图 14(a)、图 14(b)、图 14(c)、图 14(d)分别为正确答案、本问答系统返还的答案、百度搜索结果、ChatGPT 3.5 返还的答案。

经对比可知，本研究的问答系统所返还的答案相比于百度搜索结果、ChatGPT 3.5 返还的答案更接近真实答案，且答案语言结构简单，扣题严谨，没有语言冗杂。经多次试验问答系统均可稳定返回最接近真实答案的回答，由此可知本研究搭建的基于知识图谱的农作物病虫害知识问答系统能够在一定程度上满足用户对于农作物病虫害知识的问答需求。



图 14　问答结果对比图

Figure 14　Comparison of question answering results

4　结论

结合大数据时代的发展趋势，智能问答系统或专家系统在为用户提供快速准确答案方面发挥着重要作用，其在各领域的广泛应用已成为现实。与传统文本知识收集或一般网络搜索引擎检索相比，智能问答系统能够有效保障用户知识检索速度和质量。而随着知识图谱技术的迅速发展和成熟，成为了问答系统的重要数据基础。农业智能问答系统充分展示了智慧农业的推广与农业大数据化普及水平。本研究的重点在于构建基于知识图谱的农作物病虫害知识问答系统，通过建立专业领域的知识图谱，结合 BERT-BiLSTM-CRF 模型和 BERT-RNN 模型，分别完成实体识别和问题分类任务。最终，结合 Flask 框架实现了系统的可视化与实用性，使用户能够方便地使用该系统进行农作物病虫害知识的查询和获取。

在模型试验中,BERT-BiLSTM-CRF 模型和 BERT-RNN 模型,分别完成实体识别和问题分类任务。在评价指标方面,实体识别任务取得了 93.22% 的精确率(P)、92.69% 的召回率(R)和 92.21% 的 $F1$ 值,而问题分类任务取得了 94.37% 的 P、93.92% 的 R 和 92.66% 的 $F1$ 值,结果表现出色。随后在系统的测试实验中,问答系统返回答案语言结构简单,扣题严谨,没有语言冗杂,多次试验问答系统均可稳定返回最接近真实答案的回答。经测试本研究的问答系统在回答用户农作物病虫害有关问题的过程中展现出优秀的准确性与稳定性,在一定程度上满足了用户的问答需求。

本研究基于知识图谱的农作物病虫害知识问答系统的设计和实现,为农业领域智能问答系统的发展提供了一个有益的探索和范例。未来,我们将继续致力于多模态数据集的研究与构建,多语言支持和跨领域扩展等方面的研究,以进一步提升系统的功能和性能,不断完善基于知识图谱的农作物病虫害知识问答系统,并为智能问答系统的发展提供更多可能性和机遇。

参 考 文 献

[1] GUANXI Y, THIERRY B, HABIBULLAH B, et al. The technical challenges for applying unsaturated soil sensors to conduct laboratory-scale seepage experiments [J]. Sensors, 2022, 22(10): 3724－3724.

[2] 韩文霆, 李广, 苑梦婵, 等. 基于无人机遥感技术的玉米种植信息提取方法研究[J]. 农业机械学报, 2017, 48(1): 139－147.
HAN Wenting, LI Guang, YUAN Mengchan, et al. Extraction method of maize planting information based on UAV remote sensing technology [J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(1): 139－147.

[3] AMINE F, MOHAMED S, ESSAID S. An agile AI and IoT-augmented smart farming: A cost-effective cognitive weather station [J]. Agriculture, 2022, 12(1): 35.

[4] 苗小爱. 基于大数据分析的农业信息化研究[J]. 智慧农业导刊, 2023, 3(7): 4－6,10.
MIAO Xiaoai. Research on agricultural informatization based on big data analysis [J]. Journal of Smart Agriculture, 2023, 3(7): 4－6, 10.

[5] 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 Chinese Society of Agricultural Engineering, 2020, 36(24): 177－185.

[6] 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.

[7] POPPING R. Knowledge graphs and network text analysis [J]. Social Science Information, 2003, 42(1): 91－106.

[8] 陈佳云, 徐向英, 章永龙, 等. 多模态知识图谱在农业中的研究进展[J]. 农业大数据学报, 2022, 4(3): 126－134.
CHEN Jiayun, XU Xiangying, ZHANG Yonglong, et al. Research progress of multimodal knowledge graph in agriculture [J]. Journal of Agricultural Big Data, 2022, 4(3): 126－134.

[9] 徐畅, 张领先, 乔岩. 基于电子病历的作物病虫害关联挖掘及智能诊断[J]. 智能化农业装备学报(中英文), 2023, 4(4): 1－10.
XU Chang, ZHANG Lingxian, QIAO Yan. Association mining of crop diseases and intelligent diagnosis based on electronic medical records [J]. Journal of Intelligent Agricultural Mechanization, 2023, 4(4): 1－10.

[10] 王丹丹. 宁夏水稻知识图谱构建方法研究与应用[D]. 银川: 北方民族大学, 2020.
WANG Dandan. Research and application of construction method of rice knowledge graph in Ningxia [D]. Yinchuan: North Minzu University, 2020.

[11] 张永亮. 基于知识图谱的苹果病虫害智能问答系统研究[D]. 杨凌: 西北农林科技大学, 2022.
ZHANG Yongliang. Research on intelligent question answering system of apple diseases and insect pests based on knowledge graph [D]. Yangling: Northwest A&F University, 2022.

[12] 闫丽华. 基于知识图谱的葡萄病虫害自动问答系统[D]. 杨凌: 西北农林科技大学, 2021.
YAN Lihua. Automatic question answering system of grape diseases and pests based on knowledge graph [D]. Yangling: Northwest A&F University, 2021.

[13] 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.
LIU Qiao, LI Yang, DUAN Hong, et al. Knowledge graph construction techniques [J]. Journal of Computer Research and Development, 2016, 53(3): 582－600.

[14] 吴赛赛. 基于知识图谱的作物病虫害智能问答系统设计与实现[D]. 北京: 中国农业科学院, 2021.
WU Saisai. Design and implementation of intelligent

question answering system for crop diseases and pests based on knowledge graph [D]. Beijing: Chinese Academy of Agricultural Sciences, 2021.

[15] 张嘉宇，郭玫，张永亮，等. 细粒度苹果病虫害知识图谱构建研究[J]. 计算机工程与应用，2023，59(5)：270－280.
ZHANG Jiayu, GUO Mei, ZHANG Yongliang, et al. Research on construction of fine-grained knowledge graph of apple diseases and pests [J]. Computer Engineering and Applications, 2023, 59(5): 270－280.

[16] LOURDUSAMY R, MATTAM X J. Knowledge graph using resource description framework and connectionist theory [C]//IOP Publishing Ltd., 2020: 012001. DOI: 10.1088/1742－6596/1427/1/012001.

[17] 李雪. 一种基于Neo4j图数据库的模糊查询研究与实现[J]. 计算机技术与发展，2018，28(11)：78－86.
LI Xue. Research and implementation of a fuzzy query based on Neo4j graph database [J]. Compute Technology and Development, 2018, 28(11): 78－86.

[18] 王婷，王娜，崔运鹏，等. 基于人工智能大模型技术的果蔬农技知识智能问答系统[J]. 智慧农业(中英文)，2023，5(4)：105－116.
WANG Ting, WANG Na, CUI Yunpeng, et al. Agricultural technology knowledge intelligent question-answering system based on large language model [J]. Smart Agriculture, 2023, 5(4): 105－116.

[19] ZHANG Q, YU H Q, SUN P, et al. Multisource data based agricultural drought monitoring and agricultural loss in China [J]. Global and Planetary Change, 2018, 172 (JAN.): 298－306.

[20] 白皓然，孙伟浩，金宁，等. 基于改进Bi-LSTM-CRF的农业问答系统研究[J]. 中国农机化学报，2023，44(2)：99－105.
BAI Haoran, SUN Weihao, JIN Ning, et al. Research on agricultural question answering system based on improved Bi-LSTM-CRF [J]. Journal of Chinese Agricultural Mechanization, 2023, 44(2): 99－105.

[21] ZHANG Y, ZHANG M L, LUO N, et al. Understanding the formation mechanism of high-quality knowledge in social question and answer communities: A knowledge co-creation perspective [J]. International Journal of Information Management, 2019, 48: 72－84.

[22] 郑泳智，吴惠粦，朱定局，等. 基于荔枝和龙眼病虫害知识图谱的问答系统[J]. 计算机与数字工程，2021，49(12)：2618－2622.
ZHENG Yongzhi, WU Huilin, ZHU Dingju, et al. Question and answer system based on the knowledge graphs of Litchi and Longan diseases and insect pests [J]. Computer & Digital Engineering, 2021, 49(12): 2618－2622.

[23] 赵瑞雪，杨晨雪，郑建华，等. 农业智能知识服务研究现状及展望[J]. 智慧农业(中英文)，2022，4(4)：105－125.
ZHAO Ruixue, YANG Chenxue, ZHENG Jianhua, et al. Agricultural intelligent knowledge service: Overview and future perspectives [J]. Smart Agriculture, 2022, 4(4): 105－125.

[24] 李源，丁士宁，张克旺，等. 基于CiteSpace的水稻病虫害研究热点与发展趋势分析[J]. 农业展望，2022，18(5)：97－104.
LI Yuan, DING Shining, ZHANG Kewang, et al. Research hotspots and development trend of rice disease and insect pests based on CiteSpace [J]. Agricultural Outlook, 2022, 18(5): 97－104.

Research on question answering system for crop diseases and pests based on knowledge graph

ZHAO Zexing[1], WU Xiaopeng[1], WANG Yixin[2], YAN Xiaoli[1], HUANG Yuxiang[1], GAO Xiaojun[3, 4*]

(1. *College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling* 712100, *China*; 2. *College of Information Engineering, Northwest A&F University, Yangling* 712100, *China*; 3. *College of Engineering, Nanjing Agricultural University, Nanjing* 210031, *China*; 4. *National Key Laboratory of Agricultural Equipment Technology, Beijing* 100083, *China*)

Abstract: Propelled by the era of big data, intelligent question-answering systems have found widespread applications across various domains, offering users a means to swiftly and efficiently acquire answers. Compared to traditional methods of

collecting textual knowledge and utilizing web search engines, these systems demonstrate distinct advantages. With the rapid evolution of knowledge graph technology, intelligent question-answering systems have entered a new stage of development. In response to the call for smart agriculture, this study has undertaken the construction of a knowledge-based question-answering system based on knowledge graphs, aiming to provide users with question-answering services related to crop diseases and pests. The primary tasks include: (1) Acquisition of crop disease and pest data: Employing a distributed web crawling framework to retrieve data from web pages related to crop diseases and pests and preprocessing operations involve data cleaning, analysis, and structuring. (2) Construction of knowledge graph: Following data analysis, defining entity and relationship categories for the knowledge ontology and completing the pattern layer construction of the knowledge graph. Utilizing rule-based triple templates for extracting entities from semi-structured text, building the data layer, and storing triples in the Neo4j graph database. (3) Design of question-answering algorithms: Employing the BERT-BiLSTM-CRF model for question entity recognition and the BERT-RNN model for question classification. Matching templates, executing queries through Cypher statements, and processing answers into natural language forms for return. (4) Implementation and visualization of the question-answering system: Integrating crop disease and pest knowledge graphs with question-answering algorithms, the Flask framework and various web technologies are used to implement functionalities such as user questioning, entity recognition, knowledge retrieval, and answer presentation. Experimental results indicate that the entity recognition and question classification models achieved Precision (P), Recall (R), and $F1$ scores of 93.22%, 92.69%, 92.21%, and 94.37%, 93.92%, 92.66%, respectively. Compared to other search methods, the question-answering system demonstrated higher accuracy and stability. This study provides an intelligent solution for agricultural informationization, offering a new pathway for users to acquire knowledge on crop diseases and pests.

Keywords: agricultural informationization; crop disease and pest; knowledge graph; question and answering system