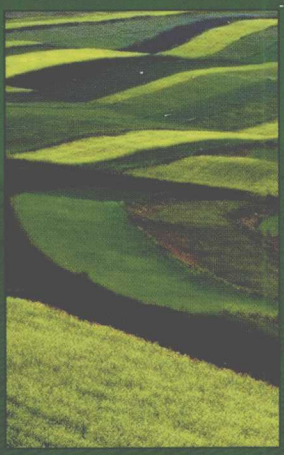

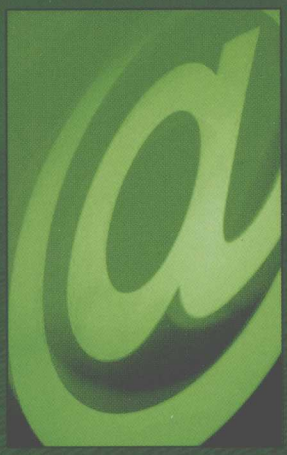

农业信息智能获取技术

岳峻　傅泽田　高文◎著

科学出版社

农业信息智能获取技术

岳 峻 傅泽田 高 文 著

科学出版社
北京

内 容 简 介

本书论述了农业领域信息的专题语义获取技术，共包含四部分内容。第一部分（第1～6章）从技术角度描述如何构建农业信息专用搜索引擎；第二部分（第7、8章）介绍本体理论与知识获取；在此基础上，第三部分（第9～13章）以蔬菜供应链为例，从技术角度提出了实现知识管理和知识语义获取的系统框架；第四部分（第14～18章）以鱼病诊断为例，提出了能够指导基于案例的鱼病诊断的智能案例获取和推理方法。

本书可供信息主管、知识主管、企业及政府管理人员、知识管理系统软件设计开发人员参考阅读，也可作为高等院校知识管理、企业管理和计算机软件等专业研究生与本科生的参考教材。

图书在版编目(CIP)数据

农业信息智能获取技术/岳峻，傅泽田，高文著. —北京：科学出版社，2011.5
ISBN 978-7-03-030860-3

Ⅰ.①农… Ⅱ.①岳… ②傅… ③高… Ⅲ.①农业科学－信息获取
Ⅳ.①G252.7

中国版本图书馆CIP数据核字(2011)第078261号

责任编辑：孙 芳 / 责任校对：郭瑞芝
责任印制：赵 博 / 封面设计：耕者设计工作室

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

丽源印刷厂 印刷

科学出版社发行 各地新华书店经销
*
2011年5月第 一 版 开本：B5(720×1000)
2011年5月第一次印刷 印张：18 3/4
印数：1—2 000 字数：360 000

定价：58.00元
（如有印装质量问题，我社负责调换）

前　言

有效获取和传播农业知识对我国现代农业的发展和新农村建设具有重要的意义。本书着重论述了利用现代信息技术进行农业信息获取的方法。全书共四部分。第一部分重点介绍了农业信息垂直搜索引擎的开发理论和方法。农业信息搜索引擎主要针对当前我国农业发展中所面临的小生产与大市场的矛盾，试图通过利用信息搜索促进农产品电子商务发展来缓解这一矛盾。针对我国农业信息搜索效率低、搜索成本高这一问题，本部分介绍了一种行之有效的服务于农业领域的农业信息垂直搜索引擎的开发理论和方法。与传统的农业信息搜索方法相比，该搜索引擎针对农业领域的特点进行了以下创新：①在分析网页时效性的基础上，提出了一种适用于农业专业领域网页数据采集模型。②在研究农业网页交易信息数据特点的基础上，提出了一种针对具有时空属性的信息表示与抽取模型。③在考虑垂直搜索引擎的索引对象具有领域性问题的前提下，提出了一种面向中文网页特点的高性能双字节倒排索引模型，保证了搜索的时间效率。④在借鉴用户个性化服务技术的基础上，针对当前搜索引擎检索质量存在的一些问题，引入检索结果自动分类与查询自动纠错技术来提高检索精度。第二至四部分别介绍了利用本体理论和智能推理理论来实现农业专业领域信息智能获取的方法。第二部分首先介绍了本体及知识获取的基本理论。第三部分和第四部分分别介绍了知识本体和智能推理理论在构建蔬菜供应链知识获取系统及鱼病诊断案例知识获取中的应用。第三部分从技术角度提出了一套能够实现蔬菜供应链知识获取系统的框架，主要包括蔬菜供应链本体模型、本体模型的形式化表示方法、领域概念获取推理方法、文本获取映射方法、蔬菜供应链知识获取系统的设计与开发。针对鱼病诊断案例的特点，第四部分提出了一套能够指导鱼病诊断案例知识获取系统的框架，主要包括诊断案例知识面向对象表示和向量空间模型、鱼病诊断本体论、鱼病诊断本体的构建与学习、诊断案例知识自动获取模型。这两部分的研究是当前语义网技术及智能推理技术在农业领域的具体应用。农业知识智能获取技术研究是当前信息技术在农业信息领域知识智能获取的探索，对开发专业领域的语义知识获取系统具有一定的借鉴意义。

在本书的撰写过程中，北京语言大学的胡亮老师和中国农业大学的李振波副

教授提供了不少有价值的研究素材，在此表示衷心的感谢！同时，本书出版获得了国家自然科学基金项目(60875039)的支持，并得到了鲁东大学信息科学与工程学院的大力支持和帮助，在此深表感谢！

由于作者水平有限，书中难免存在不妥之处，请同行和读者批评指正。

目　　录

第二部分　本体论和知识获取

第三部分　基于本体论的蔬菜供应链知识获取系统

第四部分　基于本体的鱼病诊断案例知识获取

第一部分　农业信息垂直搜索引擎

第 1 章　国内外农业信息搜索引擎现状

以计算机为代表的信息技术在农业领域的商业化应用改变了传统市场的竞争格局，带来了新的机遇与挑战。尽快提高网络经济下的农业市场效率，增强国家农业竞争力，占领世界经济制高点，已经成为各国政府关注的焦点。但与此同时，我国农产品交易市场体系功能落后，传统农产品流通领域中存在的信息不对称问题导致了中介环节多、交易成本高及流通效率低等问题，极大阻碍了国内农产品交易市场的发展。

具体来讲，农业市场交易中的信息不对称问题主要体现在三个方面：①市场形成价格稳定性差。充分竞争的市场环境是有效价格形成的必要条件。我国农产品批发市场大多数规模较小，地域分散，产品和信息隔离，难以形成交易集中、市场透明度高、竞争充分的市场环境，导致农产品价格波动较大，区域价格差异明显。②市场交易方式原始。我国农产品批发市场仍主要采用“一对一”的对手交易方式，交易规模小，次数多，信息搜寻难度大，市场透明度低，形成的价格不能正确反映供求关系。③市场价格信息系统落后。许多批发市场的信息建设基本处于空白，信息的传递效率和共享度低，对信息的搜集、加工、处理、发布能力低下，容易造成信息的扭曲与失真。

加入 WTO 后，传统的农产品市场交易方式已经无法满足我国农业所面临的持续激烈的市场竞争环境及千差万别的顾客需求，使得中国农业发展所面临的小生产与大市场的矛盾更加突出，成为阻碍我国农业和农村经济健康发展、影响农民增收乃至农村稳定的重要因素（潘明等，2007）。怎样构建新型农产品市场交易模式、降低交易成本、提高交易效益等成为我国农业进一步发展亟待解决的重大现实问题（李晓等，2001）。农业电子商务作为一种运用电子化手段来进行农业商务活动的经济运行方式，能够将农产品交易活动网络化，减少传统方式中的大量中间环节，实现企业-企业、企业-消费者的直接交易，从而最大限度地降低经济活动中的交易成本，提高经济运转的效率和效益（廖咸真等，2005）。但是，传统的农业网上交易方式是商家在自己的网站上发布农产品价格信息，消费者从网站上找到自己感兴趣的农产品，直接与供应商联系购买，这样的模式简单易行，但非常依赖网站的访问量。之后又出现了集中式的农业电子交易中心，即由大量的商家登录自己的商品信息，由网站负责整个站点的宣传工作。尽管这种模式降低了成本，丰富了农产品信息，但消费者的选择范围仍然有限，特别是随着农业电子商务的普及与电子交易市场网站数量的增长，网页规模越来越大，导致用户要从海量信息中查找自

己需要的农产品的难度也越来越大，消费者在众多网站中寻找值得信赖的商家、获得最好的价值、最好的服务等问题也成为制约农业电子商务发展的因素，一般除了选择一些影响力较大的农业网站外，大多只能通过那些具有资金优势的网站的宣传来得到农产品信息。因此，在这种背景下，要求出现一种新的交易模式(郭云升，2007)，能够从多个方面进行评比，包括网站访问量、顾客服务、商品价格、配送、客户反馈等，使得消费者在几秒钟内获得自己所需农产品的详细信息。

为了充分利用与扩展现有的农业信息资源，农业信息垂直搜索引擎综合服务平台将开发农业信息服务系统汇聚成虚拟数据库作为数据来源基础，从各农产品市场中采集原始的或经过分析处理的有用信息，进行筛选、加工、处理，再将各类信息进行整合、发布，并与其他农产品市场进行信息联网，使用户能从同一平台上获得即时、全面、有价值的信息。另外，平台还能提供各类信息增值服务，如信息的搜索、查询，同类产品销量、价格等的汇总、比较等，帮助用户减少信息搜寻成本，提高信息利用率，满足用户的多样化需求。

将所开发的农业信息垂直搜索引擎平台部署到网络上，通过系统应用为用户提供一个实际可用的农产品交易信息搜索比较平台，可以方便用户查找希望购买的农产品，降低消费者网络购物的时间成本，增强便利性；同时，提供丰富廉价的信息服务，打破了商家与消费者之间信息不对称的局面。平台包含的价格比较功能可以为商家提供最低成本的消费者来源渠道；系统还可以通过记录，分析大规模的消费者查找、比较和购买商品的行为过程，为农业企业提供精确细致的分析报告，借以提高企业的运营能力和销售技巧，从而进一步促进农业经济的全面发展，并为带动全国农产品交易的发展及农业信息化做出应有的贡献。

1.1 国内外农业相关的信息搜索引擎

1) 美国农业网络信息中心

美国农业网络信息中心(AgNIC)(http://www.agnic.org) 是美国国家农业图书馆与一些大学、研究机构及政府机构自愿组成的联合体，其每个成员都负责农业科学中某一领域的信息工作，各成员单位间互相提供信息服务，每个成员在享受服务的同时，也有为其他成员提供服务的义务，服务方式主要是通过互联网相互提供电子形式的农业信息和检索服务。

(1) 数据库规模和范围。AgNIC 建有多个较为系统和完整的与农业相关的数据库，目前拥有 1100 多个数据库的记录。AgNIC 早先提供的也是数据库检索服务，此后升级为 Web 检索服务。

(2) 信息采集方式。AgNIC 的各个成员单位在其负责的专业领域下不断维护和增加数据资源，并存储在本地服务器上，同时，他们还要将这些新增加的数据

资源在AgNIC位于Purdue大学的中心服务器上反映出来，在其负责维护的30个大类下面的某个子类中添加相关信息，做好链接使别的单位可以访问。这种数据采集方式不仅使整个系统拥有的数据资源不断增加，同时将收集数据资源的工作由农业各个领域内处于领先地位的成员单位完成，因而能够确保数据的权威性和完整性。

（3）检索功能。AgNIC提供简单检索、高级检索和词表（thesaurus）检索三种界面。简单检索界面支持布尔逻辑检索，默认逻辑关系为逻辑与。高级检索界面比较复杂，有两个检索词输入项，中间有逻辑符号选择项（AND、OR 、NOT）来确定两个检索词的逻辑关系，在下面还有三个选择项：前截断、中截断与完全匹配。高级检索还提供了检索字段选项，用户可以限定从某个字段中进行检索，共有题名、作者、摘要、关键词、主题词、全部字段等6个选择，这也为用户提供了更多的检索途径。由于AgNIC有自己的主题词表，用以描述它所收录的资源，用户输入自己的检索词后，系统会自动列出该词的相关词、上位词和下位词。例如，用户输入“crops”时，系统会提示“crops”的上位词有“plants”，下位词有“field crops”、“horticultural crops”与“specialty crops”，相关词有“crop acreages”、“crop losses”、“crop models”、“crop prices”等，这样就可以实现概念检索。

（4）结果显示形式。AgNIC检索的显示结果十分专业，在显示符合检索条件的记录总数后，再列出各条记录，每条记录首先给出题名，再给出简介、主题词、关键词，最后列出记录来源，即URL地址。

（5）结果排列规则。AgNIC检索结果的相关度排列规则是：首先看每条记录简介、主题词、关键词中出现的检索词的总数，然后根据出现频次的多少高低排序，AgNIC检索认为，检索词出现频次高的记录相关性高，排列在前，检索词出现频次相同的记录再根据记录（网页）创建时间的先后排列。

2）Agriscape

Agriscape网站（http://www.Agriscape.com）于1999年4月在美国普林斯顿建立，主要提供农业及相关产业的导航服务，其目标是发展成为农业信息、农业贸易和农业技术的信息中心。

（1）数据库规模与范围。Agriscape将所收录的内容划分为观光农业、公司、教育科研、图片、有机农业、拍卖、园艺、期刊、图书馆、政府、组织、导航目录等大类，各大类再进行细分，并提供会议、市场、新闻、天气等方面的服务。

（2）信息采集方式。与其他搜索引擎直接用robot攫取网页并自动加入数据库不同，Agriscape的信息完全是手工采集的。编辑人员对网站进行访问和获取，准确地描述网站的内容特征，并将其准确地归入不同类目之中。这种分析方法与robot从网页内容中提取关键词对网页进行描述的方法相比，确保了所收集网页信息的专指性，提高了搜索引擎的查准率，它还允许用户登记网站，在编辑人员审