

苏玉宁,姜 艺,陈贺胜,等. 基于 Ontology 的农业科学领域知识库构建[J]. 江苏农业科学,2018,46(5):194-198.
doi:10.15889/j.issn.1002-1302.2018.05.052

基于 Ontology 的农业科学领域知识库构建

苏玉宁¹,姜 艺²,陈贺胜³,朱俊武²

(1. 扬州大学农学院,江苏扬州 225000; 2. 扬州大学信息工程学院,江苏扬州 225009; 3. 扬州大学物理科学与技术学院,江苏扬州 225002)

摘要:本体(ontology)的概念起源于哲学领域,在农业信息检索领域中,由于本体可用以解决知识概念表示和知识组织体系方面的问题,因此本体概念引起了农学界专家的高度关注。发达国家在农业科学领域已经建成一些很成熟的领域本体库,并得到了实际应用,为加快这方面的工作,我国在“十一五”计划中,将开展以网络农业信息资源组织为主的农业本体构建技术研究列入其中,因此,农业本体研究是响应国家号召,大力发展和提高我国农业技术和服务水平的重要措施。以农业科学领域中的油料作物——油菜为主要对象,构建农业领域本体知识库。首先在中文科技期刊全文数据库中检索包含油菜的论文题目作为基本语料,然后对检索到的题目利用汉语词法分析系统(institute of computing technology, Chinese lexical analysis system, 简称 ICTCLAS)进行分词分割,按词频出现频率筛选农业科学领域的关键词语并进行定义和细化,最后利用 Protégé 软件构建农业科学领域知识库。构建简单的农业科学领域本体库模型,目的是为农业领域实现网络信息快速检索和提高农业信息共享水平打下坚实基础,同时为开展以农业本体服务为目的的农业本体论研究与应用作初步探索。

关键词:本体;农业科学;油菜;汉语词法分析系统(ICTCLAS);中图法

中图分类号: S126 **文献标志码:** A **文章编号:** 1002-1302(2018)05-0194-05

本体(ontology)的概念起源于哲学领域^[1],20 世纪 90 年代以来,本体的概念被逐步引入人工智能、图书情报和知识工程等领域。由于本体通过对词语的严格定义和词之间的关系来确定词汇的精确含义,因此用本体建立的词汇模型可以让机器理解 Web 页面的语义,可以解决语义层次上 Web 信息共享和交换,因此在语义 Web 中,本体具有非常重要的地位。在农业信息检索领域中,由于本体可用以解决知识概念表示和知识组织体系方面的问题,因此本体概念引起了农学界专家的高度关注。在农业科学领域,发达国家已经建成一些很成熟的领域本体并使其得到了实际应用。为加快这方面的工作,我国在“十一五”计划中,将开展以网络农业信息资源组织为主的农业本体构建技术研究列入其中。因此,农业本体研究是响应国家号召,大力发展和提高我国农业技术和服务水平的重要措施。

1 农业科学领域关键词语的获得

收稿日期:2016-10-11

基金项目:江苏高校品牌专业建设工程一期项目(编号:PPZY2015A060)

作者简介:苏玉宁(1975—),女,陕西宝鸡人,硕士研究生,馆员,研究方向为计算机技术。E-mail: ynsu@yzu.edu.cn。

1.1 关键词语的来源

要建立一个农业科学领域本体库(或知识库),这样做的目的是为实现农业信息知识 Web 页面自由检索。领域本体库必须描述关键概念以及概念与概念之间的关系。因此,建立领域本体库的首要工作是列举出农业科学领域中的所有关键概念。农业科学领域知识十分庞大,为保证建立的本体库在检索时的查全率与查准率比较理想,领域本体创建就要尽量包括本领域中尽可能多地概念,尽可能多的把本领域中一些重要概念都包括进去,只有这样才能建立起一个实际可用的农业科学领域本体库。因此,本体概念的选择至关重要。为了使本研究选择的具有代表性、说服力,本研究采用的关键词语均来自网络中文科技期刊全文数据库,这是由于科技期刊中的论文大多数为本领域专家的研究成果,他们对本领域的概念比较熟悉和了解,也是本领域专业术语的讲解者和本领域未来发展局势的判定者。关键词语是在中文科技期刊全文数据库中检索从 1989 年至 2015 年收录在农学学科,题名或关键字中包含油菜的论文题目。数据库显示共有符合条件的 19 721 条记录结果,本研究摘录论文题目中包含油菜的共 11 242 条记录,作为建立农业科学领域本体库的分词对象。论文题目检索部分结果如表 1 所示。

选择油菜作为检索条件是因为根据骨架法构建本体的流

[9]孙忠富,杜克明,郑飞翔,等. 大数据在智慧农业中研究与应用展望[J]. 中国农业科技导报,2013,15(6):63-71.

[10]许世卫. 农业大数据与农产品监测预警[J]. 中国农业科技导报,2014,16(5):14-20.

[11]王文生,郭雷风. 农业大数据及其应用展望[J]. 江苏农业科学,2015,43(9):1-5.

[12]陶雪娇,胡晓峰,刘 洋. 大数据研究综述[J]. 系统仿真学报,

2013,25(增刊1):142-146.

[13]涂新莉,刘 波,林伟伟. 大数据研究综述[J]. 计算机应用研究,2014,31(6):1612-1616.

[14]张浩然,李中良,邹腾飞,等. 农业大数据综述[J]. 计算机科学,2014,41(11A):387-392.

[15]郭承坤,刘延忠,陈英义,等. 发展农业大数据的主要问题及主要任务[J]. 安徽农业科学,2014,42(27):9642-9645.

程^[2-3],首先要确定本体应用的目的和范围。因为创建本体的大小和研究领域的大小呈正相关关系,以期建立的农业科学领域本体模型的目标大小是 1 万~2 万个论文题目作为研究范围,油菜作为关键检索词刚好满足要求。

表 1 论文题目检索结果(部分)

序号	论文题目
1	“秦杂油 19”春油菜区全程机械化栽培技术
2	“泮油 737”直播油菜经济施肥试验简报
3	“秦优 10 号”油菜优化施肥技术研究
4	“双低”油菜稻田免耕撒直播高产栽培技术
5	“双低”油菜高产栽培技术
6	“双低”油菜绘写“秀美江西”
7	“双低”杂交油菜产量构成因素与产量的相关分析(英文)
8	“四个坚持”推进油菜生产机械化
9	“下施上喷”防油菜冻害
10	“一菜两用”油菜的食用
11	“一促四防”防治油菜菌核病的效果评价
12	“油菜花父子”:情系科研三十载.
13	“油菜花父子”的追梦日记
14	“油菜—水稻”全程机械化高产栽培技术
15	“油蔬两用”油菜品种的灰色关联度评价研究
16	“浙大 619”油菜新品种高产创建和主要栽培技术探讨
17	“浙大 619”油菜新品种攻关田单产超 250 kg/667 m ² 栽培技术研究
18	10 个白菜型冬油菜品种(系)在静宁县引种试验初报
19	15% 精喹禾灵乳油防除春油菜田野燕麦试验
20	10 个白菜型冬油菜品种(系)在静宁县引种试验初报

1.2 关键词语的处理

建立本体库的目的是实现网上信息资源自由共享和检索,而要实现这样的目的,首先要让机器理解人类的自然语言,只有机器理解了人类的自然语言和文字,才能使人与机器的交流成为可能。在人类的自然语言中,词是最小的能够独立活动的有意义的语言成分^[4],所以对于中文来讲,将词确定下来是理解人类自然语言的第 1 步。在英文的行文中,单词之间是以空格作为自然分界符的,而中文只有句和段可以通过明显的分界符来划界,能够独立表达意义的词没有一个形式上的分界符。在计算机检索中常常说到中文比英文要复杂得多、困难得多,究其根本原因就是中文要通过分词这道难关,只有攻克了这道难关,才有望赶上并超过英文在信息领域的发展水平,所以中文分词意义重大。

中国科学院计算技术研究所研制的汉语词法分析系统(institute of computing technology, Chinese lexical analysis system,简称 ICTCLAS)^[5]可以进行中文分词、词性标注、命名实体识别、新词识别,同时支持用户词典、繁体中文,支持汉字内码扩展规范(Chinese internal code specification,简称 GBK)、UTF-8(8-bit unicode transformation format)、UTF-7(7-bit unicode transformation format)、统一码(unicode)等多种编码格式,是目前应用范围较广且最受欢迎的汉语分词系统。正是基于它的上述优点,本研究采用 ICTCLAS 作为分词工具,按以下 3 步确定农业科学领域关键词语。

首先,对在中文科技期刊全文数据库中检索到的与油菜有关的 11 242 条论文题目利用汉语词法分析系统进行分词操作。分词的部分结果如图 1 所示。



图 1 分词结果(部分)

其次,为了便于处理,对分词所得的所有词语词性作简单化处理。即将所有一类词性下面分的二类和三类词性统一都按一类词性处理,例如名词分为 1 个一类,6 个二类,5 个三类,分类如下:

n 名词

nr 人名

nr1 汉语姓氏

nr2 汉语名字

nrj 日语人名

nrf 音译人名

ns 地名

nsf 音译地名

nt 机构团体名

nz 其他专名

nl 名词性惯用语

ng 名词性语素

以上名词分词词性在本研究词频统计中均按一类名词词性进行统计,即 n 名词。动词有 1 个一类,9 个二类,分类如下:

v 动词

vd 副动词

vn 名动词

vshi 动词“是”

vyou 动词“有”

vf 趋向动词

vx 形式动词

vi 不及物动词(内动词)

vl 动词性惯用语

vg 动词性语素

以上动词分词所得词性在本研究词频统计中均按一类动词词性进行统计,即 v 动词。其他词性作同样处理,不再一一说明。

传统的分词算法^[6]分为三大类:基于统计的方法、基于词典匹配的方法和基于语义理解的方法。本研究对分词所得的关键词语是基于统计的方法确定的,分词所得词语部分统计结果如表 2 所示。

最后,特殊词语的处理。对所有分词词语中除了根据词

频统计外,例如数字、度量单位、英文和一些不能独立表示确定概念的单词词语即使符合筛选条件(词频≥3)也作为非关键词,不作概念分析和定义。对可以表达确定概念,而不在农业科学分类中的词语不作归类。对词频≥3 的 2 651 个词语的词性按照上述规则简单化处理后除去一些无用词语,共得到 1 524 个关键词语,作为要定义和分析的词语对象。

表 2 分词所得词语词频部分统计结果

序号	分词所得词语	词语在论文题目中使用频率 (个)	词性
1	油菜	10 624	名词(n)
2	技术	2 535	名词(n)
3	研究	1 766	动名词(vn)
4	栽培	1 666	动名词(vn)
5	甘蓝	1 378	名词(n)
6	影响	1 193	动名词(vn)
7	高产	1 142	名词(n)
8	杂交	1 092	动名词(vn)
9	试验	927	动名词(vn)
10	分析	861	名词(n)
11	产量	749	名词(n)
12	品种	666	名词(n)
13	防治	629	动名词(vn)
14	不同	610	形容词(a)
15	优质	564	区别词(b)
16	生产	545	动名词(vn)
17	病	542	名词(n)
18	直播	513	动名词(vn)
19	菌核	474	名词(n)
20	选育	450	动词(v)

2 农业科学领域分类体系选定和建立

2.1 农业科学领域分类体系的选定

目前,本体创建还没有一个统一的方法论作为指导,创建本体还处于一个无序状态之中,而要建立一个相对完整的领域本体,不但要提取和捕获这个领域中大量的概念,还要对这些概念进行语义冲突和二义性处理,这些既单调又乏味的工作必须有领域专家的参与才能较好的实现,而许多领域本体研究人员并非本领域专家,这一直是领域本体研究中难于突破的一个问题,寻找一个被本领域大多数专家认可的分类体系成为当前的主要任务。由权威机构审定通过的叙词表首先进入人们的视线。

2.1.1 叙词表 叙词表就是将文献、标引人员或用户的自然语言转换成规范化语言的一种术语控制工具,是概括各门或某一学科领域并由语义相关、簇性相关的术语组成的可以不断补充的规范化词表^[7-8](GB 13190—1991《汉语叙词表编制》)。叙词表和本体的相似之处在于它们都是用来描述和组织特定学科知识的,都包含词间关系、类间关系和概念之间关系等,所以,研究从传统叙词表向本体论转化的方法,一直是各个学科领域的热点问题。

国内对叙词表转化的研究正处于热点阶段,目前已转化为本体原型的主要有《国防科学技术叙词表》和《中国农业科学叙词表》的一部分。中国农业科学院科技文献信息中心的常春博士基于《中国农业科学叙词表》的“作物大类”,构建了

一个有关食物安全的本体原型。

叙词表是一种规范的科学语言,其术语组织结构单一、语义关系明确。但叙词表只是一个词汇库,不是知识库,另外叙词表结构保守,不能经常进行修订,缺乏新兴学科、边缘学科的代表术语,难以及时反映学科的发展趋势,面对快速发展变革的社会和快速更迭的信息技术,叙词表有些力不从心。

2.1.2 中图法^[9] 除了叙词表向本体库转化以外,将中图法转化为本体库也是研究的一个方向。由于本体创建首先要对本领域主要概念进行分类、定义和细化,而编制分类法工作量巨大,且也不易被大多数人接受,所以在传统分类法基础上改造原有分类体系为大多数研究者所采取,目前,基于中图法的分类体系主要有以下几类。

(1)郭书普提出的分类方案将行业和信息属性相结合,农业信息行业在这个分类中被简化为 4 个门类,分别为种植业、养殖业、林业及其他行业,下设 10 个大类,59 个子类;农业信息属性被分为 7 个大类,分别是农村社会经济信息、农业科学技术信息、农业生产资料信息、农产品市场信息、农业地理信息、农业空间信息和农业机构与人才,下设 140 个类目,类目下可以细分类目^[10]。

(2)骆浩文等的线分法主要面向种养业,将农业科学信息分为 5 个等级,4 个大类,10 个中类,71 个小类,966 个细类和 11 个属性类,分类采用中图法及 9 位编码法,共分 5 个层次,第 1 层次用英文字母表示,与大类的代码相对应,其他 4 层采用 01~99 表示,各码位代表相应的类别,并以此为基础编制了省级 DB/T 344—2006《种养业信息分类与代码》的地方标准^[11]。

(3)魏清风等依据中图法改造农业科学及相关学科类目编码方法时提出适合网络农业信息分类编码方法,这种编码方法将农业科学信息分为 4 级,信息编码采用 4 个码段,“S”为码段中分类首号,这是延续中图法的编码规则,码段 2 表示信息媒体类型,码段 3 为 8 位分类号(由 4 级各 2 位类号组成),码段 4 为记录序号^[12]。

以上基于叙词表和中图法的农业科学分类法为研究农业科学分类编码标准提供了理论基础,也是对这种编码的可行性进行的初次尝试,但由于农业科学自身季节性和地域性特点比较强,基于叙词表和中图法的分类法所分割的类别知识零散,使它对农业问题的表现力不强,不能涵盖所有农业过程中所涉及的问题,同时,这 2 种方法虽然被业界人员所熟知,但对业外人员来说,还是专业性太强,对使用能力较低的农业人员来说,这种分类法与农业实际需求结合还有一段距离。为了改变分类法面临的这种状况,须要根据农业信息服务系统的特点和所服务的人员对象特点,构建相适应的分类体系。

2.2 农业科学领域分类体系的建立

概念分类层次是将领域概念进行分类组织,对概念进行分类是为了便于用这些概念描述所要建立的本体领域中概念之间的类属关系,并将领域中的具体概念模块化。传统建立分类概念的层次结构^[13]有 3 种方法:自底向上法、自顶向下法和综合法。本研究中的农业科学领域本体库构建采用综合法,具体构建过程:首先,采用中图分类方法和概念分类——自顶向下法建立农业科学领域分类体系第 1 层、第 2 层和第 3 层,借鉴《中国图书馆分类法》(第 5 版)建立的农业科学领

域分类体系基本框架有 9 个一级类,75 个二级类,439 个三级类,其中第 1 层如表 3 所示。

其次,在所建立的农业科学领域体系第 1 层(即 9 个一级类)的基础上,按照中图法分类规则建立领域体系第 2 层。在农业领域体系建立过程中,分类采用中图法,即 9 位编码法,共分 3 个层次,第 1 层用英文字母“S”加数字 1~9 表示,即 S1,S2,S3,⋯,S9;第 2 层在前面大类的基础上再按数字 1~9 编码,第 2 层编码与第 1 层大类的代码相对应,即 S11,S12,S13,⋯,S19;第 3 层依次类推,即 S111,S112,S113,⋯,S119,各码位代表相应的类别。基于中图法并利用 Protégé 软件建立起的农业科学领域前 3 层部分体系框架如图 2 所示。

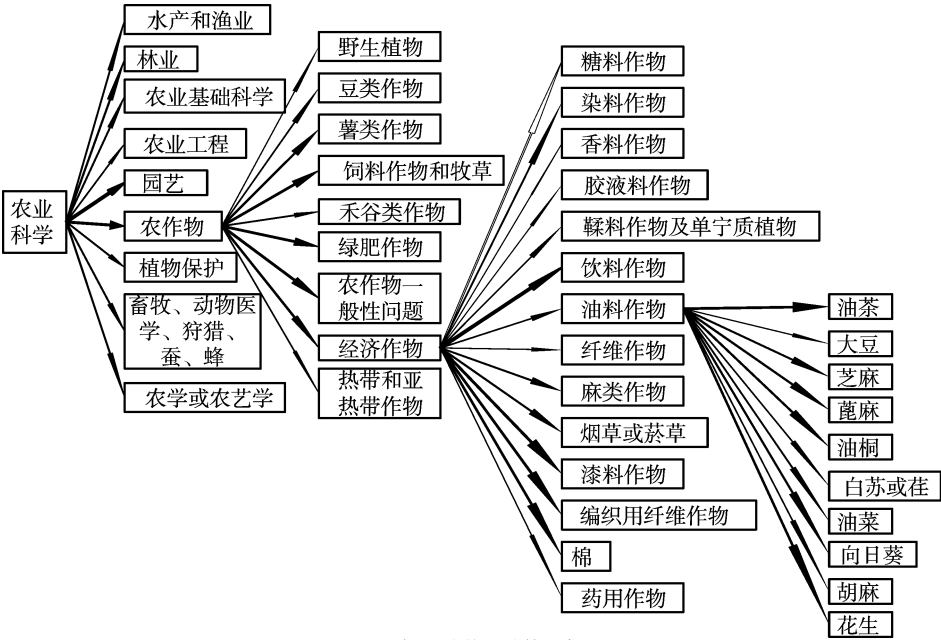


图2 农业科学领域体系框架

3 利用 Protégé^[14-15]创建和生成农业领域知识库

3.1 农业科学分类法转化成农业科学领域知识库

采用自底向上法对本研究选定的 1 524 个关键词语逐一分析,进行概念的定义,然后将这些细化的概念逐层组织在高一级的概念之下,形成一个等级层次结构。然后,使用本体描述语言可扩展标识语言(extensible markup language,简称 XML)来描述农业详表中的词语及词间关系,从而将农业科学分类法向农业领域本体库转化,由于 Protégé 软件在构建本体库时,分类关键词语中不能有特殊符号,因此,在农业科学分类法转化成农业科学领域本体库时遵循的原则是将符号“()”转化为“或”,符号“、”转化为“和”,英文字母翻译为中文词语,例如“3S”翻译为“遥感”等。

为了保证所建立的农业领域本体概念的明确性(explicit),即本体所使用的概念及在这些概念之上的约束都有明确的定义,没有二义性,对中图法分类中概念重复的地方以首次定义概念为准,即如果同一概念在几个大类中均有出现,以第一次出现为准,以后出现的这一概念均列入已出现的这一大类之中,不再另行定义概念。为了简化对农业领域本体的建立,对中图法 4 层以上概念不再细化,统一列入最近上层一类,即第 3 层。如果有些分类词语位于层次较低,深度达

表 3 中图法建立的农业科学 9 个一级类

序号	农业科学分类
1	S1 农业基础科学
2	S2 农业工程
3	S3 农学或农艺学
4	S4 植物保护
5	S5 农作物
6	S6 园艺
7	S7 林业
8	S8 畜牧、动物医学、狩猎、蚕、蜂
9	S9 水产和渔业

10 层以上,则采用逐层回推的方式,最终将此概念或术语回推到第 3 层。建立的农业本体领域库如图 3 所示。

3.2 农业领域本体构建的分析与改进

本研究建立的农业科学领域本体库是为实现农业科学信息网络检索作的基础性研究。在基于中图法建立农业科学领域本体库的构架过程中,还有许多工作要做,尤其是细化本体领域概念工作,由于时间和人力有限,对基于中图法的农业科学领域本体细化比较粗糙,而要建立真正实用可行的农业科学领域本体库需要农学领域专家的参与,这样才能够从整体上对本领域有一个正确把握,也会使建立的领域本体库范围大小更加合乎整个学科未来的发展趋势。本研究对分词所得概念分析相对偏少,这是由于分词的词语相对于整个农业科学领域来说分布比较分散,要正确划分分词所得词语在本领域所处的位置,而又完全依靠人工来做,几乎是不可能完成的事情。如果能实现计算机自动归类或大多计算机归类、人员辅助归类,工作量才会下降。

4 结束语

本研究介绍的基于中图法构建农业领域知识库的方法是半自动构建方法。借用已有的中图分类方法可以获得领域知识以及概念关系,使本体构建有一个很好的起点。目前,很少

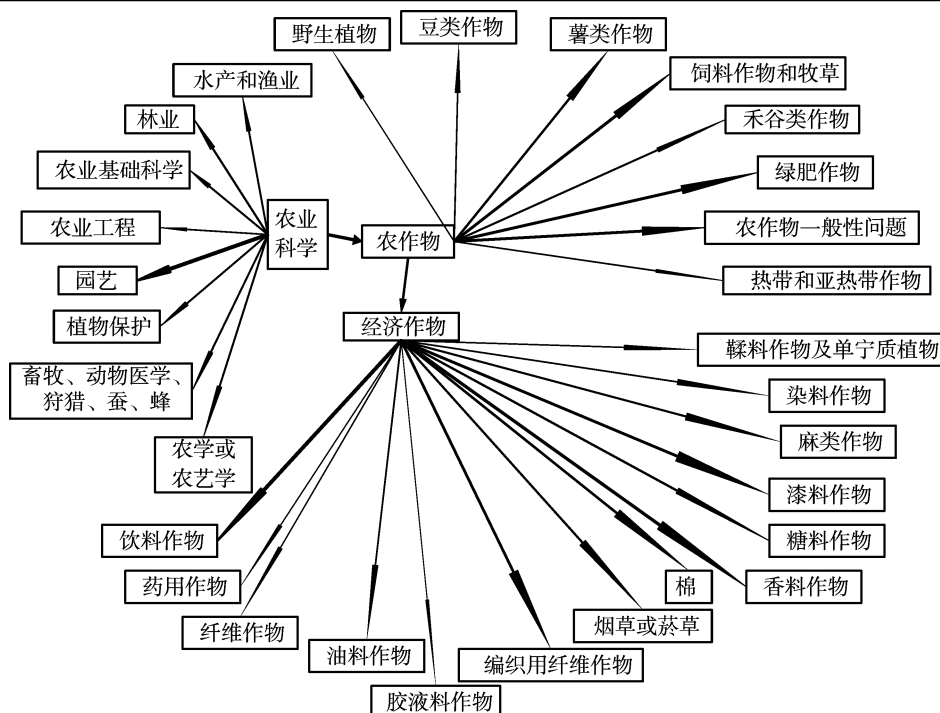


图3 农业科学领域本体部分库框

有现存的不经修改就能被复用的本体。本体的创建是个费时费力的过程^[16],目前还没有一个完整的工程化、系统化的方法来支持创建本体,为了使创建的本体能够实现有效利用,大多领域本体在创建时都邀请本领域专家参与。现存在的通用、大规模本体很少,大多本体只是针对具体应用领域创建。在领域本体的实际应用中,不同领域本体之间常常须要进行映射、扩充与合并处理,以及根据具体需要从一个比较大的领域本体中提取满足要求的小领域本体。这些都对建立领域本体提出了比较高的要求。此外,知识的快速老化,要求对先前构造的本体快速作出相应的增加和删除,以保持本体与现实存在知识的一致性,这都是在本体研究中所面对的现实问题。

参考文献:

- [1] Guarino N. Formal ontology and information systems [C]// Proceedings of formal ontology in information systems. Italy: Trento, 1998:3-15.
- [2] 刘宇松. 本体构建方法和开发工具研究[J]. 现代情报, 2009, 29(9):17-24.
- [3] 杨秋芬, 陈跃新. Ontology 方法学综述[J]. 计算机应用研究, 2002, 19(4):5-7.
- [4] 熊回香, 夏立新. 汉语分词技术综述[J]. 图书情报工作, 2008, 52(4):81-84.
- [5] 孙铁利, 刘延吉. 中文分词技术的研究现状与困难[J]. 信息技术, 2009(7):187-189.
- [6] 来斯惟, 徐立恒, 陈玉博, 等. 基于表示学习的中文分词算法探索

- [J]. 中文信息学报, 2013, 27(5):8-14.
- [7] 陆汝钤, 金芝, 陈刚. 面向本体的需求分析[J]. 软件学报, 2000, 11(8):1009-1017.
- [8] Bateman J A. Upper modeling: a general organization of knowledge for natural language processing [C]//Proceedings of the 5th International Workshop on Natural Language Generation, 1990.
- [9] 国家图书馆编辑委员会. 中国图书馆分类法[M]. 5版. 北京: 国家图书馆出版社, 2010:447-524.
- [10] 郭书普. 网络农业信息分类和编码的研究[J]. 农业图书情报学刊, 2003(6):139-141.
- [11] 骆浩文, 曾志康, 黄樑, 等. 基于网络的农业科技信息分类编码标准体系研究与应用[J]. 农业图书情报学刊, 2007, 19(3):150-154.
- [12] 魏清风, 贺立源, 黄魏, 等. 网络农业信息资源元数据研究及其著录管理系统开发[J]. 现代情报, 2009, 29(2):52-56.
- [13] 战学刚, 林鸿飞, 姚天顺. 中文文献的层次分类方法[J]. 中文信息学报, 1999, 13(6):20-25.
- [14] van Harmelen F, McGuinness D L. OWL web ontology language overview [J]. World Wide Web Consortium (W3C) Recommendation, 2004.
- [15] Stuckenschmidt H, van Harmelen F, Fensel D, et al. Catalogue integration - a case study in ontology - based Semantic translation [M]. 1st ed. Amsterdam: Vrije Universiteit Amsterdam, 2000:29-55.
- [16] 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.