领域语义信息检索研究

——以竹藤领域为例

彭琳 著

非外借

科学出版社

领域语义信息检索研究
——以竹藤领域为例

彭 琳 著

科学出版社

北 京

内 容 简 介

本书以竹藤领域为例，以实现基于植物外形特征的竹藤种类鉴别为信息服务目标；利用领域术语自动识别技术、不确定性知识表示方法、语义信息检索技术等相关理论和技术，分别对竹藤信息中的数值型数据和文本型数据的语义信息检索展开研究；完成竹藤外形特征标本数据库、竹藤领域本体库和竹藤领域语义信息检索模型的构建，实现竹藤领域信息语义关联检索。

本书可作为计算机应用、语义信息检索、农业信息技术等相关专业的研究人员、研究生、高年级本科生阅读。

图书在版编目（CIP）数据

领域语义信息检索研究：以竹藤领域为例 / 彭琳著. —北京：科学出版社，2018.8

ISBN 978-7-03-058399-4

Ⅰ. ①领… Ⅱ. ①彭… Ⅲ. ①信息检索-研究 Ⅳ. ①G254.9

中国版本图书馆 CIP 数据核字（2018）第 171154 号

责任编辑：余 丁 / 责任校对：郭瑞芝

责任印制：师艳茹 / 封面设计：蓝 正

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京画中画印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 8 月第 一 版 开本：720 × 1000 1/16

2018 年 8 月第一次印刷 印张：5 3/4

字数：105 000

定价：58.00 元

（如有印装质量问题，我社负责调换）

前　言

随着信息技术的快速发展，各行业、领域数据量的增长都达到了前所未有的速度。最大限度地集成和利用各类信息资源，快速、完整、智能地提供各种信息服务，已成为领域语义信息检索研究的新需求。从目前的研究现状来看，针对语义信息检索技术的研究大多处于起步阶段，研究多停留于探索性的理论研究，其中关于领域术语集的确定、领域本体的组织结构，以及领域语义信息检索模型的构建都尚未成熟。同时，也缺乏相应的领域语义信息检索实际应用案例。为了弥补这些研究不足，作者撰写了本书。

本书结合竹藤领域知识自身的特点和研究的需求，对利用语义检索技术进行竹藤辅助研究中的若干关键技术进行详细的分析和探讨，内容组织如下：

第一章，绪论。阐述本书的研究背景和意义，概括介绍本书的主要研究工作以及贡献，最后说明本书内容章节的组织。

第二章，相关研究综述。主要就语义信息检索、面向农业领域的语义信息检索以及植物鉴别方法的研究现状进行综述，并结合竹藤语义信息检索研究的应用需求，对这些技术目前存在的问题进行分析和讨论。

第三章，基于云模型/TOPSIS 的植物鉴别检索方法。首先对数量分类法在植物鉴别中的应用进行研究；然后结合竹藤领域知识特征，提出基于云模型/TOPSIS 的植物鉴别方法，并通过竹藤领域的实际应用验证方法的有效性。

第四章，基于互信息/条件随机场的中文领域术语识别方法。研究领域术语自动识别方法，分析现有领域术语自动识别方法存在的不足，提出基于互信息/条件随机场的中文领域术语识别方法；并将该算法应用于竹藤领域术语自动识别中，通过实际案例对算法的特点进行验证和分析。

第五章，竹藤领域语义信息检索模型。对竹藤领域知识的表达和度量、语义检索模型的设计等问题进行研究。将语义相关度扩展和领域术语权重相结合，提出基于相关度扩展的竹藤语义信息检索模型，实现对检索过程的改进和优化，达到提高检索查准率和查全率的目标。

第六章，总结与展望。总结本书的研究工作，并对进一步的研究进行展望。

本书素材主要来源于博士期间的研究工作。在此，感谢我的博士生导师——上海大学计算机工程与科学学院的刘宗田教授。他在我的学业以及撰写工作中倾注了无数的辛勤汗水和心血。虽然我离他对我的要求还有一定距离，但是在他的

细心教导下，我各方面都取得了长足的进步，也才有了此书的雏形。同时，感谢我的硕士生导师——云南农业大学大数据学院的杨林楠教授，在此书后期撰写过程中给予的支持和帮助。最后，我要衷心感谢我的父母、姐姐和姐夫以及我可爱的外甥，是他们给予我生活上和精神上无微不至的照顾和默默无闻的支持，使我能勇敢地面对研究和撰写中出现的一个接一个的困难和挑战，他们是我永远的坚强后盾。感谢所有关心我和爱我的人。

由于学科发展日新月异，新成果不断涌现，本书疏漏之处在所难免，敬请批评指正。若有任何建议，欢迎与本人联系。

彭　琳

2018年4月于昆明

试读结束，需要全本PDF请购买 www.ertongbook.com

目　录

第一章 绪 论

1.1 研究背景

随着互联网、大数据、物联网等现代科学技术的发展，伴随着时间的推移，各行业、各领域的数据量正在迅速膨胀。2006 年，全球产生了 161EB（1EB= 1024PB）的数据，2007 年产生了 280EB 数据，2011 年全球被创建和复制的数据总量为 1.8ZB（1ZB=1024EB），预测到 2020 年，全球将拥有 35.2ZB 的数据量。各种类型的海量数据的产生，对人类的数据驾驭能力提出了新的挑战。如何最大限度地集成和利用各类信息资源，从不同领域的海量数据中发现新知识，快速、完整、智能地提供各种信息服务，已成为领域语义信息检索研究的热点。

由于不同专业、不同领域、不同人员对信息的认识不同，因此从理解层面将分散的领域知识，依据应用需求的不同，提取、融合、处理后提供给用户的领域语义信息检索技术的研究还不够成熟。本书拟选取竹藤领域为研究对象，以竹藤种类的快速鉴别为实际应用需求，以实现竹藤领域语义信息关联检索为目的，对领域语义信息检索相关理论和方法展开研究。

竹藤是竹类和藤类植物的合称。竹藤是植物王国中的一个大家族，在世界森林资源中占有相当重要的地位。它是集经济效益、生态效益与社会效益于一体的重要的非木质森林资源，是木材短缺情况下的主要替代性资源。目前，全世界约有竹林面积 1700 万 hm^2，竹种 1200 余种，主要分布于东亚及其邻近地区，少数分布于非洲和南美洲等，因其具有生长迅速、产量和经济效益高等特点，故被誉为“绿色金矿”而倍受各产竹国的高度重视[1]。藤是棕榈科鳞果亚科省藤族中 13 个属 600 多种植物的统称，属棕榈科，有天然藤和人工藤 2 种。藤与其他棕榈科植物最大的区别是它们多数为藤本，攀缘于其他植物上，与棕榈科其他品种的高大树干相比，有很大的分别。从外表来看，藤的叶片与竹有点相像，但不论是叶脉或茎的形态都完全不同。藤类植物原产于热带地区，主要分布在旧大陆，即亚洲和非洲。在亚洲，大部分集中在亚洲的南部，如老挝、斯里兰卡、孟加拉国、马来西亚、菲律宾、印度尼西亚、印度和中国等。在非洲的分布范围较广，如西非的贝宁、塞内加尔、加纳、科特迪瓦、尼日利亚，中非的刚果、喀麦隆、加蓬

和中非共和国等。各个藤种的分布范围很不均匀，自数千至一二百万平方公里不等[2]。

中国是世界竹子的分布中心之一，是竹子种类最丰富、分布最广的国家。据第 4 次森林资源普查结果，我国现有竹林面积 379.08 万 hm^2，主要分布在我国南方的 17 个省（市）。全世界共有竹种 70 多属，1000 多种，而我国除引种栽培者外，已知有 37 属，500 余种（含变种），许多竹种为我国特有，其中特有竹分类群就有 10 属 48 种[3]。同时，我国也是藤类植物的重要分布地区之一。藤是种密实坚固又轻巧坚韧的天然材料，具有不怕挤、不怕压、柔韧有弹性的特性[4]。在我国南方，自古以来人们习惯于使用藤条做成橱、柜、几、案、屏、架、椅、桌和床等家具。竹藤产品及其副产品在我国的农业生产乃至整个国民经济及人民生活中都有着广泛的用途和发展前景[5-14]。

竹藤属于特殊的植物类群，具有特殊的经济、生态和社会价值。对竹藤进行研究不仅有助于人们更好地理解植物区系的起源、种系分化及其演化进程，而且对保护我国生物多样性和可持续地合理开发利用竹藤生物资源具有极为重要的实践意义。

目前已有研究表明，植物的分布与经度、海拔、大气环流、地形、温度、降水等多种环境因素有着密切的联系[15-20]。利用地理信息系统（geographic information system，GIS）技术、差距分析（gap analysis，GAP）、多元分析方法等，对植物周边环境进行综合考察，通过对区域内植物多样性及其分布格局、成因、区划和“热点”地区确定等方面进行研究，才是植物研究和保护工作的重点。但是，目前我国竹藤的分布格局、生理特性，以及相关的地理、气候和野外生境等信息资料存在着存放分散、集成整合度低，数据标准不统一、不系统、共享性差、查找难等问题，极大地制约了对竹藤的大尺度、系统研究。具体地讲，主要存在以下问题：

① 竹藤领域知识涉及专业较多，各类信息资源分布在不同数据库和专业网站中，各种资源检索系统方法不尽相同，研究人员需要掌握各种不同界面的数字资源系统检索技术，花费大量时间和精力去分别浏览、检索、汇总各类信息，造成了这些信息资源综合利用程度偏低，制约了竹藤的大尺度、全面系统的研究。例如，目前包含竹藤相关信息的文本和图片就分散存放在国家农业科学数据共享中心（www.agridata.cn）、中国数字植物标本馆（www.cvh.ac.cn）、中国植物主题数据库（www.plant.csdb.cn）、中国植物科学网（www.chinaplant.org）、中国植物图像库（www.plantphoto.cn）等几十个专业数据库中。

② 虽然已有一些数据库包含了丰富的竹藤专业知识，存储了大量的竹藤相关科技信息，但由于这些数据库专业性强，并且只能采用关键词检索方式进行查询，

因此增加了操作者对数据库的操作难度，降低了数据库的使用效率，造成了资源浪费。其主要表现为：

第一，对操作者要求高。这些数据库的使用者必须能熟练地运用计算机，了解数据库检索界面，掌握检索策略；同时，必须对主题、关键词、机构、全文、题名等一般的检索概念和检索途径要有所了解和掌握。但是，这些数据库的使用者不仅包括有着较强专业知识的领域专家和农技人员，还包括农民和非农业科技工作者，这些人员大部分很难根据要求准确地输入检索词。

第二，关键词检索方式要求检索提问必须严格按照规定的格式输入，只有当完全匹配时才可得到搜索结果，这种在字面上与检索提问标识保持一致的检索方式，很难实现在内容上和概念上检索到满足用户需求的检索结果，将导致检索结果的查全率和查准率较低。

③ 竹藤领域信息存在于不同数据库和专业网站中，这些信息以 TXT、HTML、XML、RTF、PDF、PSZ/PS 等不同数据格式存在，以中文、英文、拉丁文等不同语言形式表达，而目前的检索工具大多不能提供异构数据的信息检索；同时，这些来自互联网及专业数据库的信息一般只能实现基于关键字的简单检索，只有查询词出现在文档中才可能被检索到，这种查询方法不具备语义查询能力，经常出现与用户查询请求相关的文档，由于使用同义词而无法被检索出来的情况。在实际应用中，由于关键词的不匹配，用户不得不频繁地人为更换查询词，才能检索出想要的结果。

④ 在竹藤相关研究过程中，需要对收集到的大量资料进行筛选，挑选出典型的内容用于研究。这一工作目前主要凭借人的主观进行，具有一定的主观经验因素。同时，由于竹藤相关文档内容丰富，语义关系复杂，因此要全面收集与研究课题有关的竹藤资料，必须对竹藤知识间的关联关系进行分析。其关联关系包括两个层次：一是竹藤文本表示的语义知识间的关联关系；二是竹藤文本与相关文献间的关联。关联分析可提高竹藤领域资料收集的综合性和全面性。

⑤ 对竹藤种类快速鉴别方法的研究还不够。在对竹藤进行相关研究和开发的过程中，对未知植物的快速鉴别是一切研究工作的基础。鉴别工作对于区分竹藤种类、探索竹藤种类间的亲缘关系、阐明竹藤系统的进化规律具有重要意义。目前，植物鉴别方法主要包括传统鉴别法、分子鉴定法、光谱鉴定法、数量分类法和图像识别法五种。针对竹藤植物，利用竹藤植物自身的外形特征和其专业语言特点的快速鉴别方法的研究还未见报道。

针对以上问题，本书以实现基于植物外形特征的竹藤种类鉴别为目的，利用领域术语自动识别技术、不确定性知识表示方法、语义信息检索技术等相关理论和技术，分别对竹藤信息中的数值型数据和文本型数据的语义信息检索展开研

试读结束，需要全本PDF请购买 www.ertongbook.com

究，对面向语义信息检索的竹藤信息资源集成共享的解决方案进行探索。以期构建竹藤领域语义信息检索模型，实现竹藤领域信息语义关联检索。本书研究成果不仅为领域语义信息检索模型的构建提供了一种新方法和思路，也为领域语义信息检索的深入研究提供了技术支持和实例借鉴。

1.2 研究意义

随着计算机科学技术的发展，利用信息化技术手段最大限度地集成和利用各类信息资源，快速、完整、智能地提供各种信息服务，已成为研究和保护竹藤的新需求。这对竹藤异构信息集成共享技术和方法的研究与应用，为研究人员提供基于语义信息检索的文本信息和数字信息相结合的个性化服务，具有重要的现实意义。

本书围绕目前我国竹藤信息资源数据格式多样、语义检索效率低下、快速鉴别手段单一等问题，分别对竹藤信息中数值型信息和文本型信息展开研究。针对竹藤领域中的数值型信息，本书在构建植物的外形特征标本数据库的基础上，利用云模型对被测植物的外形特征进行数字化描述，然后再与植物外形特征标本数据库进行比对，从而实现被测植物的初步鉴别，实现其语义检索。针对竹藤领域中的文本型信息，本书选取《中国植物志》中的竹藤相关内容作为竹藤语料库原始语料，建立竹藤语料库，提出一种适用于竹藤领域术语特点的术语自动识别方法，构建一种适用于竹藤语义信息检索模型，最终实现竹藤领域语义信息检索。

本书研究内容解决了竹藤信息资源利用的技术障碍，提高了竹藤种类鉴别效果；为竹藤的宏观深层次研究，提供了新的技术支持和理论依据；同时，鉴于国内领域语义信息检索相关研究还处于起步阶段，尤其是对于不确定数值型数据的语义信息检索和文本型数据的语义信息检索模型的研究还缺乏完善理论基础和应用实例，本书做了一定的探讨。本书所提出的基于云模型/TOPSIS 的植物鉴别方法、领域术语识别方法和领域语义信息检索模型具有一定的通用性，可推广应用于其他专业领域知识的语义信息检索中，为领域语义信息检索的研究和应用提供了理论基础和技术经验，也为领域语义信息检索系统的构建奠定了基础。

1.3 研究内容

本书以竹藤为研究对象，围绕竹藤信息资源研究与应用的实际需求，结合领

域术语自动识别技术、云模型、语义信息检索技术等相关理论和技术，以实现竹藤领域信息语义检索为目的展开研究。重点研究内容如下。

1. 基于植物外形特征的鉴别检索方法研究

为了使研究人员可以对未知植物进行快速鉴别，并从竹藤本体知识库中检索出未知植物相关信息，本书在对传统信息检索方法中的关键词检索方法和基于图像识别技术的植物鉴别方法进行分析研究的基础上，对基于植物外形特征的植物鉴别检索方法开展研究。利用云模型对植物外形特征进行数字化描述，构建基于植物外形特征的植物数字特征表达式，实现植物外形特征信息的定性与定量之间的不确定转换，为基于植物外形特征的植物鉴别检索方法提供理论基础。并将该方法应用于竹子的快速鉴别中，取得了较好的鉴别效果。

2. 竹藤领域术语自动识别方法研究

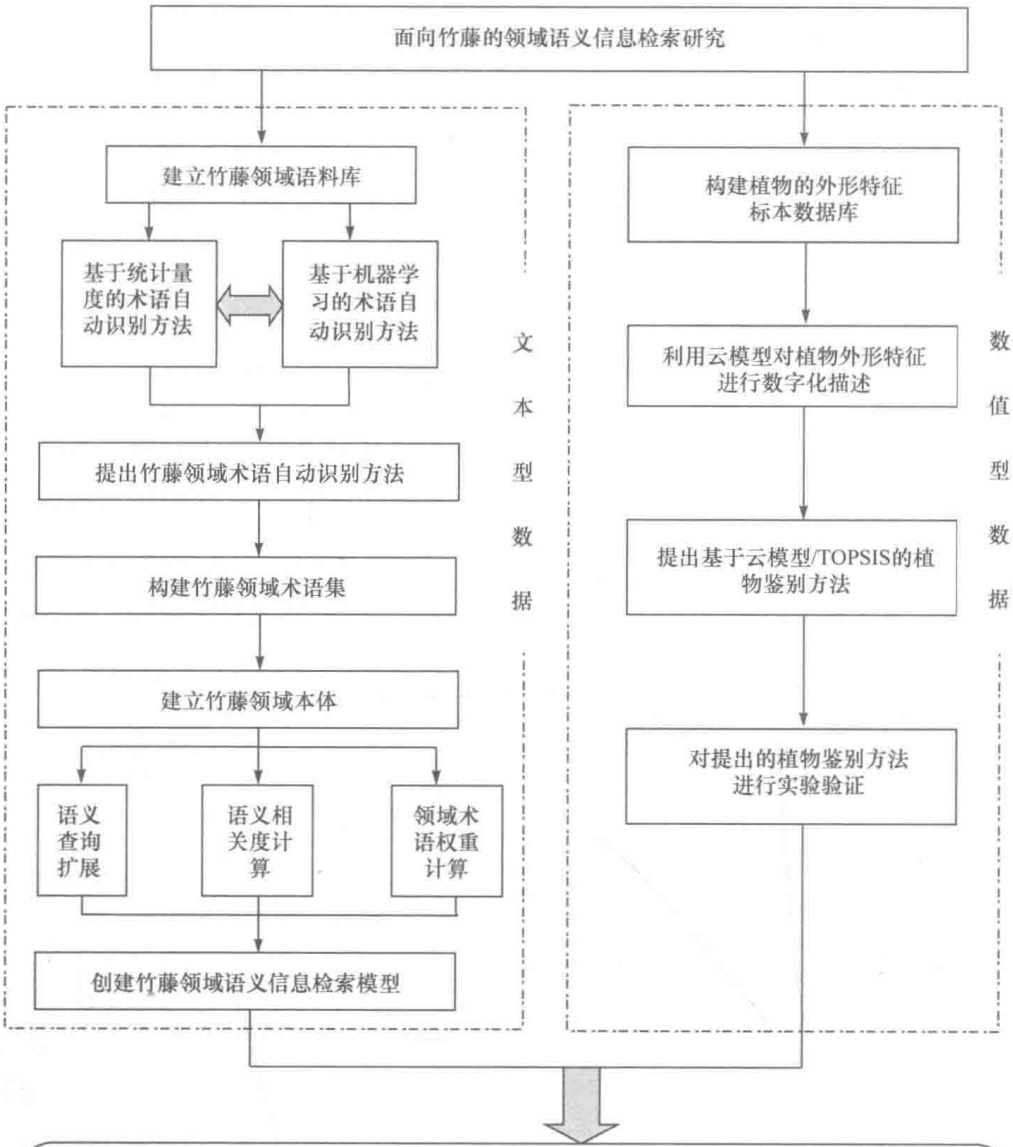
针对目前我国农业本体中的领域术语集普遍借助于主题词表、叙词表和农业领域专家，依靠人工进行构建，代价十分巨大而且进展缓慢的现状，本书针对竹藤领域的语言学特征，在创建竹藤领域语料库的基础上，分别对基于统计量度的术语自动识别方法——互信息方法、基于机器学习的术语自动识别方法——条件随机场，以及两类提取方法相结合的混合术语自动识别方法进行比较研究，提出了一种适用于竹藤领域术语特点的术语自动识别方法，给出竹藤领域术语自动识别的一般方法和处理规范，为其他专业领域术语的自动识别提供参考。

3. 基于相关度的竹籐语义信息检索模型研究

目前我国农业语义信息检索领域的相关研究人员大多将研究重点集中在通过传统的分类法和叙词表，利用基于概念层次和关系规则的查询扩展方式，来提高检索精度和检索效果，忽视了农业领域知识特点，对适用于农业领域知识特点的语义信息检索模型研究相对有限。针对此，本书引入统计语言模型思想，将领域术语权重和概念查询扩展相结合，提出了适用于竹籐领域的语义信息检索模型。

本书以竹藤领域语料库的构建为基础，以植物外形特征的数字化表示模型的建立和面向竹藤领域的术语自动识别方法的研究为手段，以实现面向竹藤领域的语义信息检索为目的展开研究。

1.4 技术路线



- 利用云模型对竹藤外形特征信息进行数字化描述，实现竹藤外形特征信息的定性与定量之间的不确定转换；
- 构建竹藤外形特征标本数据库，实现基于竹藤数值型信息的关联检索；
- 给出适用于竹藤领域的领域术语自动识别方法；
- 建立竹藤领域本体库，实现竹藤领域知识的数字化表示；
- 构建竹藤领域语义信息检索模型，实现竹藤领域信息语义关联检索。

1.5 本书贡献

本书面向竹藤语义信息检索展开研究，以实现竹藤的快速鉴别为应用目标，对植物鉴别技术、领域术语自动识别、不确定知识的表示以及语义信息检索技术进行了综述和分析。针对竹藤领域知识特征，研究了竹藤语义信息的表达和度量，提出了基于云模型的植物鉴别检索方法和基于互信息/条件随机场的中文领域术语识别方法，构建了竹藤领域语义信息检索模型。具体贡献如下。

1. 在植物快速鉴别方法的研究方面

利用云模型对植物外形特征信息进行数字化描述，实现了植物外形特征信息的定性与定量之间的不确定转换，为基于植物外形特征信息的植物鉴别提供了理论基础。

针对竹藤数值型信息，实现了对竹藤种类的识别和检索。依据竹藤外形特征的数值信息，实现了对竹藤的识别和检索，为植物鉴别方法中的数量分类法研究提供了新思路。

2. 在领域术语自动识别方法的研究方面

提出了一种基于互信息和条件随机场的领域术语自动识别方法。该方法将基于统计和机器学习的两类术语识别方法结合在一起，有效地解决了单纯利用统计方法进行术语识别时的数据稀疏问题，同时仅采取了三个特征，对条件随机场模板进行训练，有效地降低了条件随机场的运算时间。

3. 在领域语义信息检索模型的研究方面

提出了基于相关度的竹藤领域语义信息检索模型。从竹藤领域术语权重、语义相关度两个方面，描述检索词概念和竹藤知识之间的相关关系，较好地解决了检索者真实检索意图与竹藤知识之间的“语义鸿沟”问题。

实现了竹藤领域文本信息的关联检索和语义查询扩展。依据竹藤文本中的语义信息，实现了对竹藤种类的识别和检索。

第二章　相关研究综述

竹藤信息检索研究是对竹藤领域知识管理、分析和应用的重要内容。本书将不确定性知识表示方法和语义信息检索技术应用于竹藤领域知识表示和检索中，以竹藤领域知识表示和竹藤领域术语自动识别方法的确定为基础，以竹藤领域语义信息检索模型的构建为核心，以实现基于语义信息检索的竹藤种类快速鉴别为目的展开研究。主要研究工作包括：竹藤领域数值型数据的不确定性表示，竹藤领域文本型数据的语义信息检索方法以及竹藤种类快速鉴别方法等。本章将围绕研究中涉及的语义信息检索、农业语义信息检索、基于本体的农业语义信息检索，以及植物鉴别方法等领域的国内外研究现状展开综述和分析。

本章的组织结构为：2.1 节介绍语义信息检索；2.2 节介绍农业领域语义信息检索；2.3 节介绍基于本体的农业领域语义信息检索；2.4 节介绍植物鉴别方法；2.5 节对本章进行小结。

2.1　语义信息检索

针对当前网络信息缺乏结构化和语义化的问题，万维网的缔造者 Berners-Lee 于 2000 年 12 月在 XML2000 会议上提出了语义网（semantic Web）的概念[21]。语义网作为对当前网络的扩展，它的目标并不是要完全取代现有的网络，而是让网络上的信息能够被计算机理解，从而实现语义层上的智能应用。语义网的出现为实现语义信息检索提供了可能。

语义信息检索的概念是由 Guha 等于 2003 年首次提出[22]，他们认为语义信息检索是研究基于语义网的搜索技术，其目的是通过语义网技术提高当前的搜索性能，并构建下一代基于语义网的新型搜索引擎。语义信息检索一经提出，就引起国内外学术界的高度重视，许多研究者从不同角度对其进行了一些开创性研究。其中，Cohen、Lei 等围绕语义信息检索的框架结构展开研究。耶路撒冷希伯来大学的 Cohen 等[23]设计了一个专门针对 XML 文档的搜索引擎 XSEarch，提出了一套完整的理论；韩国中央大学的 Cho 和 Lee[24]提出语义检索框架中应该包括本体构建、爬虫、索引器、查询语句引擎和可视化五个部件，并将整个搜索引擎分为在线和离线两部分；英国开放大学的 Lei 等[25]介绍了一个语义搜索引擎 Semsearch，同时提出了语义搜索引擎的五层次结构；Rodrigo 等[26]提出了语义网

图形系统，这个系统形成了以语义网为基础的语义搜索引擎，在获得查询结果的同时还可以获得关联信息。王进、Gary 等针对语义检索中搜索优化的问题进行了重点研究。中国科学技术大学的王进[27]提出了一种基于本体的语义信息检索模型，模型采用语义聚类方法对文档进行分类，然后将用户查询要求对应到某一类别中，从而提高语义检索的效率；亚利桑那州立大学的 Gary 等[28]采用了 Marker-passing 的搜索算法，以外部刺激的方式并行地搜索整个语义网络，有效地提高了语义搜索效果。

但从目前的研究现状来看，这些研究都还处于起步阶段，研究大都只停留于探索性的理论研究，其中语义信息检索模型、语义检索系统的构造方法和实现机制都还未成熟。

2.2　农业领域语义信息检索

近年来，我国农业领域研究人员对语义信息检索在农业中的应用也进行了大量探索性研究。例如，中国科学院地理科学与资源研究所的甘国辉和徐勇在国家“十一五”科技支撑计划课题“农村信息协同服务技术研究与集成应用”和 863 计划课题“农业语义检索技术研究”中，对科技文献信息、空间信息和农业网络信息的信息融合技术、农业领域本体构建和农业领域知识的语义检索策略与方法等农业知识语义信息检索关键技术进行了研究；广东省农业科学院科技情报研究所的郑业鲁、李泽和王众等参与了联合国粮食及农业组织（Food and Agriculture Organization of the United Nations，FAO）的农业本体服务研究（agricultural ontology server，AOS）项目中的渔业本体（fishery ontology）的构建工作，并对农业语义检索技术展开相关研究；中国农业科学院农业信息研究所的孟宪学等在农业科学叙词表转化得到农业本体基础上，设计并实现了基于农业本体的智能检索原型系统，进一步完善了传统信息检索系统的功能（国家自然科学基金项目成果）；中国农业科学院农业信息研究所的苏晓路等利用文献计量方法对《中国农业科技文献数据库》中的分类和主题标引进行了分析，对其中的主题词与类目之间的关系进行研究，构建了基于主题词的农业初级本体，并以该本体作为检索知识库，建立了农业科技智能检索系统（国家“十五”重点科技攻关计划项目成果）；平顶山工学院的张玉花与西安工业大学的李宝敏在国家“星火计划”项目“西北农业专家远程信息化服务体系示范”课题中，共同实现了在农业果品领域中的语义智能检索[29]；中国农业科学院农业信息研究所的杨晓蓉在其博士学位论文《分布式农业科技信息共享关键技术研究与应用》中针对农业异构数据源的检索问题，对基于农业领域词典的中文分词方法和基于农业领域本体的语义扩展方法进行了研究，

试读结束，需要全本PDF请购买 www.ertongbook.com