# 一个基于枸杞病虫害领域本体的语义检索模型

李贯峰[1],李卫军[2]
(1. 宁夏大学 信息工程学院,宁夏 银川 750021;
2. 北方民族大学 网络信息技术中心,宁夏 银川 750021)

**摘 要**:由于缺少信息在语义上的处理和表示,传统的以关键字和主题词为检索途径的信息检索方法会导致检索结果不全面和不准确,无法完全满足用户的检索要求。为了提升检索系统的检索质量,将本体引入至语义检索过程中,提出了一种基于枸杞病虫害领域本体的语义检索模型,并对模型涉及的一些关键技术进行了研究。该模型构建了枸杞病虫害领域本体,并修复了本体不一致问题,确保领域知识能准确的组织和表示,利用本体固有的树形结构,结合语义距离、上下位概念重合度及概念节点层次深度等影响语义相似度计算的因素,提出了一个概念相似度算法,结合所建立的语义推理规则,构建了基于枸杞病虫害领域本体的查询与检索模型。实验结果表明,该语义检索模型能较好地弥补传统检索方式的不足,提高信息检索的查全率和查准率。

**关键词**:本体;枸杞病虫害;不一致性检测;语义相似度;语义检索
**中图分类号**:TP391 **文献标识码**:A **文章编号**:1673-629X(2017)09-0048-05
doi:10.3969/j.issn.1673-629X.2017.09.011

# A Semantic Retrieval Model with Domain Ontology Based on Wolfberry Disease and Pests

LI Guan-feng[1], LI Wei-jun[2]
(1. School of Information Engineering, Ningxia University, Yinchuan 750021, China;
2. Network Information & Technology Center, Beifang University of Nationalities, Yinchuan 750021, China)

**Abstract**: The traditional information retrieval methods based on keywords and subject are lack of processing and presentation on the semantic level and thus lead to incomplete and inaccurate retrieval results, which cannot meet the user's retrieval needs totally. In order to improve the quality of retrieval system, a model of semantic retrieval based on ontology for wolfberry disease and pests domain has been presented and its key technologies have been investigated in the processing of the introduction of ontology into semantic retrieving. It constructs the domain ontology of wolfberry diseases and pests, and modifies its inconsistent problem to ensure consistency and accuracy of wolfberry diseases and pests knowledge. Under the guidance of hierarchical tree structure of the domain ontology, a concept similarity method considering semantic distance, superior concepts coincidence degree and depth of concept nodes is proposed. Combined with the semantic inference rules a semantic retrieval model based on domain ontology of wolfberry disease and pests is realized. The experimental results demonstrate that the semantic retrieval model has well overcome the deficiency of the traditional retrieval method and effectively improved the recall and precision of information retrieval.

**Key words**: ontology; wolfberry disease and pests; inconsistency detection; semantic similarity; semantic retrieval

## 0 引 言

宁夏枸杞自古享誉中外,是宁夏最具潜力的优势特色产业之一。目前,宁夏枸杞种植面积 85 万亩,枸杞干果总量达到 13 万吨,约占全国总产量的 55%,年综合产值超过 80 亿元,是宁夏第一大出口农产品。在枸杞栽植和生产过程中,枸杞病虫害问题一直是宁夏枸杞产业发展的主要问题。随着信息技术与互联网技术的迅速发展,如何准确全面地获取枸杞病虫害信息资源,是目前枸杞产业信息服务中一个亟待解决的问题。传统以关键字、主题词等字符串匹配原理为核心的信息资源检索方法由于缺少在语义层面上的处理和表示,用户输入的检索内容与信息资源库中的目标内

**收稿日期**:2016-10-13 **修回日期**:2017-01-18 **网络出版时间**:2017-07-11
**基金项目**:宁夏自治区高校科研基金资助项目(NGY2014009)
**作者简介**:李贯峰(1979-),男,硕士,副教授,研究方向为知识工程。
**网络出版地址**:http://kns.cnki.net/kcms/detail/61.1450.TP.20170711.1455.056.html

容不相匹配,不能从根本上解决用户查询意图与检索资源之间的语义匹配问题,造成了检索结果的不全面、不准确,甚至系统无法返回符合用户需要的检索结果,从而影响检索结果的客观性。

本体(Ontology)是一种描述概念以及概念间关系的知识建模工具。本体具有良好的概念层次结构和对逻辑推理的支持,因而将本体引入信息检索系统中,能够为改进信息检索性能提供组织形式和语义上的保证[1]。一方面本体提供了对概念的语义支持,保留了概念之间的语义关系,从而实现基于语义理解的智能检索;另一方面引入了推理机制,本体通过属性和公理描述概念之间的逻辑关系和设计的推理规则实现推理,从而实现隐含知识的发现。

近年来,以本体为知识模型的语义检索技术已成为一个研究热点,国内外学者开展了大量的研究工作[2]。文献[3]为提高多式信息检索系统的性能,利用医学本体扩展了用户检索关键词;文献[4]提出了一种基于领域本体的混合查询方法,利用查询重写和推理的方法处理动态和静态的知识,实现了有效的知识检索;文献[5]提出了面向领域本体的查询扩展模型,总结出了 5 种应用于语义检索系统中的查询扩展方法;文献[6]借助所建立的新闻领域本体和启发式规则,提出了一种语义检索方法,获得了较高的查准率;文献[7-8]主要研究了基于距离、内容和属性的相似度计算方法,用于计算领域本体的概念相似度。

虽然基于本体的语义检索方法取得了一定的进展,但是大多数方法是利用本体进行关键字的语义扩展查询,忽略了属性和实例等语义关系及应用程序层面的本体建立。此外,没有充分利用本体的推理功能,以发现本体中概念和实例之间隐含的语义关联,弱化了检索效果。针对上述问题,需要在基于本体的语义检索中建立新的语义检索模型,并通过引入推理来发现隐含的语义关联。为此,利用农业领域本体中概念之间的语义联系和结构差异,结合语义推理和语义相似度,提出了一种基于本体的农业领域语义查询模型,是对传统的语义检索的补充和提升。

## 1　本体的构建及一致性检测

### 1.1　本体的构建

农业本体是农业领域中概念、概念间的相互关系以机器能理解的形式化语言表示和组织农业知识的模型。从本质上说,本体是一个客观事实的集合,而这些集合是实现语义信息检索的基础。本体的构建是一项复杂的系统工程,目前没有统一的本体构建的方法和规则。Gruber 提出本体构建的 5 个原则,即本体的定义具备清晰性、完整性、一致性、最大单向可扩展性和最小编码相关性[9]。对于领域本体的构建,还应遵循标准化建设原则、本体的复用原则、协作原则和评建结合的原则。借鉴相关构建本体的方法[10],依据农业领域知识的特点,给出农业本体构建流程,如图 1 所示。

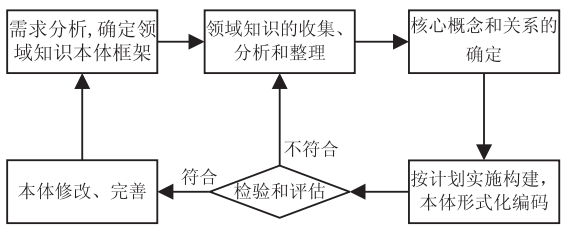


图 1　农业领域本体构建方法流程

根据上述的构建步骤,在学习了很多相关枸杞病虫害书籍和大量文献资料的基础上,结合枸杞病虫害领域专家建议,以宁夏地区常见的枸杞蚜虫、枸杞红瘿蚊、枸杞瘿螨等 51 种枸杞害虫和根腐病、炭疽病、白粉病等 15 种枸杞病害为研究对象,以诊断和防治为研究目标,抽取领域中的重要概念、属性及实例,用 Protégé 工具构建了一个枸杞病虫害本体。本体的类结构如图 2 所示,共计 37 个本体类,基本涵盖了实际生产中主要的枸杞病虫害种类。该本体中有 7 个数据属性和 12 个一级对象属性用于描述枸杞病虫害的基本信息,包括 51 个害虫实例,15 个病害实例及其他类的实例。

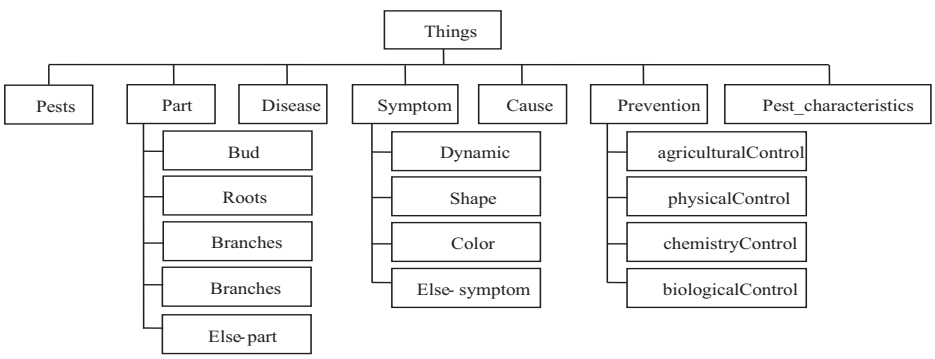


图 2　枸杞病虫害本体类关系模型

### 1.2　本体一致性检测推理

本体构建时手动添加本体或本体合并难免会出现本体不一致,当概念相似度的计算基于本体的一致性时,语义检索才有意义。因此研究如何处理本体的不

一致很有必要[11]。推理是计算机对本体知识理解的一种重要表现,利用本体推理可以获取更准确的语义关系。基于规则的推理(rule-based reasoning)是一种将领域专家的专业知识和经验抽象成为推理规则的基于谓词逻辑的产生式系统。该推理方法比较直观,推理过程简单,同时推理效率比较高,因此采用基于本体的规则推理方法来实现农业本体推理。推理思路是首先要详细分析一下领域本体的语义关系,接下来在本体知识库中进行本体推理规则设计,制定规则库,然后依靠推理机按照一定的推理算法从既有事实推理出隐含知识,并用于语义检索。推理规则是实现语义检索的关键技术之一,利用领域本体中的语义关系和语义性质,如互逆性、传递性等的逻辑特点,设计出有效的推理规则,以应用于推理过程。推理规则语法:

RuleName: $T_1, T_2, \cdots, T_n \rightarrow P$

其中,RuleName 为规则名; $T_i(i = 1,2,\cdots,n)$ 为已经存在的三元组知识; $P$ 为可以推导出的三元组知识。

在推理规则中,如果左边前提知识为真,则可以得到右边的结论知识。将本体的推理规则分为两类:通用规则和领域规则。通用规则是指与领域无关的推理规则,即所有本体都要用到的规则。例如用于确定多概念间的父子关系的传递性规则定义如下:

(? c1, rdfs: subClassOf, ? p), (? p, rdfs: subClassOf, ? c2)→(? c1, rdfs:subClassOf,? c2)。

确定通用规则后,还需要考虑概念属性的具体语义,将通用规则具体化,形成领域规则。领域规则是指与领域相关的规则,实例之间的关系类型取决于其所在领域,需要领域专家参与确定,是对通用规则的补充。例如,枸杞根腐病的症状为:枸杞病株叶片泛黄、萎垂;剖检病株根、茎部,能够看到患部变褐至黑褐色,部分皮层腐烂、脱落,露出木质部,构建了相应的诊断推理规则,格式如下:

(? x, rdf:type, Wolfberry), (? y, rdf:type, Disease), (? z, rdf:type, Roots), (? y, harmsOn, ? z), (? z, hasColor, black brown), (? z, hasDS, rot), (? u, rdf:type, Leaf), (? y, harmsOn, ? u), (? u, hasColor, yellow), (? u, hasDS, sag), (? v, rdf:type, Branches), (? y, harmsOn, ? v), (? v, hasColor, brown), (? v, hasDS, Cortex fall off)→(? y, rdf:type, Ceitocybe bescens)。

其中,x、y、z、u、v 分别为类 Wolfberry(枸杞)、Disease(病害)、Roots(根部)、Leaf(叶片)、Branches(茎)的实例;harmsOn、hasColor、hasDS 等为属性关系。

目前,利用领域本体语义关系进行的推理主要是使用一些推理机来完成的,通过推理引擎去解析本体库中的知识概念,运用推理机根据相应的概念和推理规则进行规则匹配,从而获得新的知识概念。语义推理可分为前向链推理和后向链推理两种方法[12],使用前向链推理算法,采用 Jena 作为推理机进行推理,利用其提供的 DIG 接口实现推理,推理过程如下:

(1)构建领域本体概念集合以及推理规则集合。

(2)从已知概念展开,根据需要来选择用到的推理规则。

(3)若无规则匹配-触发时,则推理终止;若出现多条推理规则,利用相关策略进行选择。

(4)当有规则被触发时,进行推理,并将新事实添加到概念集中。

(5)重复第(2)步。

## 2 概念相似度计算方法

在语义检索过程中,为了获取准确和全面的检索结果,通常使用本体中的术语来表达用户的检索需求,判断本体中的术语与用户检索条件在语义上的匹配程度,即需要计算术语间的相似度。语义相似度[13]是指两个或两个以上的不同概念间具有相近的特征。若有本体中的两个概念 $c_i$ 和 $c_j$,它们之间的相似度用函数 $\mathrm{sim}(c_i, c_j): S \times S \rightarrow [0,1]$ 表示。目前的语义相似度计算方法主要是基于本体的概念,没有综合考虑影响术语间语义相似度的因素和充分利用本体结构知识的问题,不能满足本体库中语义相似度计算的需要。利用本体固有的树形结构,结合语义距离、上下位概念重合度、概念节点层次深度等影响语义相似度计算的因素,建立一种新的基于本体的语义相似度计算模型,使之能够满足本体知识库中语义相似度计算的需要。

(1)基于概念语义距离的语义相似度。

语义距离是度量本体中两个概念在语义上的近义程度的方法,在本体树结构中,通过计算两个概念节点间的最短路径来衡量语义距离。语义距离与语义相似度之间是一种简单的反比关系。对于词汇 $c_i$ 和 $c_j$,如果 $\mathrm{dis}(c_i, c_j)$ 为其语义距离,则语义相似度为:

$$\mathrm{sim}_1(c_i, c_j) = \frac{\alpha}{\mathrm{dis}(c_i, c_j) + \alpha} \tag{1}$$

其中, $\mathrm{dis}(c_i, c_j) = \mathrm{Sd}(c_i, \mathrm{LCA}(c_i, c_j)) + \mathrm{Sd}(c_j, \mathrm{LCA}(c_i, c_j))$, $\mathrm{Sd}(c_i, c_j)$ 为概念节点 $c_i$ 和 $c_j$ 在本体树中的最短距离, $\mathrm{LCA}(c_i, c_j)$ 为 $c_i$ 和 $c_j$ 的最小共同祖先节点;$\alpha$ 为一个可调节的参数。

(2)基于上位概念重合度的语义相似度。

上位概念重合度度量领域本体中两个概念之间在语义上的重合程度,它指两个概念相同的上位概念数量与所有的上位概念数量间的比率,显示了两个概念的祖先节点的相似度。上位概念重合度与语义相似度呈正比,两个概念的上位概念越多,重合度就越大,相

应的语义相似度越大,反之亦然。对于两个概念 $c_i$ 和 $c_j$ , $N(c_i)$ 和 $N(c_j)$ 分别为概念 $c_i$ 和 $c_j$ 的上位概念集合,集合中元素的数量与本体树结构中节点 $c_i$ 和 $c_j$ 与根节点"Thing"的最短路径中所包含的节点数相等。$N(c_i) \cap N(c_j)$ 表示 $c_i$ 和 $c_j$ 相同的上位概念集合, $N(c_i) \cup N(c_j)$ 表示 $c_i$ 和 $c_j$ 所有的上位概念集合。由于在信息论中采用非线性函数来评估语义相似性更好,因此,利用对数函数计算概念 $c_i$ 和 $c_j$ 之间的上位概念重合度,公式如下:

$$\mathrm{sim}_2(c_i,c_j) = \log_2\left(1 + \frac{|N(c_i) \cap N(c_j)|}{|N(c_i) \cup N(c_j)|}\right) \quad (2)$$

(3)基于概念层次深度的语义相似度。

利用概念的层次结构可以计算概念之间的语义相似度。一般来说,本体树结构中处于同一层次的两个概念所含的信息量相似,当两个概念间层次和增加,语义相似度会变大,反之,当两个概念所在层次差增加,其语义相似度会减小。对于两个词汇 $c_i$ 和 $c_j$ ,利用概念层次结构计算语义相似度的公式如下:

$$\mathrm{sim}_3(c_i,c_j) = \frac{L(c_i) + L(c_j)}{2 \times d_{\max}(|L(c_i) - L(c_j)| + 1)} \quad (3)$$

其中, $L(c_i)$ 和 $L(c_j)$ 分别为概念 $c_i$ 和 $c_j$ 的层次; $d_{\max}$ 为本体树的深度。

综合考虑本体结构树中各个因素的影响,结合上述语义相似度计算方法,最终的语义相似度计算方法如下:

$$\mathrm{sim}(c_i,c_j) = \begin{cases} \alpha\mathrm{sim}_1(c_i,c_j) + \beta\mathrm{sim}_2(c_i,c_j) + \gamma\mathrm{sim}_3(c_i,c_j), c_i \neq c_j \\ 1, c_i = c_j \end{cases} \quad (4)$$

其中, $\alpha$ 、 $\beta$ 、 $\gamma$ 为调节系数,取值范围均为(0,1],且 $\alpha + \beta + \gamma = 1$。

## 3　基于本体的语义检索模型

### 3.1　语义检索模型

建立基于领域本体的语义检索模型,首先根据枸杞病虫害领域具体的知识结构,构建了领域本体。然后采集枸杞病虫害领域文档,通过预处理将文档进行标注,建立枸杞病虫害知识资源库。利用枸杞病虫害领域本体中概念之间的语义联系和结构差异,结合语义推理和概念相似度建立语义检索模型。该模型主要由系统界面、本体库、知识资源库、语义扩展和推理、语义检索等模块组成,如图 3 所示。

(1)用户界面:该功能主要实现查询用户和语义检索系统的信息交互,用户利用检索界面输入相应的查询关键词,系统处理后返回查询结果。

(2)本体库:本体是语义检索的核心,对于原查询词的语义扩展和资源库语义信息的标注至关重要。为了使用户能够对领域知识理解一致,实现知识的共享和本体的重用,通过从相关书籍、领域专家和本体学习等途径获取本体信息,构建本体。本体库定义了农业领域中的概念、关系以及实体和属性集合。

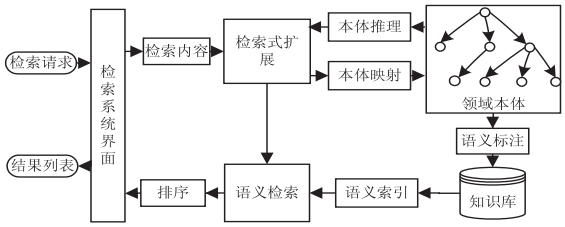


图 3　基于本体的语义检索模型

(3)知识资源库:该模块提供了可供语义检索的信息资源。利用网络爬虫在 Web 上爬取网页文档,然后在页面中找出本体中的实体,包括本体中的概念、属性和实例等,进行相应的语义标注,用领域本体中各种概念的语义关系来描述文档的语义,为资源文档建立基于本体的索引,以便对信息资源进行语义检索。

(4)查询扩展处理:对用户输入的查询词进行分词等预处理后,该功能模块根据领域本体库信息,把原查询词与本体的内容进行映射,进行语义相似度计算和查询语义扩展。与此同时,利用本体中的各种语义关系,设计推理规则,进行知识推理,得到新的更能反映用户检索意图的检索式,从而提高了检索精度[14]。

(5)语义检索模块:按新的检索式对知识资源库进行检索,根据查询实例与文档的相关度和相似度进行排序,并将排好序的查询结果返回给用户。

### 3.2　语义检索过程

根据用户的检索要求进行语义检索,过程如下:

(1)对输入的用户检索请求进行分词处理;

(2)利用分词后的结果,判断检索词是否为本体库中的概念和实例,如果是则进行知识检索,如果不是,则根据农业领域本体中存在的语义关系和设计的推理规则,结合语义相似度计算方法对用户检索词进行语义扩展;

(3)语义扩展后,用得到的检索词进行检索操作。语义检索引擎根据和原检索词相近或相似的新的检索词进行语义检索;

(4)按相似度从大到小排序后输出检索结果,并将结果列表输出到用户页面。

## 4　实验及结果分析

为了对提出的模型进行实验验证,从 http://www.nyyy.cc/、http://wolfberry.forestry.gov.cn/和 http://www.qhgq.org/三个大型的枸杞农业网站中获取相关网页,以这些 Web 页面作为信息资源库。实验前

对资源进行了相应的语义标注,使其能满足语义检索的要求。实验使用传统的基于关键字的检索方法(M1)和基于本体的语义检索方法(M2)分别对语料库进行检索,以对比两种检索方法的性能。与传统的信息检索系统一样,基于本体的语义检索模型的目标也是在资源耗费较少的前提下快速检索到准确而全面的结果,因此对检索系统的评价也从效果和效率方面进行。效果方面采用的评价指标包括查准率、查全率和 $F$ 值。其中,查准率是检出的正确结果总量与检出的结果总量的比率,查全率是检出结果的总量与系统中相关结果总量的比率, $F$ 反映了查准率和查全率的平衡的综合评价指标,通常与检索系统性能呈正比关系[15]。效率方面主要对时间开销和响应速度进行测试比较。

从表 1 中可以看出,基于本体的语义检索方法不论查全率还是查准率,均优于基于关键字的检索方法,因为基于关键字匹配的检索技术仅仅是关键字字型的匹配,不提供语义支持和规则推理,无法获取语义关联的结果和隐含的知识。而提出的方法主要实现了关键字语义层面上的匹配和推理,可以检索出与关键字语义相关的知识,因此各个评价指标总体上比基于关键字的检索方法要高。

表 1 语义检索实验结果 %

| 算法 | 查准率 | 查全率 | $F$ |
|---|---|---|---|
| M1 | 60.50 | 20.39 | 30.50 |
| M2 | 72.29 | 33.40 | 45.69 |

随着本体中的类和实例等数量(本体树结构中节点数)的增加,需要耗费更多的时间来遍历本体树结构,因此整个语义检索系统耗时就增加了,时间开销曲线如图 4 所示。

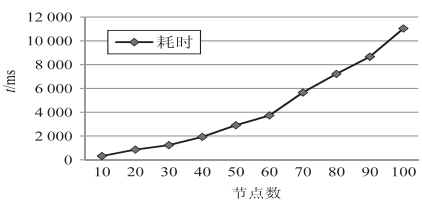


图 4 语义检索方法时间开销

## 5 结束语

传统的信息检索模型是基于字符串匹配,缺乏语义,极大限制了检索的查准率和查全率。为此,利用本体的语义结构和语义推理的能力,在研究基于枸杞病虫害领域本体的语义检索系统模型及其关键技术的基础上,提出了一种基于枸杞病虫害领域本体的语义检索模型。其主要工作包括:构建了领域本体库并使用语义规则和推理引擎对本体进行一致性检测,改进了语义相似度计算方法并实现了基于领域本体的语义检索模型。实验结果表明,与基于关键字的检索模型相比,该模型有效可行,是完善知识检索方法的一种尝试,为农业科技知识服务平台提供了一种有效的检索方法。随着本体应用的不断深入,还需要对现有检索模型进行进一步优化,以提高检索的整体效率。



**参考文献:**


[1] 杨月华,杜军平,平 源. 基于本体的智能信息检索系统[J]. 软件学报,2015,26(7):1675-1687.

[2] Zammali S,Arour K,Bouzeghoub A. Using ontologies to build testbed for peer-to-peer information retrieval systems[C]// 27th international conference on advanced information networking and applications. [s. l. ]:IEEE,2013:1033-1040.

[3] Díaz-Galiano M C,Martín-Valdivia M T,Ureña-López L A. Query expansion with a medical ontology to improve a multimodal information retrieval system[J]. Computers in Biology & Medicine,2009,39(4):396-403.

[4] Yoo D. Hybrid query processing for personalized information retrieval on the semantic web[J]. Knowledge-Based Systems,2012,27(3):211-218.

[5] Liu Z Y,Chen J X,Li X,et al. Design and application for the model of semantic query expansion based on domain ontology[J]. International Journal of Modelling, Identification and Control,2012,16(3):277-284.

[6] Kallipolitis L,Karpis V,Karali I. Semantic search in the world news domain using automatically extracted metadata files[J]. Knowledge-Based Systems,2012,27(3):38-50.

[7] Batet M,Sánchez D,Valls A. An ontology-based measure to compute semantic similarity in biomedicine[J]. Journal of Biomedical Informatics,2011,44(1):118-125.

[8] 王旭阳,万 里. 信息检索中语义相似度算法研究[J]. 计算机工程与应用,2014,50(10):124-127.

[9] Studer R,Benjamins V R,Fensel D. Knowledge engineering, principles and methods[J]. Data and Knowledge Engineering, 1998,25(2):161-197.

[10] 郑业鲁,何绮云,钱 平,等. 基于本体的农业知识管理系统构建方法[J]. 中国科学:信息科学,2010,40(S):196-204.

[11] Huang Z,Harmelen F V. Using semantic distances for reasoning with inconsistent ontologies[C]//International conference on the semantic web. [s. l. ]:Springer-Verlag,2008:178-194.

[12] 李贯峰,李卫军. 基于 SWRL 的枸杞病虫害本体知识推理研究[J]. 江苏农业科学,2016,44(11):399-402.

[13] 刘宏哲,须 德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学,2012,39(2):8-13.

[14] 苏依拉,吉亚图,窦葆媛. 基于蒙古语课程领域语义 Web 的推理与检索方法的研究[J]. 计算机工程与科学,2016,38(2):376-385.

[15] 张乃静,鞠洪波,纪 平. 基于本体的林业领域语义查询扩展模型[J]. 计算机系统应用,2016,25(3):151-156.