doi:10.6041/j.issn.1000-1298.2025.12.050

基于多源异构数据和知识图谱的作物病害关联分析

王　波[1]　丁俊琦[1]　吴奇峰[2]　张领先[1]

(1.中国农业大学信息与电气工程学院,北京 100083;2.新疆农垦科学院,乌鲁木齐 832003)

摘要:多源异构农业信息的整合不仅有助于理解和揭示病害侵染与防控路径,也可为病害处方推荐研究提供重要的数据支持。针对电子病历等多源异构农业数据解析中的融合、对齐和异质性问题,本文构建作物病害处方知识图谱,可视化解析病害关联关系。基于病三角原理,解析病原体、宿主及环境在病害侵染与防治过程中的关键作用,结合数据特点构建包含 18 类本体概念、17 条关系和 6 条属性边的作物病害处方知识图谱本体层;结合基于规则匹配与基于深度学习的知识抽取方法构建知识图谱实体层,共包含 1 121 个实体实例和 8 292 条关系实例;分别开展基于度中心性和介数中心性的 Top20 关键节点识别和基于 Adamic－Adar 的病害-防控产品 Top5 关联预测,可视化解析病害、症状、防控产品等关键实体与属性的关联关系,并从防治方案选择、绿色补贴过滤和相关信息查询 3 方面制定了优化病害防控产品推荐和信息检索功能的 6 条规则。研究成果可为基于电子病历数据挖掘的作物病害多目标决策提供支撑。

关键词:作物病害;电子病历;病三角原理;多源异构数据;知识图谱;关联分析

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1000-1298(2025)12-0546-14　　**OSID**:

Association Analysis of Crop Diseases Based on Multi-source Heterogeneous Data and Knowledge Graph

WANG Bo[1]　DING Junqi[1]　WU Qifeng[2]　ZHANG Lingxian[1]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *Xinjiang Acadamy of Agricultural and Reclamation Sciences*, *Urumqi* 832003, *China*)

Abstract: The integration of multi-source heterogeneous agricultural information not only help understanding and unveiling the pathways of diseases infection and prevention but also offers vital data support for research on diseases prescription recommendations. Aiming at the problems of fusion, alignment and heterogeneity in the analysis of multi-source heterogeneous agricultural data such as electronic medical records, the knowledge graph of crop diseases prescription was constructed, and the diseases association on this basis was visually analyzed. Initially, starting from the principle of the disease triangle, the critical roles of pathogens, hosts, and environment in the process of disease infection and control were analyzed, constructing the ontology layer of the crop diseases prescription knowledge graph with 18 ontology concepts, 17 relationships, and 6 attribute edges based on data characteristics. Subsequently, the entity layer of the knowledge graph was built by integrating rule-matching and deep-learning knowledge extraction methods, encompassing 1 121 entity instances and 8 292 relationship instances. Lastly, identification of Top20 key nodes based on degree and betweenness centrality, along with Top5 disease-prevention product association predictions using the Adamic－Adar index, were conducted to visually analyze the associations between key entities and attributes, including diseases, symptoms, and prevention products. Six rules were established to enhance the recommendation and information retrieval functions of disease prevention products from three perspectives: prevention scheme selection, green subsidy filtering, and related information inquiry. The research result can provide reference for electronic medical record data mining and association analysis of crop diseases.

Key words: crop diseases; electronic medical records; disease triangle principle; multi-source heterogeneous data; knowledge graph; association analysis

收稿日期:2024－08－23　修回日期:2024－10－21

基金项目:国家自然科学基金项目(62376272)、2024 年中国高校产学研创新基金-云中大学专项(二期)课题(2024MU050)和新疆农垦科学院农业科技创新工程专项(NCG202507)

作者简介:王波(2000—),男,博士生,主要从事农业信息技术和自然语言处理研究,E-mail:1601207481@qq.com

通信作者:张领先(1970—),男,教授,博士生导师,主要从事农业信息化研究,E-mail:zhanglx@cau.edu.cn

0　引言

大量积累的电子病历及作物病害相关资料蕴含了丰富的作物、环境、病原、病害和处方关联信息。这些信息的抽取和整合不仅有助于理解和揭示病害侵染与防控路径，也为病害处方推荐研究提供重要的数据支持。然而，数据的异质性和分散性对其集成与应用造成困难。如何从多源异构的农业数据中有效抽取知识，构建系统化的知识体系，成为了农业信息化领域亟待解决的问题。

传统的数据整合方法如数据仓库和数据湖提供了一定程度的解决方案[1]。数据仓库[2]适用于结构化数据整合，需要大量的前期准备和维护成本，且不易适应数据源的快速变化；而数据湖[3]虽能快速存储大量结构化与非结构化数据，却面临数据管理和质量控制问题。知识图谱[4]通过构建实体之间的关系网络，能够有效整合不同数据源的信息，并提供一个直观且灵活的关联分析和知识检索框架。相比于传统的关系型数据库，基于图的知识表示更直观，能够揭示实体间复杂的关系。

知识图谱已在医疗[5-6]、中医[7-8]等行业得到应用，在农业领域也展现出了巨大的潜力。已有研究人员基于领域文献等文本数据等开展农业实体、关系识别研究[9-11]，或构建水稻栽培[12]、茶叶种植[13]、棉花病虫害检测[14]等领域知识图谱。目前农业领域知识图谱构建研究多聚焦于构建过程中的某一个重要环节，如本体建模[15]、信息抽取[16]、知识融合[17]、知识加工[18]等；农业知识图谱应用研究集中于信息检索[19]、问答系统[20]、推荐系统[21]、专家诊断系统[22]、作物预测[23]等。还未见以作物病害处方为核心构建知识图谱，一方面，来自电子病历及作物病害相关资料的多源异构数据转换为结构化知识的难度大，且需要大量的人力资源；另一方面，难以寻求知识图谱和作物病害处方的结合点。另外，目前国外关于疾病-处方关联分析的相关研究主要聚焦于生物医学领域[24-29]。国内关于疾病-处方关联分析的研究主要聚焦于中医领域[30-31]。目前未见作物病害领域开展疾病-处方关联分析研究，借鉴国内外医学领域相关研究经验，基于构建的作物病害处方知识图谱可开展作物病害关联挖掘研究。

综上，疾病-处方关联分析在生物医学和中医领域已经取得了很多成果，基于知识图谱的实体关联网络构建方法在农业领域也有众多研究，实现了针对文本数据的农业知识有效抽取，但是难以应对作物电子病历等多源异构农业数据的分散性、异质性问题，对于以作物病害处方为核心的知识图谱也缺乏研究。因此，本文从病三角原理出发，构建作物病害处方知识图谱并开展关联分析，以揭示病害关键因素及防控路径。

1　数据获取与预处理

1.1　作物病害多源数据获取

（1）电子病历数据。2012 年，北京市在全国首次引入国际应用生物科学中心植物诊所的先进理念[32]，建立了包括市区乡（村）的三级植物诊所。北京市植保站依托所建立的 115 家植物诊所陆续获取了 2017 年 10 月 30 日至 2021 年 7 月 13 日的病害问诊记录数据，共计 35 859 条。这些数据中不仅有结构化的信息，如就诊时间、诊所位置和发病时间等，还包括非结构化的自由文本信息，如主要症状、诊断结果和治疗建议等。

（2）北京市农作物病虫害绿色防控产品推荐名录。该推荐名录包括农药产品、天敌产品、理化诱控产品、物理阻隔产品、授粉昆虫和植保机械共 296 项产品，记录了产品的登记证号、生产企业、防治对象等信息。

（3）植保站病害防治明白纸。北京市植保站动态开发和修订，至今共完成了覆盖全市 22 种主要作物在内的 127 种病害的明白纸，以农业防治和高效、低毒的化学农药等绿色防控技术为主，基本实现了对主要作物常见病害问题的全覆盖。

（4）北京市推广应用绿色防控产品绿色补贴记录。记录了 2020 年 3 月—2021 年 7 月北京市推广应用绿色防控产品绿色补贴情况，包括对多家公司的绿色防控产品销售的绿色补贴情况。具体数据涵盖每一笔应用绿色防控产品销售绿色补贴的供应商和经销商等信息。

（5）国家农业科学数据中心网站。为了获取更多关于作物病害的属性、侵染机理和防治方案的信息，从国家农业科学数据中心获取数据。

1.2　作物病害电子病历数据预处理

针对电子病历结构化数据部分以及北京市农作物病虫害绿色防控产品推荐名录等半结构化数据，首先将数据解析成更容易处理的格式，如结构化文档转换为电子表格；其次进行数据清洗，采用人工观察、统计方法、可视化方法等检测异常值，通过删除、插值或使用默认值填充等操作处理异常值和缺失值；再次进行特征工程，包括一致化处理（如将日期数据转换为年月日格式）和特征提取，去除无用或隐私数据；最后执行数据去重操作，删除数据集中的重复记录以确保数据唯一性。

针对电子病历文本数据，首先通过正则表达式

提取出电子病历中的文本部分;然后进行文本清洗:去除多余的空格和换行符,去除无关字符、特殊符号和 HTML 标签等无用信息,以及去除“的”、“了”等停用词以减小文本数据维度;最后进行同义词替换等文本标准化操作,例如将口语化表达替换为规范用语。

2 作物病害关联关系与推理规则分析

2.1 作物病害关联关系与推理规则分析流程

本研究以病三角原理为理论基础,构建作物病害处方知识图谱,基于构建的知识图谱进行关联关系与推理规则分析,主要流程如图 1 所示。

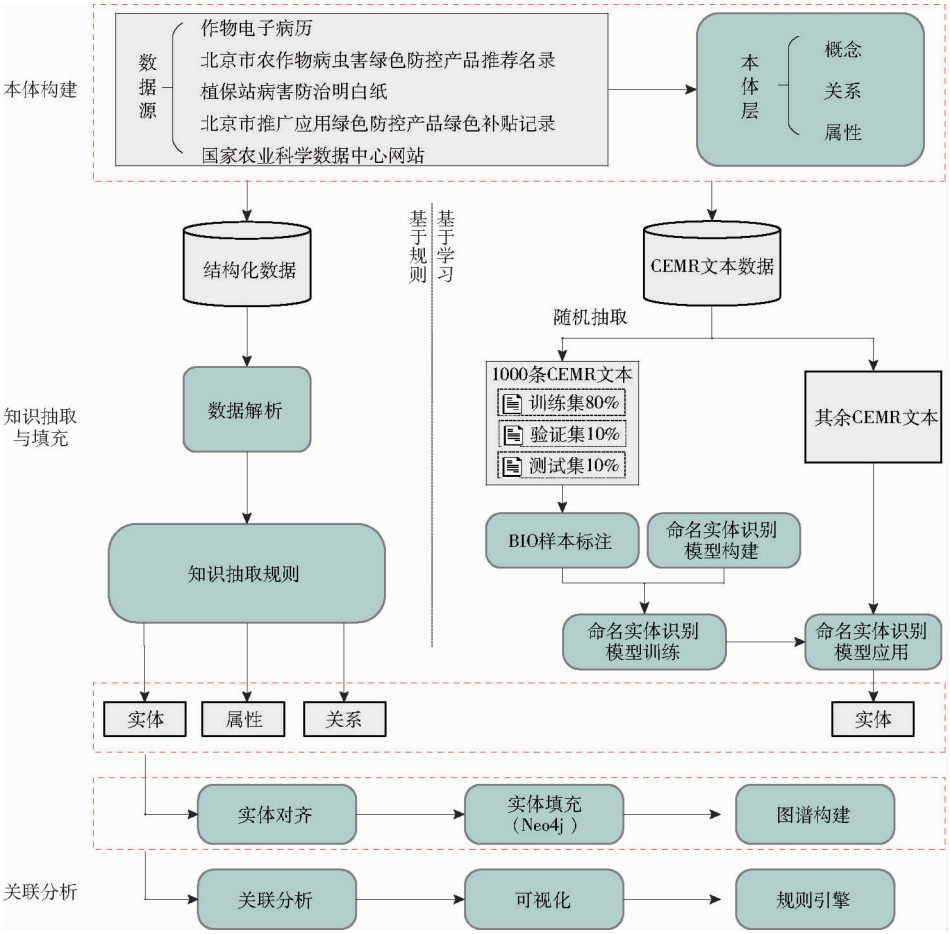


图 1 作物病害关联关系与推理规则分析流程

Fig. 1 Analysis process of crop disease association and reasoning rules

2.2 作物病害处方知识图谱本体层构建

(1)病三角原理。病三角原理阐述了病原体、宿主植物以及环境三者之间的关系是如何共同作用,引发并影响病害的发生和扩散[33-34],为制定有效的病害治疗方案提供了理论基础(图 2)。病原体是作物病害的传染源,包括病毒、细菌、真菌和线虫

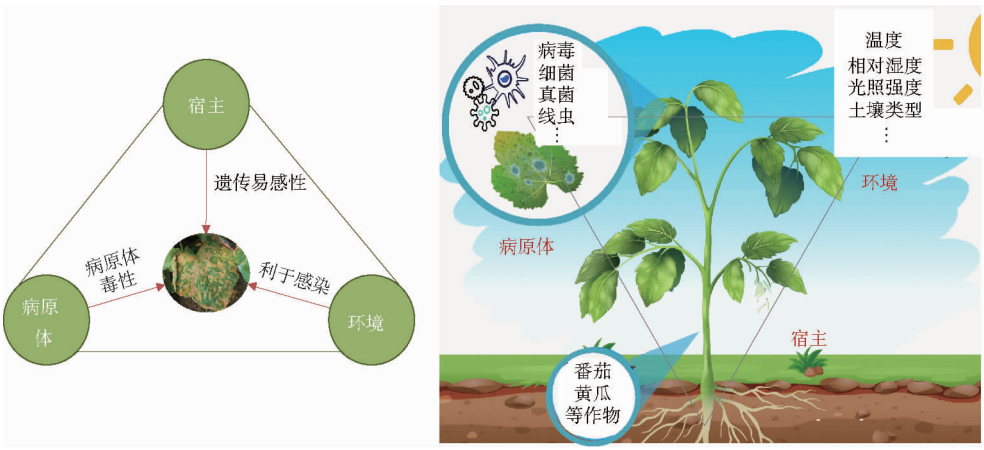


图 2 病三角原理

Fig. 2 Disease triangle principle

等微生物和生物体。根据不同的病原体种类，宿主植物会表现出不同的症状；宿主是指受到病害影响的植物，其易感性程度对病害的发生至关重要。考虑不同植物在不同生长阶段的抗病性变化；环境条件包括温度[35]、相对湿度[36]、光照强度、土壤类型等因素，对病原体的生长、传播及宿主植物的抵抗力有重要影响。大量研究证明了作物病害的环境依赖性[37]，但环境因素在现有智能病害管理研究中时常被忽视。

根据病三角原理，在制定病害治疗策略时需要全面考量这 3 个要素，这也从作物电子病历中体现出来。电子病历记录了每次植物医生问诊过程中的时间地点、病害发生情况、植物医生提供的诊断结果和有针对性的病害防治方案等重要信息，这些信息反映了病原体的侵染表征、作物的遗传易感性和病害发生的外界环境之间的复杂关系，即“病原体-宿主-环境”关联关系。因此，在整合与提取多源数据中的病害防控知识时，从病三角原理出发，揭示病害关键影响因素并形成知识网络，最终构建作物病害处方知识图谱。

（2）本研究采用自顶向下与自底向上相结合的方法，以病害-处方关联为核心，一方面从病害三角原理出发确定“病原体-宿主-环境-病害-处方”的核心范围框架（自顶向下），另一方面从数据出发，明确框架范围内的实体及处方相关实体（自底向上）。最终构建的本体层如图 3 所示。①概念。作物病害处方

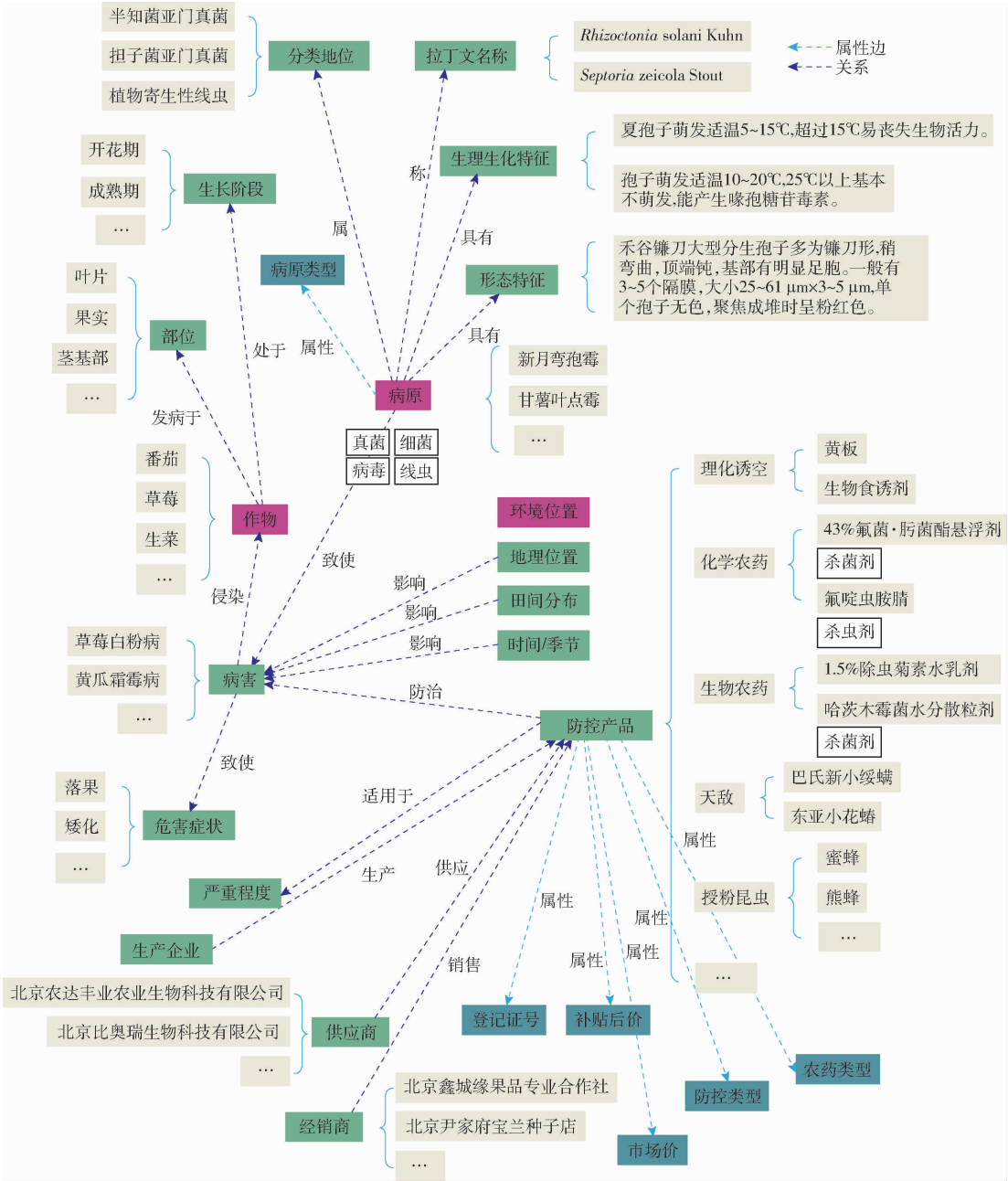


图 3　知识图谱本体层示意图

Fig. 3　Schematic of ontology layer of knowledge graph

知识图谱本体层以病害和防控产品作为核心概念,根据病三角原理和多源数据特征向外扩展,共纳入 18 类本体概念。与病害相关联的概念覆盖了作物、病原和环境相关的 12 个类别:作物、部位、生长阶段、病原、分类地位、拉丁文名称、生理生化特征、形态特征、危害症状、地理位置、田间分布、时间/季节;与防控产品相关联的概念包括严重程度、供应商、生产企业、经销商 4 个类别。②关系。农作物病害处方知识图谱定义了 17 条不同类型的直接关系,包括防治、适用于、影响、侵染、致使、属等。在知识图谱中,通过这些直接定义的关系可以进一步推理出新的关系链条。这些直接和间接关系构成知识图谱的关系网络,实现知识的串联和推导。③属性。作物病害处方知识图谱本体层共定义了 6 条属性边:病原概念有病原类型这 1 条属性边,防控产品概念有防控类型、市场价、绿色补贴后价和登记证号 4 条属性边。

2.3 作物病害处方知识抽取

由于数据包括多个数据来源且包含结构化、半结构化和非结构化几种不同类型,因此在知识抽取环节采用基于规则匹配和基于深度学习 2 种方法以适应不同的数据类型。表 1 展示了从各个数据源中待抽取信息的类型、特点和数据量。

表 1 多源异构数据信息解析

Tab. 1 Analysis of multi-source heterogeneous data information

数据源	待抽取信息	数据类型	数据量
电子病历	作物、部位、生长阶段、病害、危害症状、地理位置、田间分布和时间/季节、防控产品	结构化 + 文本	35 859 条
北京市农作物病虫害绿色防控产品推荐名录	防控产品、防控类型、生产企业、防治对象、登记证号	半结构化	212 条
植保站病害防治明白纸	作物、病害、防控产品	半结构化	127 张
北京市推广应用绿色防控产品绿色补贴记录	防控产品、防控类型、市场价、绿色补贴后价、供应商、经销商	结构化	5 036 条
国家农业科学数据中心网站	病害、病原、分类地位、拉丁文名称、生理生化特征、形态特征、危害症状	半结构化	147 条

在对 5 种数据进行知识抽取后,以病害为桥梁,将电子病历数据、植保站病害防治明白纸和国家农业科学数据中心网站数据进行融合;以防控产品为桥梁,将电子病历数据、北京市农作物病虫害绿色防控产品推荐名录和北京市推广应用绿色防控产品绿色补贴记录进行融合。数据对齐与整合过程如图 4 所示。

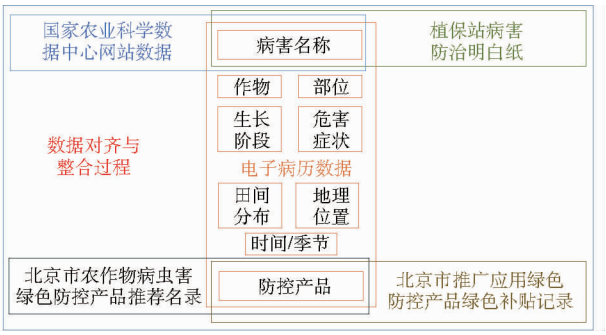


图 4 数据对齐与整合过程

Fig. 4 Data alignment and integration process

2.3.1 基于深度学习的作物病害处方知识抽取

对于文本形式的非结构化数据,即电子病历中的问诊记录部分,采用基于深度学习的方法进行知识抽取,再与电子病历的结构化数据部分进行信息融合。电子病历问诊记录部分是非结构化数据,对应着一条完整的电子病历样本,其它结构化数据部分的实体和关系已经确定,对于非结构化数据只需识别一些实体进行补充。本研究基于 BERT - BiLSTM - CRF 模型对非结构化数据进行命名实体识别。

(1)BIO 数据标注。B(Begin):表示实体的开始部分。例如,在标注一个作物“黄瓜”时,“黄”字会被标注为“B - CROP”。I(Inside):表示实体的内部部分。继续上述例子,“瓜”字会被标注为“I - CROP”。O(Outside):表示非实体部分。采用人工标注的方式进行语料库标注,数据标注实体分为 3 种类别:作物、症状、部位。随机选取电子病历样本中的 1 000 条数据,标注其中的问诊记录部分。共标注数据 1 000 条,按照比例 8∶1∶1 划分训练集、验证集、测试集。

(2)命名实体识别模型构建。采用 BiLSTM - CRF 模型作为命名实体识别的基准模型,在此基础上融合 BERT 预训练语言模型,构建基于 BERT 字嵌入向量的 BiLSTM - CRF 模型来抽取电子病历非结构化数据中的实体。如图 5 所示,模型结构包含字嵌入层、BiLSTM 层和 CRF 层 3 个模块。利用 BERT 预训练模型获得句子中每个字的嵌入向量;将获取的向量输入 BiLSTM 模型进行进一步语义编码,通过不同方向的文本表示充分获取目标文本序

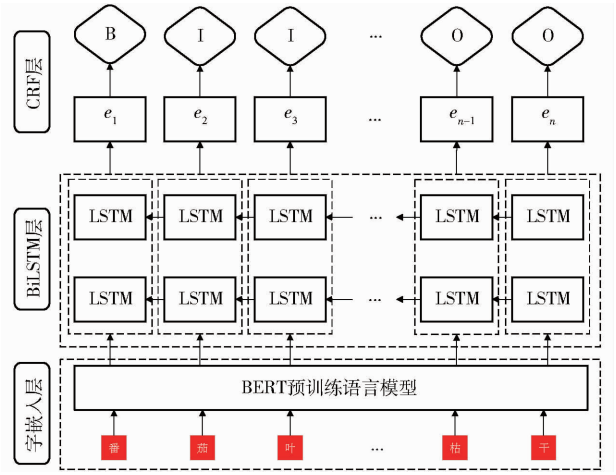


图 5　命名实体识别模型

Fig. 5　Named entity recognition model

列特征信息；通过 CRF 层输出概率最优标签序列。在 CRF 层，采用状态转换矩阵来预测当前标签，获得全局最优的标记序列。最后，使用维特比算法（Viterbi）将得分最高的序列 y^* 作为模型最终的标注结果。

2.3.2　基于规则匹配的作物病害处方知识抽取

对于植保站病害防治明白纸、北京市推广应用绿色防控产品绿色补贴记录等结构化和半结构化数据，基于规则匹配的方法通过预定义的模式匹配规则来识别关键字段，如防控产品的登记证号、市场价和绿色补贴后价等。具体的规则要根据不同数据源的固定格式和标签分别设计，以确保高准确性的信息抽取。以防控产品的实体抽取为例，防控产品名称通常包括化学成分和剂型，以及浓度或活性成分含量的百分比表示，据此设定匹配的正则表达式如表 2 所示。

表 2　防控产品正则表达式

Tab. 2　Prevention and control product regular expression

类型	正则表达式	示例
百分比浓度加化学成分名称和剂型	\d + % [\u4e00 - \u9fa5] + (?: 可溶液剂 | 悬浮剂 | 水乳剂 | 可溶粉剂 | 微乳剂 | 水分散粒剂 | 可湿性粉剂 | 颗粒剂 | 乳油)	5% 桉油精可溶液剂
量词加“孢子/毫升”或“CFU/克”等单位	\d + [\u4e00 - \u9fa5] * 孢子/毫升 | \d + [\u4e00 - \u9fa5] * CFU/克	100 亿孢子/毫升短稳杆菌悬浮剂
特定成分名称后跟剂型和倍数表示	[\u4e00 - \u9fa5] + (?: \· [\u4e00 - \u9fa5] +) * \s * (?: 可溶液剂 | 悬浮剂 | 水乳剂 | 可溶粉剂 | 微乳剂 | 水分散粒剂 | 可湿性粉剂 | 颗粒剂 | 乳油) \s * \d * 倍？液？	1.8% 虫菊·苦参碱水乳剂

2.4　作物病害知识图谱构建

知识图谱的构建策略包括自顶向下（Top - Down）和自底向上（Bottom - Up）2 种方式（图 6）。自顶向下的知识图谱构建策略先定义高层的概念和分类，然后逐步细化到具体的实体和关系。自底向上的知识图谱构建策略从具体的数据和信息开始，然后逐步抽象和归纳出高层的概念和结构，其优势在于快速、全面抽取数据中的关键信息。将二者结合的混合策略需要基于数据特征构建初步的模式层，并在数据挖掘中不断更新模式层，最后进行实体的填充形成最终的知识图谱。这种策略构建的知识图谱结构既有良好的结构又包含丰富的数据，更能贴合数据本身的特性，因此本文采用这种策略构建作物病害处方知识图谱。

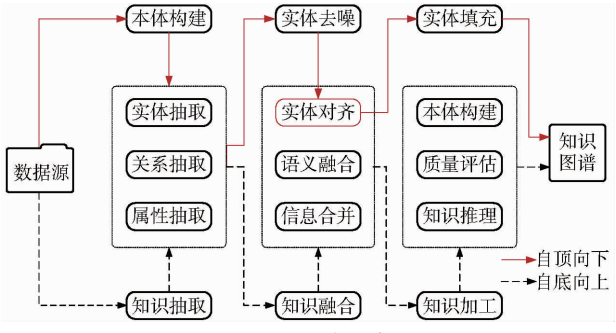


图 6　知识图谱构建方式

Fig. 6　Construction of knowledge graph

2.5　作物病害关联分析算法

（1）基于中心性度量算法的关键节点识别。采用度中心性（Degree centrality，DC）和介数中心性（Betweenness centrality，BC）2 种算法衡量作物病害处方知识图谱中心性。

度中心性是图关联分析中最直观的中心性度量指标，通过节点的度衡量节点重要性，即该节点直接相连边的数量。对于总数为 g 的网络，节点 i 的度中心性 C_D 表示为它与其它 $g-1$ 个节点的直接联系总数，其计算公式为

$$C_D(N_i) = \sum_{j=1}^{g} x_{ij} \quad (i \neq j) \tag{1}$$

式中　N_i——第 i 个节点

x_{ij}——与第 i 个节点直接相连的边的数量

介数中心性 C_B 用于衡量 1 个顶点出现在其他任意 2 个顶点对之间的最短路径的次数，其计算公式为

$$C_B(N_i) = \sum_{j,k=1}^{g} s_d(j,i,k) \quad (j \neq k) \tag{2}$$

其中,$s_d(j,i,k)$表示节点 j 到节点 k 的最短路径经过了节点 i,即节点 i 处于节点 j 到节点 k 的最短路径上。

(2)基于 Adamic - Adar 的病害-防控产品关联预测方法。对于知识图谱中潜在的病害-防控产品关联,设计基于 Adamic - Adar 的病害-防控产品关联预测方法。主要步骤为:对于病害 a,计算其与其他病害节点之间的 Adamic - Adar 指数并排序,获取最相关的病害 b;查询病害 b 在作物病害处方知识图谱中关联的防控产品 c,并在病害 a 与防控产品 c 之间建立预测关联;对于防控产品 c,计算其与其他防控产品节点之间的 Adamic - Adar 指数并排序获取最相关的防控产品 d;在病害 b 与防控产品 d 之间建立预测关联,以及病害 a 与防控产品 d 间建立预测关联。

Adamic - Adar 是一种衡量节点间相似性的链路预测算法。Adamic - Adar 指数基于 2 个节点的共同邻居节点数量和权重预测二者之间关系,其中权重取决于共同邻居节点的度数。节点 i 和节点 j 之间的 Adamic - Adar 指数 AA 计算式为

$$AA(i,j)=\sum_{z\in\Gamma(i)\cap\Gamma(j)}\frac{1}{\lg(|\Gamma(z)|)} \tag{3}$$

式中 $\Gamma(i)$、$\Gamma(j)$——与节点 i、j 相邻的节点集合

$\Gamma(z)$——与节点 z 相邻的节点集合

2.6 作物病害规则引擎制定

基于作物病害处方知识图谱可以实现病害防控产品推荐和信息检索,为此制定以下规则引擎(表 3)以提升复杂查询需求的检索准确性和效率,优化用户体验。

表 3 基于规则的匹配算法

Tab.3 Rule-based matching algorithm

规则	规则名称	Cypher 查询语言
防治方案选择	规则 1	MATCH(n) -[r:'防治'] ->(m {name:'某作物病害'}) WHERE '生物农药' IN n.'防控类型` RETURN n, m
绿色补贴过滤	规则 2A	MATCH(n) -[r:'防治'] ->(m {name: '某作物病害'}) WHERE exists(n.'绿色补贴后价') RETURN n, m
	规则 2B	MATCH(n) -[r:'防治'] ->(m {name: '某作物病害'}) RETURN n, m
相关信息查询	规则 3A	MATCH (pathogen:病原) -[:'致使'] -> (d {name: '某病害'}) OPTIONAL MATCH (pathogen) -[: '属'] ->(classification: '分类地位') OPTIONAL MATCH (pathogen) -[: '称'] ->(latin: '拉丁文名称') OPTIONAL MATCH (pathogen) -[: '具有'] ->(traits: '生理生化特征') OPTIONAL MATCH (pathogen) -[: '具有'] ->(morphology: '形态特征') OPTIONAL MATCH (pathogen) -[:'致使'] ->(symptoms: '危害症状') RETURN pathogen, pathogen.'病原类型', classification, latin, traits, morphology, symptoms
	规则 3B	MATCH(q) -[p:'生产'|'销售'|'供应'] ->(m {name: '某防控产品'}) RETURN p, m, q
	规则 3C	MATCH (m {name: '某防控产品'}) RETURN m.'防控类型', m.'登记证号', m.'市场价', m.'绿色补贴后价'

(1)防治方案选择规则。如果用户根据个人需求选择特定的防治方案,如生物农药/化学农药/理化诱空/天敌/授粉昆虫,查询返回所有能够治疗该病害的具有相应防空类型属性的防控产品(规则 1)。

(2)绿色补贴过滤规则。为满足不同用户对绿色补贴产品的需求差异,制定绿色补贴过滤规则。如果用户想要查看有绿色补贴的防控产品,查询需要过滤出只有绿色补贴的产品(规则 2A);如果用户对是否享有绿色补贴没有特定要求,则返回所有相关防控产品信息(规则 2B)。

(3)相关信息查询规则。针对病害、防控产品以及供应链信息的不同查询需求,制定 3 个子规则。如果用户要获取某病害相关的病原知识,当查询到该病害时,返回该病害相关的病原节点、病原类型属性以及病原相关的分类地位、拉丁文名称、生理生化特征、形态特征和危害症状等节点(规则 3A);如果用户想要了解某防控产品的供应链信息,返回与之关联的生产企业、供应商和经销商等节点(规则 3B);如果用户想要获取某防控产品的更多信息,返回其所有标签信息,如防控类型、登记证号、市场价和绿色补贴后价(规则 3C)。

3 实验与结果分析

3.1 实验环境与模型参数

实验硬件环境为 Intel(R) Xeon(R) Bronze 3106 CPU@1.70 GHz 处理器,32 GB 内存,操作系统为 Windows 10,GPU 为 NVIDIA GeForce RTX 2 080 Ti (11 GB)。在训练过程中,经多次调试与训练,BERT - BiLSTM - CRF 模型参数设置如表 4 所示。

本研究采用精确率(Precision,P)、召回率(Recall,R)和 F1 值来评价命名实体识别效果。

表 4　模型参数设置

Tab. 4　Model parameters setting

参数	数值/类型
序列最大长度	256
训练轮次	50
选择批处理大小	32
梯度下降算法	AdamW
随机失活率	0.5
学习率	0.03
LSTM 隐藏层	128

表 5　命名实体识别模型实验结果

Tab. 5　Experimental results of named entity recognition model　%

模型	精确率	召回率	F1 值
BiLSTM－CRF	80.08	81.14	80.56
BERT－BiLSTM－CRF	87.66	90.57	89.07

3.2　作物病害关联可视化结果

（1）作物病害实体抽取模型实验结果

实体抽取模型实验结果如表 5 所示。BERT－BiLSTM－CRF 模型精确率、召回率、F1 值分别达到了 87.66%、90.57%、89.07%，F1 值比 BiLSTM－CRF 模型高出 8.51 个百分点，能够实现有效的命名实体识别。

（2）基于病三角原理的病害影响因素关联可视化

作物种类、生长阶段和部位等实体体现了病三角原理中的宿主信息。不同作物在不同生长阶段和部位容易受到不同类型的病害侵袭。例如，叶类蔬菜在嫩叶/树枝生长期可能更容易受到叶斑病等病害的影响，而果实类作物可能在结果期遭受感染。这些因素有助于确定作物的易感性和防治时机。对知识图谱中侵染关系进行检索和可视化展示如图 7 所示，获取到 171 种病害与 129 种作物之间的 510 条关系。对于生长阶段和部位实体与病害间的关联关系，部分实例的可视化展示如图 8 所示。

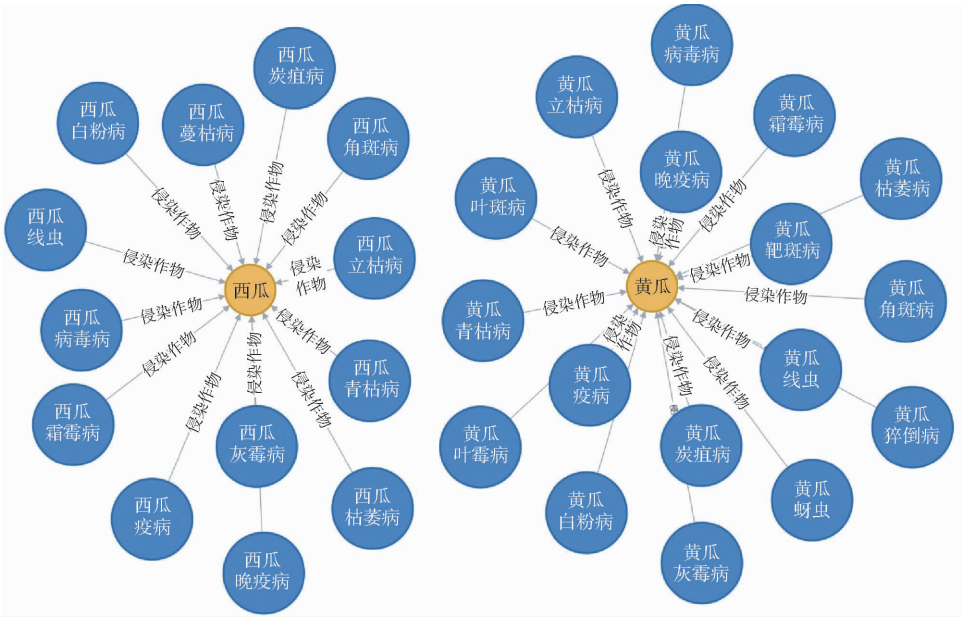


图 7　作物-病害侵染关系示例

Fig. 7　Example of infection relationship between crops and diseases

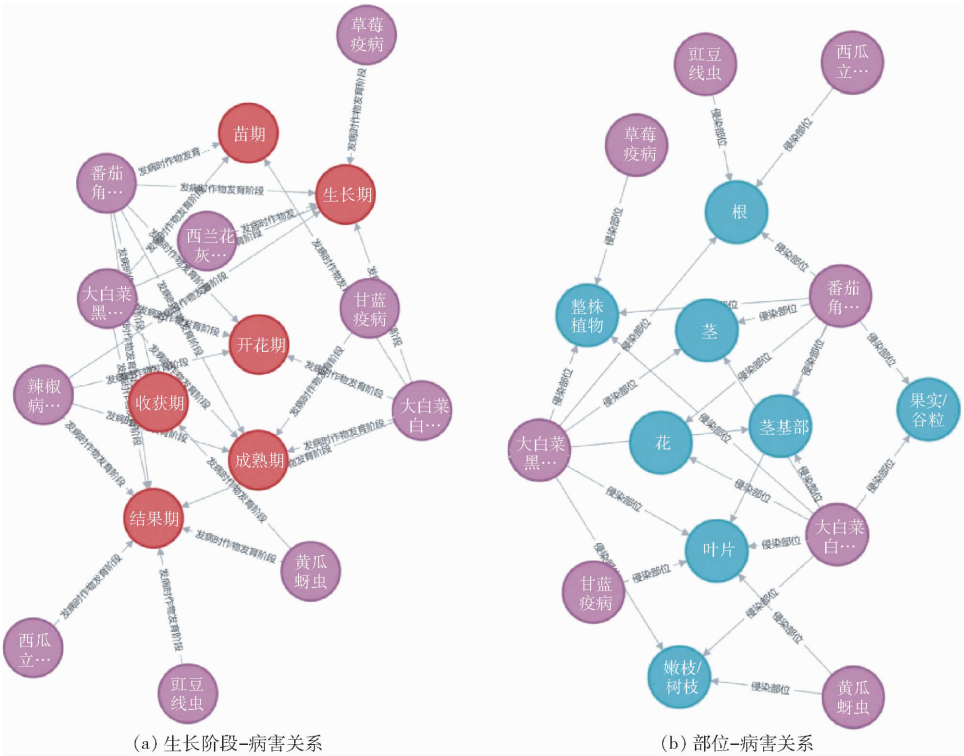
危害症状是病原体侵染的外在表征，也是区分病害的重要因素。例如对于番茄霜霉病和番茄白粉病这 2 种病害，二者都会表现出落叶、黄化、矮化等相同的 9 种症状，但只有番茄白粉病会表现出水疱状和梢枯的症状，只有番茄霜霉病会表现出落果、溃疡、叶烧和小叶 4 种症状（图 9a）。图 9b 共展现了 182 种病害和 29 种主要危害症状的关系，主要危害症状包括丛枝、发现昆虫/螨、叶斑、叶烧、咬痕、小叶、干枯、条斑、梢枯、水疱状等。

（3）病害与防控产品关联可视化

知识图谱中病害与防控产品之间的关系，涉及到 158 种病害与 140 种防控产品，揭示了不同作物病害与相应防控产品之间的关系网络。图 10a 展示了农药双炔酰菌胺可以治疗的不同种病害。图 10b 以西瓜炭疽病和西瓜疫病这 2 种常见病害为例，尽管在防控措施上有所不同，但也存在共用的防控产品。西瓜炭疽病通常需要特定的杀菌剂来预防菌丝体的生长和孢子的形成，而西瓜疫病可能更多地依赖于广谱性的生物农药来抑制病原体的扩散。共同的防控产品如甲基硫菌灵这种广谱生物农药。

（4）防控产品的生产、销售、供应关联可视化

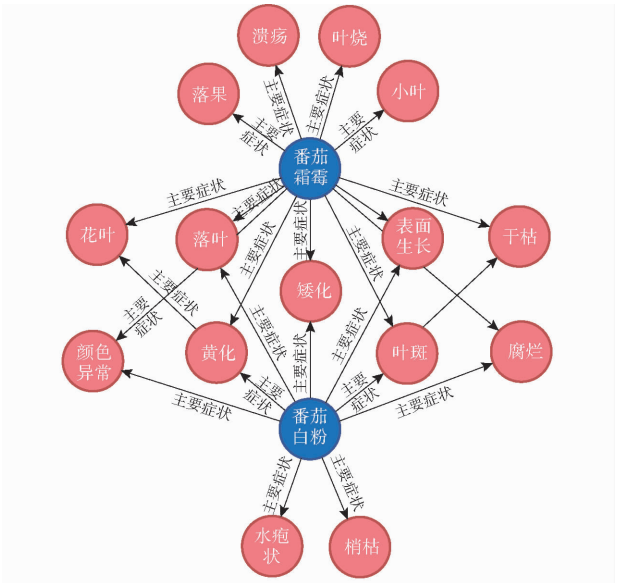
图 11 呈现了防控产品从生产到销售再到供应的整个链路的关联关系。图中不仅标识了关键生产



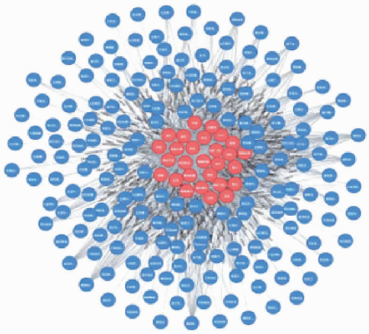
(a) 生长阶段-病害关系　　(b) 部位-病害关系

图 8　生长阶段-病害关系示例与部位-病害关系示例

Fig. 8　Example of growth stage-disease relationship and example of position-disease relationship



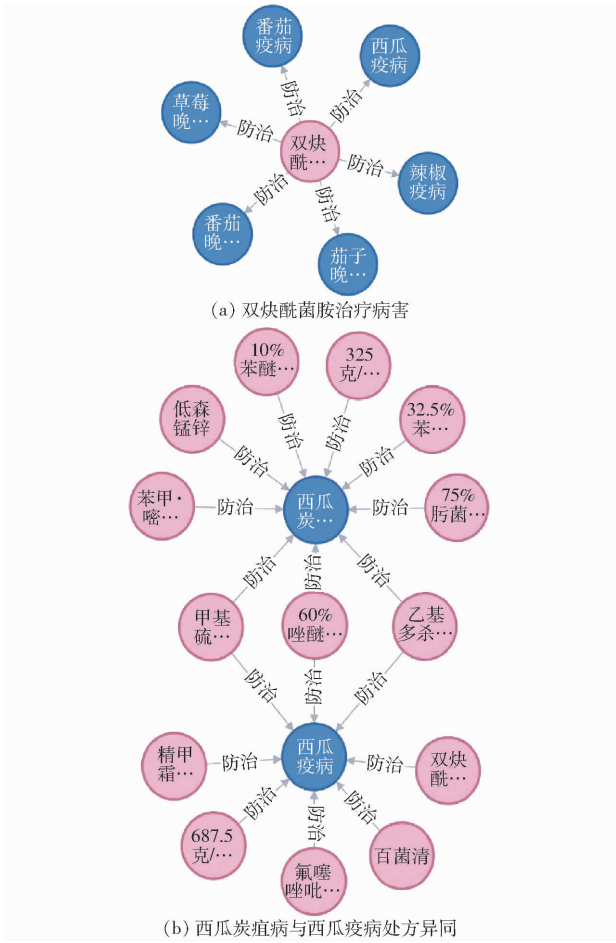
(a) 番茄白粉病与番茄霜霉病主要症状异同



(b) 182种病害和29种主要症状

图 9　病害-危害症状关联关系

Fig. 9　Association between disease and hazard symptoms



(a) 双炔酰菌胺治疗病害　　(b) 西瓜炭疽病与西瓜疫病处方异同

图 10　病害-防控产品关系示例

Fig. 10　Example of disease and prevention and control product query

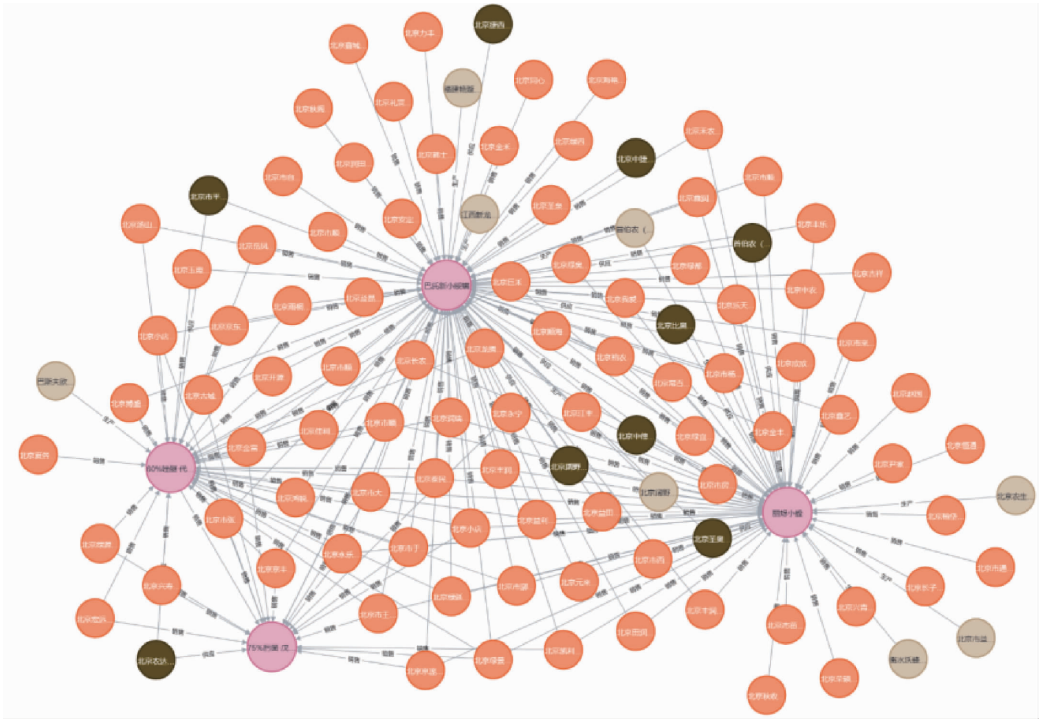


图 11　部分生产、销售和供应关系展示

Fig. 11　Partial production, sales and supply relationship display

厂家，也绘制产品流向市场的路径，以及这些产品如何被经销商分发到农业生产一线。例如生物农药 3 亿 CFU/克哈茨木霉菌可湿性粉剂由美国拜沃股份有限公司生产，北京中德恒星科技有限公司为供应商，并最终由北京海艳农乐商贸有限公司、北京我爱农农业生产资料门市部、北京裕农盛农资商店 3 家经销商供应给农户。

3.3　作物病害关联分析结果

（1）基于度中心性和介数中心性的关键节点识别

分别利用度中心性和介数中心性 2 种指标对作物病害处方知识图谱进行分析。度中心性衡量节点直接相连的边的数量，计算结果排序后的前 20 个节点（Top20）如图 12 所示。关键度中心性节点为病害严重程度、发病部位、作物生长阶段、病害名称和症状等相关实体，这些实体直接连接的节点数量较多，影响力较大。介数中心性衡量了节点作为连接不同节点之间路径桥梁程度，计算结果排序后的前 20 个节点（Top20）如图 13 所示。在介数中心性排名前列的节点全部是病害和防控产品，这些实体对于不同节点之间的路径传播起着关键作用。

（2）基于 Adamic－Adar 的病害-防控产品关联预测

以番茄霜霉病为例，计算其与其它病害节点之间的 Adamic－Adar 指数并排序。前 5 种病害结果如表 6 所示，Adamic－Adar 指数最高的病害为黄瓜霜霉病。这 2 种病害均是由霜霉菌引起的真

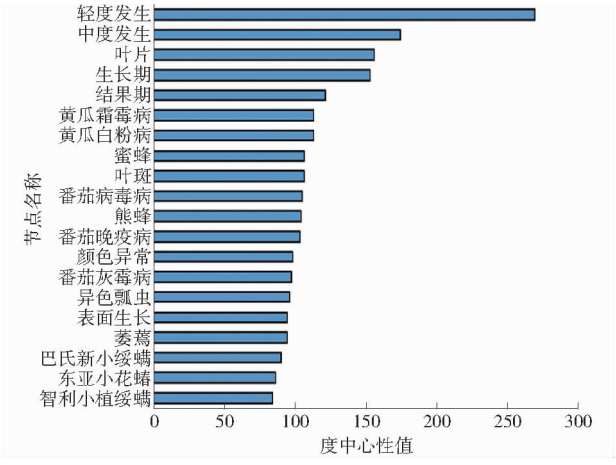


图 12　Top20 度中心性节点

Fig. 12　Top20 degree centrality nodes

菌性病害，常发病于潮湿、温暖环境，危害症状非常相似，包括叶片黄褐斑、霉斑等，防治方法也有相似之处。

在作物病害处方知识图谱查询黄瓜霜霉病的防控产品，排除其中与番茄霜霉病已有关联的产品，即作为番茄霜霉病的预测关联防控产品，如 3% 多抗霉素可湿性粉剂。对于 3% 多抗霉素可湿性粉剂，计算其与其它防控产品节点之间的 Adamic－Adar 指数并排序（表 6）。Adamic－Adar 指数最高的其它防控产品为 1 000 亿孢子/克枯草芽孢杆菌可湿性粉剂。因此将 1 000 亿孢子/克枯草芽孢杆菌可湿性粉剂作为黄瓜霜霉病和番茄霜霉病的预测关联防控产品。这些新的病害-防控产品关联预测，为研究

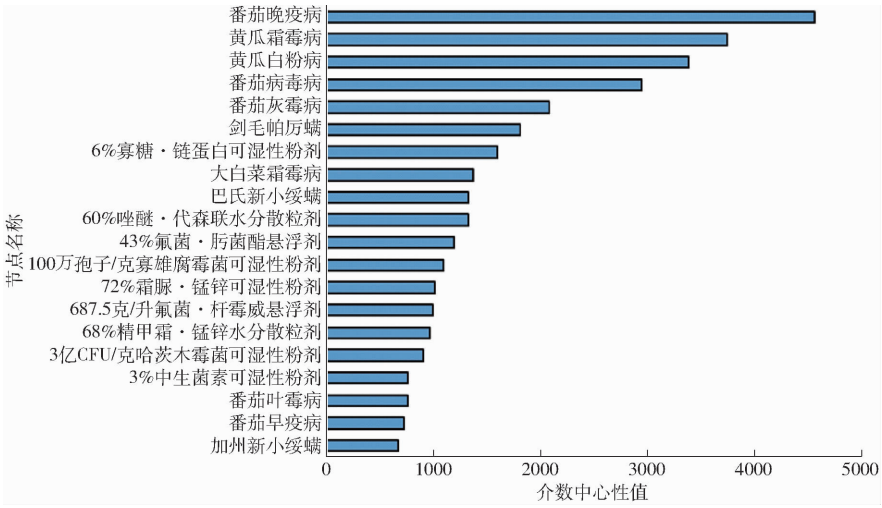


图 13 Top20 介数中心性节点

Fig. 13 Top20 intermediate centrality nodes

表 6 Adamic－Adar 指数 Top5 节点

Tab. 6 Adamic－Adar index Top5 nodes

节点	其它节点	Adamic－Adar 指数
番茄霜霉病	黄瓜霜霉病	15. 52
	番茄晚疫病	12. 88
	番茄灰霉病	11. 87
	番茄病毒病	11. 25
	番茄早疫病	11. 24
3% 多抗霉素可湿性粉剂	1 000 亿孢子/克枯草芽孢杆菌可湿性粉剂	3. 09
	68% 精甲霜·锰锌水分散粒剂	2. 55
	722 克/升霜霉威盐酸盐水剂	2. 31
	60% 唑醚·代森联水分散粒剂	2. 27
	64% 噁霜·锰锌可湿性粉剂	2. 27

和探索作物病害的新处方提供了可参考方向。

3.4 作物病害规则引擎应用结果

结合作物病害智能诊断和作物病害处方知识图谱,利用基于规则的匹配算法,可以实现防控类型选择、绿色补贴过滤、相关信息查询等多种处方筛选功能。根据诊断结果和作物病害处方知识图谱的规则引擎,在 Neo4j 图数据库可视化查询指定病害的防治产品,并选择防控类型等条件。例如实现草莓白粉病的化学农药防控处方推荐(图 14),或者草莓白粉病的生物农药防控处方推荐(图 15),同时获取防控产品相关的市场价、补贴后价、登记证号、防控类型等信息。

对于补贴敏感型生产者,可以通过绿色补贴过滤获取有绿色补贴的防控处方,如图 16 所示。

对于任意防控产品,也可以获取其生产、销售、供应关系,得到生产企业、供应商和经销商等信息(图 17)。

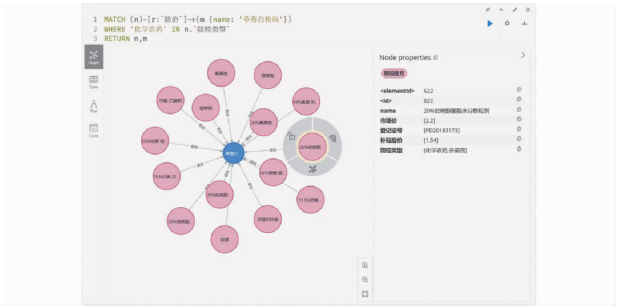


图 14 草莓白粉病的化学农药防控处方推荐

Fig. 14 Recommendation of chemical pesticide prevention and control prescription for strawberry powdery mildew

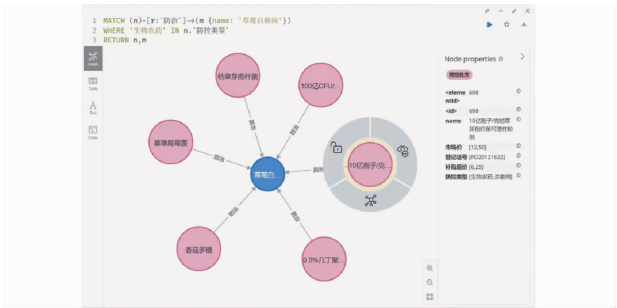


图 15 草莓白粉病的生物农药防控处方推荐

Fig. 15 Recommendation of biological pesticide prevention and control prescription for strawberry powdery mildew

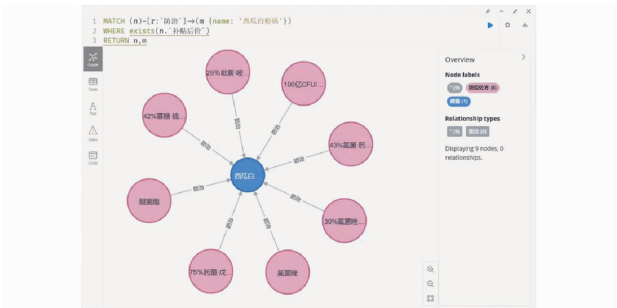


图 16 西瓜白粉病的绿色补贴过滤处方推荐

Fig. 16 Green subsidy filtering prescription recommendation for watermelon powdery mildew

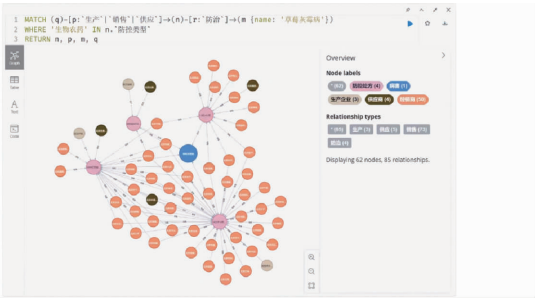


图 17　草莓灰霉病防控产品供应链信息查询

Fig. 17　Query on supply chain information of strawberry gray mold prevention and control products

4　结论

(1)针对电子病历等多源异构农业数据的知识抽取难题,从病三角原理出发,结合基于规则匹配与基于深度学习的知识抽取方法构建了作物病害处方知识图谱,可视化解析作物、病害、环境与处方的关联关系并制定了查询规则引擎。

(2)引入病三角原理,解析病原体、宿主、环境在病害侵染与防治过程中的关键作用;并以此为理论基础,结合数据特点构建作物病害处方知识图谱本体层,其中包含 18 类本体概念、17 条关系和 6 条属性边。

(3)根据不同数据源的固定格式和标签分别设计抽取规则,采用基于规则匹配的方法进行知识抽取;针对电子病历中的非结构化数据(问诊记录文本),采用基于深度学习的方法进行知识抽取。构建基于 BERT－BiLSTM－CRF 的命名实体识别模型,模型的精确率、召回率和 F1 值分别达到 87.66%、90.57% 和 89.07%。将其与基于规则匹配的方法结合,在知识融合后将所有实体、属性和关系实例储存进 Neo4j 图数据库。最终构建的作物病害处方知识图谱共包含1 121个实体实例和8 292条关系实例。

(4)基于作物病害处方知识图谱的关联分析。开展了基于度中心性和介数中心性的关键节点识别和基于 Adamic－Adar 的病害–防控产品关联预测;通过对作物病害处方知识图谱的查询和展示实现多种病害、处方及其特征属性关系的可视化分析,探究了各种实体(如病害、危害症状、防控产品、供应商等)和属性(如防控产品的市场价、登记证号等)之间的关系;制定规则引擎优化了病害防控产品推荐和信息检索功能。

参　考　文　献

[1] ERRAMI S A, HAJJI H, El KADI K A, et al. Spatial big data architecture: from data warehouses and data lakes to the lakehouse[J]. Journal of Parallel and Distributed Computing, 2023, 176: 70－79.

[2] NAMBIAR A, MUNDRA D. An overview of data warehouse and data lake in modern enterprise data management[J]. Big Data and Cognitive Computing, 2022, 6(4): 132.

[3] SAWADOGO P, DARMONT J. On data lake architectures and metadata management[J]. Journal of Intelligent Information Systems, 2021, 56(1): 97－120.

[4] FENSEL D, ŞIMŞEK U, ANGELE K, et al. Introduction: what is a knowledge graph[J]. Knowledge Graphs: Methodology, Tools and Selected Use Cases, 2020(1): 1－10.

[5] 石立莹,闫景瑞,李咏梅. 基于基础医学知识图谱的自主学习教学设计[J]. 基础医学教育,2023(10):862－865.
SHI Liying, YAN Jingrui, LI Yongmei. Teaching design of autonomous learning based on basic medical knowledge graphs[J]. Basic Medical Education, 2023(10): 862－865. (in Chinese)

[6] 王彩云,郑增亮,蔡晓琼,等. 知识图谱在医学领域的应用综述[J]. 生物医学工程学杂志,2023,40(5):1040－1044.
WANG Caiyun, ZHENG Zengliang, CAI Xiaoqiong, et al. A review of the application of knowledge graphs in medical field[J]. Journal of Biomedical Engineering, 2023, 40(5): 1040－1044. (in Chinese)

[7] 林睿凡,周洪伟,刘亮亮,等. 基于本体方法构建《伤寒论》版本知识图谱[J]. 中国中医基础医学杂志,2023,29(10):1630－1633.
LIN Ruifan, ZHOU Hongwei, LIU Liangliang, et al. Constructing the knowledge graphs of the version of treatise on febrile diseases based on ontology[J]. Chinese Journal of Basic Chinese Medicine, 2023, 29(10): 1630－1633. (in Chinese)

[8] 赵静,肖勇. 基于知识图谱的中医药信息化研究可视化分析[J]. 中国卫生信息管理杂志,2023,20(5):838－846.
ZHAO Jing, XIAO Yong. Visualization analysis of traditional Chinese medicine informatization research based on knowledge graph[J]. Chinese Journal of Health Information Management, 2023, 20(5): 838－846. (in Chinese)

[9] 郭旭超,唐詹,刁磊,等. 基于部首嵌入和注意力机制的病虫害命名实体识别[J]. 农业机械学报,2020,51(增刊 2):335－343.
GUO Xuchao, TANG Zhan, DIAO Lei, et al. Recognition of Chinese agricultural diseases and pests named entity with joint radical embedding and self-attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(Supp.2): 335－343. (in Chinese)

[10] 赵鹏飞,赵春江,吴华瑞,等. 基于注意力机制的农业文本命名实体识别[J]. 农业机械学报,2021,52(1):185－192.

ZHAO Pengfei, ZHAO Chunjiang, WU Huarui, et al. Named entity recognition of Chinese agricultural text based on attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(1): 185 - 192. (in Chinese)

[11] 韦紫君,宋玲,胡小春,等. 基于实体级遮蔽 BERT 与 BiLSTM - CRF 的农业命名实体识别[J]. 农业工程学报,2022,38(15):195 - 203.
WEI Zijun, SONG Ling, HU Xiaochun, et al. Named entity reeognition of agricultural based entity-level masking BERT and BiLSTM - CRF[J]. Transaetions of the CSAE, 2022, 38(15): 195 - 203. (in Chinese)

[12] LIN Y, LI D, PENG P, et al. A reasoning method for rice fertilization strategy based on spatiotemporal knowledge graph[J]. Transactions in GIS, 2024, 28(4): 902 - 924.

[13] 刘永波,黄强,高文波,等. 融合 BERT - WWM 和注意力机制的茶叶知识图谱构建[J]. 西南农业学报,2022,35(12):2912 - 2921.
LIU Yongbo, HUANG Qiang, GAO Wenbo, et al. Construction of tea knowledge graph by BERT - WWM and attention mechanism[J]. Journal of Southwest Agriculture, 2022, 35(12): 2912 - 2921. (in Chinese)

[14] GAO R, DONG Z, WANG Y, et al. Intelligent cotton pest and disease detection: edge computing solutions with transformer technology and knowledge graphs[J]. Agriculture, 2024, 14(2): 247.

[15] DEEPA R, VIGNESHWARI S. An effective automated ontology construction on the based agriculturen domain[J]. ETRI Journal, 2022, 44(4): 573 - 587.

[16] 李书琴,庞文婷. 嵌入词汇信息的 BERT - CRF 玉米育种实体关系联合抽取方法[J]. 农业机械学报,2023,54(11):286 - 294.
LI Shuqin, PANG Wenting. Joint extraction method of entity and relation in maize breeding based on BERT - CRF and word embedding[J]. Transactions of the Chinese Soeiety for Agricultural Machinery, 2023, 54(11): 286 - 294. (in Chinese)

[17] LI D, SHEN C, DAl X, et al. A kind of agricultural content networking information fusion method based on ontology[C]// Cloud Computing and Security: 4th International Conference (ICCCS), 2018.

[18] YAN W, ZHANG Z, ZHANG Q, et al. Deep data analysis-based agricultural products management for smart public healthcare [J]. Frontiers in Public Health, 2022, 10: 1 - 9.

[19] 张海瑜,陈庆龙,张斯静,等. 基于语义知识图谱的农业知识智能检索方法[J]. 农业机械学报,2021,52(增刊):156 - 163.
ZHANG Haiyu, CHEN Qinglong, ZHANG Sijing, et al. Intelligent retrieval method of agricultural knowledge based onsemantic knowledge graph[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(Supp.): 156 - 163. (in Chinese)

[20] 刘明鹏,王忠明,马文君. 基于本体的造林树种知识图谱构建与应用[J]. 北京林业大学学报,2023,45(8):109 - 122.
LIU Mingpeng, WANG Zhongming, MA Wenjun. Construction and application of ontology based knowledge graph for afforestation tree species[J]. Journal of Beijing Forestry University, 2023, 45(8): 109 - 122. (in Chinese)

[21] 戈为溪,周俊,袁立存,等. 基于知识图谱与案例推理的水稻精准施肥推荐模型[J]. 农业工程学报,2023,39(2):126 - 133.
GE Weixi, ZHOU Jun, YUAN Licun, et al. Recommendation model for rice precision fertilization using knowledge graph and case-based reasoning[J]. Transactions of the CSAE, 2023, 39(2): 126 - 133. (in Chinese)

[22] 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. (in Chinese)

[23] SHI Y, ZHOU K, ZHOU M, et al. Temporal knowledge graph for food risk prediction[J]. lEEE Transactions on Artificia Intelligence, 2023, 5(5): 2217 - 2226.

[24] FLEMING-DUTRA K E, HERSH A L, SHAPIRO D J, et al. Prevalence of inappropriate antibiotic prescriptions among US ambulatory care visits, 2010—2011[J]. JAMA, 2016, 315(17): 1864 - 1873.

[25] HOANG K H, HO T B. Learning and recommending treatments using electronic medical records[J]. Knowledge-based Systems, 2019, 181: 104788.

[26] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2020, 36(4): 1234 - 1240.

[27] LIU Y, YU Z, YANG Y. Diabetes risk data mining method based on electronic medical record analysis[J]. Journal of Healthcare Engineering, 2021, 2021: 1 - 11.

[28] LIANG D, KRISHNAN R, HOFFMAN M, et al. Variational autoencoders for collaborative filtering[C]// Proceedings of the 2018 World Wide Web Conference, 2018.

[29] GOH W, TAO X, ZHANG J, et al. Feature-based learning in drug prescription system for medical clinics[J]. Neural Processing Letters, 2020, 52(3): 1703 - 1721.

[30] WANG X, ZHANG Y, WANG X, et al. A knowledge graph enhanced topic modeling approach for herb recommendation[C]// International Conference on Database Systems for Advanced Applications, 2019.

[31] 郭文龙,罗熊,姜惠娟,等. 基于 TF－IDF 算法的方剂构成相似度可视化研究[J]. 中国中医药信息杂志,2019,26(7):104－108.
GUO Wenlong, LUO Xiong, JIANG Huijuan, et al. Research on visualization of formula composition similarity based on TF－IDF algorithm[J]. Chinese Journal of Traditional Chinese Medicine Information, 2019, 26(7): 104－108. (in Chinese)

[32] 赵磊,张涛,郑书恒,等. 运行植物诊所破解绿色防控技术推广难题[J]. 中国植保导刊,2018,38(2):84－85.
ZHAO Lei, ZHANG Tao, ZHENG Shuheng, et al. Running a plant clinic to solve the problem of promoting green prevention and control technology[J]. China Plant Protection Guide, 2018, 38(2): 84－85. (in Chinese)

[33] CHAPPELKA A H, GRULKE N E. Disruption of the 'disease triangle' by chemical and physical environmental change[J]. Plant Biology, 2016, 18: 5－12.

[34] MUNIR M. Plant disease epidemiology: disease triangle and forecasting mechanisms in highlights[J]. Hosts Virus, 2018, 5(1): 7－11.

[35] 潘红丽,李迈和,蔡小虎,等. 海拔梯度上的植物生长与生理生态特性[J]. 生态环境学报,2009,18(2):722.
PAN Hongli, LI Maihe, CAI Xiaohu, et al. Plant growth and physiological ecological characteristics on altitude gradient[J]. Journal of Ecological Environment, 2009, 18(2): 722. (in Chinese)

[36] 刘秉儒. 生物多样性的海拔分布格局研究及进展[J]. 生态环境学报,2021,30(2): 438.
LIU Bingru. Research and progress on the altitudinal distribution pattern of biodiversity [J]. Journal of Ecological Environment, 2021, 30(2): 438. (in Chinese)

[37] ALEXANDER H M. Disease in natural plant populations, communities, and ecosystems: insights into ecological and evolutionary processes[J]. Plant Disease, 2010, 94(5): 492－503.

(上接第 498 页)

[44] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv: 2010.11929, 2021.

[45] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: hierarchical vision Transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 10012－10022.

[46] LIU Z, HU H, LIN Y, et al. Swin Transformer V2: scaling up capacity and resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 12009－12019.

[47] WANG W, TRAN D, FEISZLI M. What makes training multi-modal classification networks hard? [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[48] GOYAL L, SHARMA C M, SINGH A, et al. Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture[J]. Informatics in Medicine Unlocked, 2021, 25: 100642.

[49] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5): 602－610.

[50] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding [J]. arXiv: 1810.04805, 2019.

[51] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 38th International Conference on Machine Learning, 2021: 8748－8763.