

郑颖,金松林,张自阳,等. 基于领域本体的农作物病虫害问题分类研究[J]. 江苏农业科学,2016,44(9):145-148.
doi:10.15889/j.issn.1002-1302.2016.09.041

基于领域本体的农作物病虫害问题分类研究

郑颖¹,金松林¹,张自阳²,霍云凤²,王斌²

(1. 河南科技学院信息工程学院,河南新乡 453003; 2. 河南科技学院生命科技学院,河南新乡 453003)

摘要:问题分类是问答系统的重要组成部分,其作用是将问题划分到对应的类别里以提高问答系统的准确率。本研究提出了一种基于领域本体的农作物病虫害问题分类方法,该方法首先构建农作物病虫害领域本体,将领域本体中的领域词添加到分词系统中以提高分词的准确率。然后提取特征词,并利用同义词词林和领域本体对特征词进行扩展。最后,针对农作物病虫害领域的特殊性将问题分为 4 类,利用语义和规则相结合的问题分类方法对问题分类。试验结果表明,该方法有助于提高问题分类的准确率。

关键词:农作物病虫害;领域本体;特征词扩展;问题分类

中图分类号: TP391;S126 **文献标志码:** A **文章编号:** 1002-1302(2016)09-0145-03

长期以来,病虫害一直是影响农作物产量的主要问题,每年因病虫害损失的粮食约有 250 亿 kg,有效预防和控制病虫害的发展对于提高农作物产量有着重要的意义。问答系统是一种能够对用户输入的问题进行快速分析并准确地返回答案的智能系统。为农民提供一个病虫害领域的问答系统,可以为农民在农作物种植过程中出现的疑难问题提供实时指导,进而减少粮食的损失。问答系统的工作流程一般分为 3 个阶段:问题分析、答案检索和返回答案。问题分类是问句分析阶段需要解决的关键问题,它对答案的抽取有着指导意义^[1],例如问句“玉米螟虫最佳防治时机是什么时候?”如果能够分析出该问句为询问时间类,答案的抽取则具有一定针对性,答案抽取的准确率也会提高。

传统问答系统对问句的分析只利用问句的表层特征信息,并没有考虑问句的语义特征,导致问答系统抽取到的答案准确率较低^[2]。本体是一种语义层次的领域知识建模工具,对概念及概念之间的关系进行明确定义^[3]。本体用形式化定义领域内的各种资源及资源之间的联系,不仅使知识的语义信息更加丰富^[4],而且还具有重用性和知识推理的特点。本研究针对农作物病虫害领域的特殊性,基于本体理论,搜集农作物种植过程中的病虫害知识,构建农作物病虫害知识本体,将本体运用在病虫害问答系统中问题分类的整个过程中,提高问题分类的准确率。本课题研究的主要问题有农作物病虫害本体构建、特征词扩展及问题分类。

1 基于领域本体问题分类总体框架

对问题进行分类首先需要将问题变为计算机能够理解的

收稿日期:2016-03-18

基金项目:国家科技支撑计划(编号:2015BAD26B01);科技部创新方法专项(编号:2015IM010400);河南省教育厅科学技术研究重点项目(编号:14A520080)。

作者简介:郑颖(1987—),女,河南新野人,硕士,助教,主要从事自然语言处理、农业信息化研究。E-mail:zhengying198766@126.com。

通信作者:王斌,硕士,副教授,主要从事小麦育种及农业信息化研究。Tel:(0373)3040337;E-mail:wangbin6339@163.com。

形式化语言,常用的处理方法是将句子变为由特征词组成的向量空间模型,处理过程包括:预处理和特征词抽取及扩展,领域本体贯穿于整个阶段。问题分类工作流程如图 1 所示。

(1)预处理:语义分析的基础,包括分词、词性标注和去停用词。领域特征词普通分词系统还无法准确划分,因此需要将领域本体中的概念添加在分词系统中以提高分词的准确率。

(2)特征词抽取及扩展:特征词对句子理解起着关键作用。在候选答案句中可能包含特征词的同义词或者近义词,如果不进行特征词扩展有可能遗漏问题的答案,因此有必要对其适当扩展。

(3)问题分类:制定问题分类,并对每个类别制定特征词表和规则,采用基于语义和规则相结合的方法对问题分类。

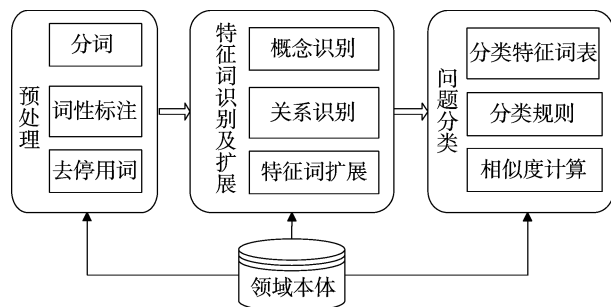


图1 基于领域本体的农作物病虫害问题分类流程

2 基于领域本体问题分类研究

2.1 农作物病虫害本体构建

领域本体是面向某一个特定领域的概念及概念之间关系的规范化描述。对于特定领域,其专业知识强,本体构建必须收集相关的领域知识,并且需要该领域的专家进行指导,这样才能使构建的本体更加合理^[5]。本研究将农作物在种植过程中病虫害问题的相关概念及其之间的关系组织起来,形成可重用的农作物病虫害领域本体。

2.1.1 本体构建思路 本体的构建工作主要包括:领域相关知识的获取、领域概念的获取和领域概念的关系^[6]。只有充分了解领域内的相关知识才能构建出高质量的本体。《农业

科学叙词表》^[7] 提供了丰富的农业领域知识,利用《农业科学叙词表》中的领域概念及概念之间的关系可以减少构建本体的工作量。另外,由于《农业科学叙词表》形成已久,其中的概念和知识没有及时更新,因此,利用网络爬行工具对中国农业信息网、农林网等专业领域网站进行知识抽取以便及时补充和更新知识概念。最后,由领域专家对知识进行检查整理,去除抽取错误的知识,合并重复知识,确保构建本体的准确性。构建方法如图 2 所示。

2.1.2 农作物病虫害本体构建 农作物病虫害本体包括类、属性和实例 3 个组成部分。类即概念,是本体的重要组成部分,属性和实例都是对类的说明。对本体的构建首先需要将类按照合理的层次组织起来,类的获取参考《农业科学叙词表》中的分类,并根据本研究构建的农作物病虫害本体的实际需求,将顶层类分为农作物、病害、虫害、防治方法 4 类,农作物类又划分为禾谷类、豆类、经济类、薯类、蔬菜类和果树类。本体概念层次如图 3 所示。属性是描述类之间的关系,

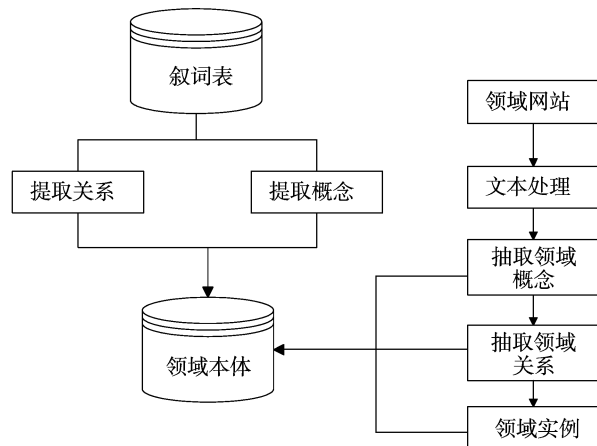


图2 基于领域本体的构建方法

例如特征和颜色表示值—属性关系。创建的本体中也应该包含实例,例如“小麦”“高粱”都是禾谷类作物的实例。

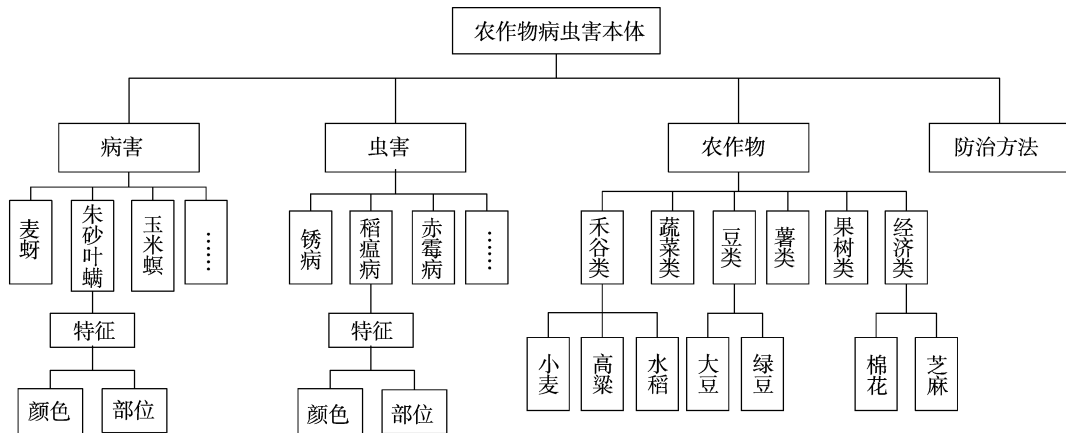


图3 基于领域本体概念层次

2.2 预处理

分词是文本处理的关键问题,其效果好坏直接影响语义分析的结果。汉语中词语之间并无分割,因此需要借助工具将相连的词语分隔开。本研究使用的分词工具是张华平博士开发的分词工具 NLPPIR (ICTCLAS2015),该工具可以自动完成分词及词性标注。另外,NLPPIR 还具有添加用户词典的功能,本研究对农作物病虫害领域的问题分类,涉及到很多专业领域词汇,为了确保分词的准确性,将构建的领域本体中的领域概念添加到用户词典中,并标注其词性。在汉语中,有许多类似“的”“了”“啊”等停用词,这些停用词对文本理解没有实际意义,但其出现频率却非常高,去除停用词可以大大缩小特征词的处理空间。例如,对于句子“麦蚜,一直是小麦灌浆期集中在穗部危害的主要害虫,如果控制不当对小麦的粒重影响很大。”进行预处理后表示成空间向量为{麦蚜/n,一直/d,小麦/n,灌浆期/t,集中/v,穗部/n,危害/n,主要/b,害虫/n,控制/v,不/d,当/v,小麦/n,粒重/n,影响/vn,很/d,大/a}。

2.3 特征词抽取及扩展

2.3.1 特征词抽取 经过预处理后的句子虽然降低了词语的维数,但是词语数量仍然较多,处理过多的词语会对结果造成一定误差,因此,需要对预处理后的句子进一步提取特征词。特征词抽取主要考虑词语在文本中出现的频率,如 TF-IDF

特征词抽取方法。在特定领域内,一些词语出现的频率虽然不高,但是它们对句子的理解有重要作用,如果忽略了这些领域词语将会直接影响到系统的准确性。因此,本研究利用构建的本体来识别领域内特征词,提高抽取特征词的精确度。首先利用 TF-IDF 的方法抽取特征词,然后利用小麦病虫害本体进行领域特征词抽取,将 2 次抽取的结果进行合并即为特征词。例如,句子“麦蚜,一直是小麦灌浆期集中在穗部危害的主要害虫,如果控制不当对小麦的粒重影响很大。”的特征词向量为{麦蚜/n,小麦/n,灌浆期/t,穗部/n,粒重/n}。

2.3.2 特征词扩展 汉语中语言表达丰富,很多农作物在不同地区的习惯名称也不相同,如果问句中的特征词与答案中的特征词在语义上一致,但是表达名称不一样时就会降低检索答案的准确率。例如“大豆”和“黄豆”表示的是一种植物,“番茄”和“西红柿”也表示的是同一种植物,在这些情况下就需要对词语进行同义词扩展,本研究选用哈工大信息检索实验室《同义词词林扩展版》对同义词进行扩展。该词典收录了近 7 万条词语,采用树形结构组织词语,处于树形结构同一行的词语意思相近,因此与特征词处于同一行的词语都可以作为该特征词的扩展词。

另一方面,有些词语不是同义词,但是所要表达的意思很接近,例如问句“百农 AK58 纹枯病主要发病期在什么时候?”

特征词“百农 AK58”在答案库中没有出现,可能就搜索不到正确答案,但是如果通过构建的农作物病虫害本体可知“百农 AK58”是“小麦”的下位词,由此可知“百农 AK58”是小麦的一个品种,该问句扩展为“小麦纹枯病主要发病期在什么时候?”就很容易搜到答案。因此,本研究对特征词扩展的另一种方法是利用病虫害本体中的上下位关系对特征词扩展。

对于特征词的扩展也应当谨慎,如果扩展范围太广泛会检索到很多无关信息,则会影响答案的准确率。本研究只对特征词中的名词和动词进行扩展,其他词语暂不进行扩展。

2.4 问题分类

目前对于问题还没有统一的分类方法,具有代表性的是哈工大的基于答案类型的开放领域问题分类方法,该方法将问题分为人物、时间、地点、数量、实体、描述和未知 7 个大类^[8]。本课题研究的是农作物病虫害预防领域问题,具有领域特殊性,因此不能将哈工大的问题分类直接应用在本研究中。本研究按照领域专家的建议将该领域问题分为 4 类:病虫害种类、发病时期、病症预防、病症描述。目前,问题类型识别常用基于语义和基于规则的分类方法。基于规则的方法需要针对每类问题制定一定的规则,如果制定的规则过多则需要大量人力,如果规则太少则无法准确判断问题的类别,并且还会出现一个问题可以匹配到多个分类规则里面。基于语义的分类方法在研究中也取得不错的分类效果,但是如果问题过于简单,或者虽然较长但是所包含的特征词较少时分类效果也不理想。基于以上原因,本研究对基于语义和基于规则的分类方法相结合以提高分类的准确性。基于语义和规则相结合的分类方法思路如下:

- (1) 根据领域专家的建议为每类问题建立对应的特征词汇总表和规则库,部分特征词汇总表和规则库如表 1 和表 2 所示;
- (2) 抽取问句特征词,并对特征词中名词和动词进行扩展词,形成问句特征向量为 $T_w = \{W_1, W_2, W_3, \dots\}$;
- (3) 将 T_w 分别与问题类型 C_1, C_2, C_3, C_4 中的特征词汇进行相似度计算(特征词间的相似度计算按照刘群等基于知网提出的计算方法^[9]),计算结果分别为 $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, 其中 $\alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \alpha_4$, 如果 $\alpha_1 - \alpha_2 \geq \beta$, 则类别 C_1 即为问题所属类别,分类结束,否则继续;
- (4) 将问句分别与类别 C_1 和 C_2 进行规则匹配,选择最匹配类别作为问句所属类别。

表 1 农作物病虫害问题的领域分类

问题类型	特征词汇
病虫害种类	病害、虫害……
发病时期	病虫害名称、发病、时间(时期)……
病症预防	病虫害名称、预防、防治、治理、措施……
病症描述	发病部位(根、茎、叶、花、果实)、病症(霉状物、粉状物、粒状物、脓状物)、病状(变色、腐烂、萎蔫)……

3 试验分析

问题分类中参数 β 的取值直接影响分类的准确率,因此通过试验确定参数 β 的取值使分类结果最优。

3.1 试验数据

本研究所用试验语料来自于农林网、农业信息网等农业

表 2 基于领域本体农作物病虫害的问题规则示例

问题类型	规则
病虫害种类	*(指代农作物)有哪些病虫害?
发病时期	*(指代病虫害)什么时候/时期出现?
病症预防	怎么/如何预防/治理*(指代病虫害)?
病症描述	……特征是什么病(虫)害?

类网站,对语料进行清洗、分类,所用各类问题语料数量如表 3 所示。

表 3 基于领域本体农作物病虫害的试验语料分类情况

问题类型	数量
病虫害种类	156
发病时期	197
病症预防	273
病症描述	105

3.2 试验结果及分析

分别将参数 β 设定不同数值,采用基于语义和规则相结合的分类方法判断试验问题语料所属分类的准确率如图 4 所示。由图 4 试验结果表明,参数 β 的最佳取值为 0.2,通过试验结果还可以看出,随着参数 β 取值增大分类准确率降低。结果表明:当待判定问题与领域分类特征库中两类分类计算结果较为接近时,才有必要根据规则判断其真正所属类别;当待判定问题与分类特征库中两类分类计算结果相差较大时,因为构建的规则并不全面而导致问题所属类别判断存在误差较大。因此下一步的工作中需要将问题分类的规则库进行扩充,使其规则更加丰富,提高分类结果的准确性。

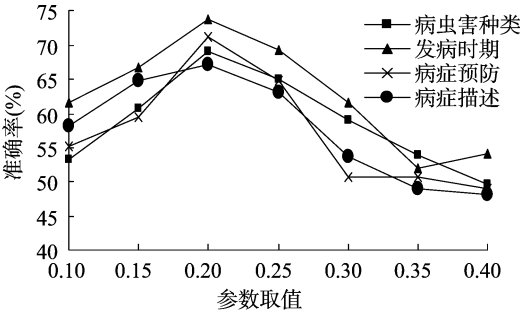


图 4 参数不同取值时试验结果

将参数 β 的取值设置为 0.2 时,分别采用基于语义的分类方法、基于规则的分类方法和本研究所用的分类方法进行比较分类的准确率,结果如表 4 所示。从表 4 可以看出,本研究的分类方法比基于规则和基于语义的方法分类准确率都有提高,特别是相对于基于规则的分类方法准确率有较大的提高,而相对于基于语义的方法准确率提高较小,分析是因为采用规则的分类方法对每类问题制定的规则有限,直接影响了分类的准确率。另外,目前对于问题的分类仅分为 4 类,问题类别划分不够细致,这也会影响到问题分类的准确率。

表 4 农作物病虫害问题分类方法的准确率比较结果

问题类型	准确率(%)		
	本研究方法	基于规则的方法	基于语义的方法
病虫害种类	69.1	53.8	63.7
发病时期	73.9	62.5	69.4
病症预防	71.2	49.7	65.8
病症描述	67.2	51.8	60.9

吉沐祥,姚克兵,王建华,等. 设施草莓病虫害全程绿色防控技术模式[J]. 江苏农业科学,2016,44(9):148-151.
doi:10.15889/j.issn.1002-1302.2016.09.042

设施草莓病虫害全程绿色防控技术模式

吉沐祥¹,姚克兵¹,王建华¹,杨勇¹,彭燕琼²,李国平¹

(1. 江苏丘陵地区镇江农业科学研究所,江苏句容 212400; 2. 江苏农林职业技术学院,江苏句容 212400)

摘要:为了提高草莓安全品质,避免农药残留污染,保证食用安全,针对设施草莓生产现状与实际问题,结合国内外的新技术研究成果,总结设施草莓病虫害全程绿色防控技术模式。分别从技术模式、防控目标、关键技术、效益分析、适用规模与主要投入品等几个方面进行阐述,其关键技术包括农业生态防治技术、物理防控技术、生物防治技术、化学农药防治技术。

关键词:设施草莓;病虫害;绿色防控;技术模式

中图分类号:S436.68⁺4 **文献标志码:**A **文章编号:**1002-1302(2016)09-0148-04

设施草莓发生的病虫害较多,主要病害有炭疽病、灰霉病、白粉病、枯萎病、黄萎病、根腐病等,主要虫害生物有蚜虫、蓟马、红蜘蛛、斜纹夜蛾、蛱蝶、地老虎等^[1]。随着消费者对鲜食果品质量安全要求普遍提高,绿色和有机果品备受推崇,避免农药残留污染,提高果实安全品质已成为优质草莓生产的重要内容。“舌尖上的安全”是事关人们健康的大事,已引起政府的高度关注,随着人们生活水平的迅速提高,以鲜食为主的大棚草莓消费量不断增大,其食用安全格外重要。因而,保证草莓鲜果品质,尽量减少草莓鲜果农药残留,是我们科研

工作者和生产者的首要任务^[2]。设施草莓病虫害绿色防控必须以农业防治为主导,重视物理防治、生态调控和生物方法等综合措施,实施健身栽培,提高植株抗病虫害能力,辅以化学防治,做到经济、安全、有效、简便地控制病虫害,提高草莓安全品质,避免农药残留污染,保证食用安全,真正使消费者吃上放心草莓^[3]。为此,笔者根据目前设施草莓生产实际,结合国内外的新技术研究成果,总结设施草莓病虫害全程绿色防控技术模式。

1 技术模式

农业防治(轮作换茬、摘除病叶老叶、平衡施肥等)、连作田休闲期太阳能夏季高温消毒、性诱剂(诱杀斜纹夜蛾等)、黄/蓝板诱杀或网室阻隔(蓟马、蚜虫等)、生物防治、低毒低残留化学药剂防治病虫害。

2 防控目标

2.1 防控效果

设施草莓病虫害绿色防控总体防效达到90%以上,减少

收稿日期:2016-04-18

基金项目:国家科技富民强县专项(编号:BN20156222);上海市农委示范推广计划[编号:沪农科推字(2015)第2-7号];江苏省农业科技自主创新资金[编号:CX(15)1029]。

作者简介:吉沐祥(1963—),男,江苏宝应人,研究员,主要从事果树植保与农药研究开发工作。Tel:(0511)80978086;E-mail:jilvdun2800@163.com。

通信作者:李国平,硕士,研究员,主要从事果树与丘陵农业资源开发利用研究。E-mail:jrlgp@126.com。

4 结论

本研究首先构建农作物病虫害领域本体,将领域本体应用在预处理、特征词抽取及扩展中,根据领域的特殊性将问题分为4类,利用基于语义和规则相结合的分类方法对问题进行分类。试验结果表明,本研究方法对农作物病虫害领域问题分类时具有一定的有效性。但是,本研究仍存在问题,例如领域本体如何实现自动更新、问题类型规则不完善等,这些都将是下一步工作的重点。

参考文献:

- [1]郑实福,刘挺,秦兵,等. 自动问答综述[J]. 中文信息学报,2002,16(6):46-52.
- [2]廖梦. 面向问答系统的金融本体构建技术研究[D]. 哈尔滨:哈尔滨工业大学,2014.
- [3]邓志鸿,唐世渭,张铭,等. Ontology研究综述[J]. 北京大学学

- 报:自然科学版,2002,38(5):730-738.
- [4]Li S P, Yin Q W, Hu Y J, et al. Overview of researches on ontology[J]. Journal of Computer Research and Development, 2004, 41(7): 1041-1052.
- [5]潘彩霞,薛佳妮,于辉辉,等. 基于本体的鱼病诊断专家系统的构建[J]. 广东农业科学,2015,42(1):157-160.
- [6]王超,李书琴,肖红. 基于文献的农业领域本体自动构建方法研究[J]. 计算机应用与软件,2014,31(8):71-74.
- [7]农业部情报研究所. 农业科学叙词表[M]. 北京:中国农业出版社,1994.
- [8]Zhang W, Chen J J, Niu Y Q. Research on Chinese question classification based on Hownet and dependency parsing[C]. The 3rd International Workshop on Intelligent System and Application. Wuhan, China, 2011:483-486.
- [9]刘群,李素建. 基于《知网》的词汇语义相似度计算[C]. 台北第三届汉语词汇语义学研讨会论文集,2002:59-76.