农业本体论 ——农业知识组织系统的建立

贺纯佩 , 李思经

(中国农业科学院 科技文献信息中心 ,北京 100081)

摘 要:这篇论文描述了农业本体论(Agricultural Ontology Service [AOS])在中国的发展 ,农业本体论服务的特点 ,所采用的技术和创新点。作者还对农业本体论服务的应用进行了评价 ,并论述了农业本体论服务应用给农业领域将带来的社会效益和经济效益。本论文还描述了农业本体论服务所开发的农业文献机器翻译系统 ,以及将多语种词表 AGROVOC 翻译成中文等应用。作者还展望了农业本体论在将来农业信息网络检索中的应用。

关键词:农业 ;本体论 ; 叙词表 ; 主题标引 ;中国

中图分类号:G354 **文献标识码**:A **文章编号**:1002 - 1248(2004)10 - 0041 - 04

Agricultural Ontology : Establishment of Agricultural Knowledge Organization System in China

HE Chun-pei , LI Si-jing

(Scientech Documentation and Information Center (SDIC) ,

Agricultural Ontology : Establishment of Agricultural Knowledge Organization System in China

Abstract : The paper describes the development of Agricultural Ontology Service (AOS) in China , as well as the characteristics , key technology and innovation points of AOS. The authors also make some comments in the application of AOS and the social and economic benefit AOS has rendered to the agricultural sector. The paper mentions the application of AOS in machine translation software packages and the translation of multilingual thesaurus AGROVOC into Chinese language. The authors look forward to the application of agricultural ontology service in network information searching in the future.

Key words : agriculture ; ontology ; thesaurus ; subject indexing ; China

随着第四媒体的突飞猛进的发展 ,网络信息数量以指数形式增长。因特网的利用 ,在当今信息时代给我国农业跨越式发展提供了前所未有的机遇。作为当今世界上最大的信息资源库的因特网已经发展成为我国农业信息传播、获取和发布最为活跃的实用领域。但是 ,当前网络信息检索在当前的结果常常是海量的信息。其中大多数信息与查询的目标相差甚远。在查全率和查准率方面存在很大缺陷 ,严重影响了农业信息的有效获取和利用。农业科学叙词表是计算机信息资源有效管理工具之一 ,对农业信息准确查询作出过重要贡献。在当今网络信息环境下 ,在研究农业科学叙词表的基础上。研究和建立农业科学本体论将促进农业信息查询和利用的效率 ,将本体论应用于农业信息管理系统将促进我国“三农”问题的解决 ,加快我国农业科技和经济的发展。

农业本体论 ——又称为“农业知识概念系统”,在今天的网络时代用于管理海量信息和知识系统变得越来越重要。为了在网络环境下不断容易地提供信息 ,特别是企业内部网络和知识库系统中 ,找到正确的信息变得越来越困难。当今的各种检索工具在检索信息方面相当差 ,而且 ,基本上都是依赖关键词进行检索 ,或浏览各种文献的题目部分。这种没有重点的检索方法常常导致不好的检索结果。

1 国内本体论研究进展

国内有关本体论的研究主要是有关哲学方面的研究 ,也有关于人工智能 ,计算机科学方面的研究。一些计算机科学较强的大学或研究所 ,例如哈尔滨工业大学计算机科学与工程系的廖明宏 ,中国科学院计算所的武成岗等等 ,有一些本体论研究的报道。农业本体论服务项目(AOS)的基本原理是 :在学科专家的参与下 ,建立某一学科的专业本体论 ;收集网络信息 ,按照本体论的原理建立知识库 ;将用户检索请求转换

收稿日期:2004 - 07 - 27

为本体论规则下的概念,在知识库中进行匹配,查找概念含义水平上的信息;然后将检索结果返回给信息查询者。

北京大学的邓志鸿系统地描述了本体论的内容以及目前的应用和研究状况。他们正在探 ontology 技术在图书分类体系和主题词表的技术舢板感建立概念模型,并利用该概念模型进行智能导航。

杨晓青报道了利用 RDFS(Resource Description Framework Schema)建立本体论的方法,提出语义的互操作性将成为语义 Web 的必要条件,同时这种互操作性必须通过现有的 RDF 的方法加以实现,而不是使用 XML 的标记。良好的 RDF(Resource Description Framework)数据模型可以通过人工智能和知识工程中建立语义交互性的方法直接使用。

王泰森(2003)提出了一个基于本体论的全文自动标引方案,从中可以具体理解本体论在全文自动标引中的作用。具体思路是首先将文件转化为文本文件用分词系统将文件语句的词分开,滤去与文章主题内容无关的和高度重复的词,剩下的词通过计算其出现位置和出现频率,经过系数加权,得到不同的函数值,经与本体论词间关系对照,抽出那些函数值较大或平均值超过阈值的一组词,作为自动标引的词。

2 本体论的概念

本体论的英文名称为 Ontology,是一个哲学上的概念,现在广泛应用于知识工程、知识表达、信息检索和知识管理等领域。国际上,本体论的研究非常活跃,甚至被应用到企业集成、自然语言翻译、电子商务、生物信息系统等领域。Neches R. 等人将 ontology 定义为"给出构成相关领域词汇的基本术语和关系,并利用这些术语和关系构成的规定这些词汇外延的规则定义"。GruberT. R. 认为"ontology 是概念模型的明确的规范说明",并在随后又明确"ontology 是对概念化的精确描述"。Guarino N. 又称"本体论的最终目标是精确地表示那些隐含信息,使得它们可被软件系统重用和共享。"在农业信息资源管理领域,联合国粮食与农业组织给了本体论一个简明的定义:本体论是指"包括一个领域中各种标准术语,并对这些术语进行准确定义且明确这些术语间的各种关系"。

本体论在国际上的研究非常活跃,而在国内的研究仍然处于起步阶段。特别是在农业信息管理方面的理论和应用基本上还处于空白阶段。本文的目的是介绍现有本体论研究的进展,分析国际国内本体论研究的特征,通过具体构建一个领域的本体论,示范和讨论本体论的构建方法。同时采用具体实例提出从农业叙词表向农业本体论的转换方法。以及通过对多语种叙词表翻译方法的分析等,最终简明地说明本体论是什么,在农业信息管理中有什么价值,以及如何实现其价值等等。

3 农业本体论服务对相关概念的定义

农业本体论研究涉及到对一些概念的理解和定义,这里介绍在 AOS 研究中的一些相关概念:

本体论(ontology):在人工智能和知识表示领域,本体论的使用范围已经扩大到信息系统模型。本体论使用普通方法描述学科知识并对一个特定的领域提供统一的理解。本体论是一个包括各种术语,术语的定义以及术语之间关系说明的系统。

学科知识(domain knowledge)::学科系指一个限定的知识领域,是实践经验的扩展和深入,也是个人、研究机构在一个特定活动范围内对记载知识的理解。

分类法(taxonomy):分类方法是一种信息组织工具。它可以使用户通过规范的结构和名称术语了解和浏览信息,帮助用户查找信息。

公开标准(Open standards):包括通用标准标记语言(SGML, Standard Generalized Markup Language)、可扩展标记语言(XML, Extensible Markup Language)和超文本标记语言(HTML, Hypertext Markup Language)等等。这些标准方案的目的不是为了确立限制性标准,而是为开放标准活动,以便讨论和促进可通用的标准,在用户中推荐和改进其通用性的方法。

语义网(Semantic Web):通过语义网,万维网上的网站能够了解其它网站上的信息,不是因为严格使用的标准协议或各种各样的格式,而是由于数据映像所产生的所有数据知识。XML 和 RDF 就是促成这一网站的工具。

叙词表(thesaurus):农业叙词表包括英国应用生物科学中心开发的"CAB 叙词表"、联合国粮食与农业组织开发的"AGROVOC 叙词表"和中国农业科学院文献信息中心开发的"农业科学叙词表"等等。这些叙词表都可以作为生产农业本体论的工具之一。

4 农业本体论的用途

农业本体论包括农业的专业概念或范例,是一种对应用知识的正式描述,也是对各种概念的定义。在农业本体论中或农业知识组织系统中,它显示农业各种概念之间的关系,可以用电脑进行加工处理。

农业本体论可以用来对网站进行语义的组织,可以形成知识图,可以引导人们发现知识,是我们能够不使用复杂的布尔逻辑检索式便可检索到我们所需要的信息。

农业本体论可以使我们能够使用电脑进行文字加工,可以使我们能够在网上进行文字挖掘检索,是一种自动标引和文字解释的工具。可以利用全文检索引擎生产出有意义的知识分类。

农业本体论可以在网上进行知识检索,从机读元数据建立动态的目录。还可以进行自然语言加工,使机器翻译效果更佳,还可以使我们能够使用自然语言进行信息检索。

建立农业本体论的目标是可以创造一个专业的共享概念的正式定义。

加工并创建一个学科的本体论。

本体论的获得有两个途径:建立核心本体论;用基础词汇表和学科特异性文本自动抽取本体论知识。

然后将这两种方法的结果合并成一个某学科的本体论。然后进行提纯和扩展。对这种方法获得的本体论进行评价和评估。

5　本体论的优点、模型和方法论

511　本体论的优点

本体论能够帮助我们改善对知识的管理。语义解释的文献,例如经过本体论术语和概念而不是简单的关键词标引的文献能够提供很多优点。首先,本体论文摘能够提供满意的检索结果而不受文献内容变化的影响。文献内容可能改变检索时使用的语言,但是,文献一旦经过本体论语义解释,就不会影响到检索的结果。第二,由于本体论的解释使用的是专业性本体论关键词的语义解释与特有的专业领域相关,因此,检索效率也大大地提高了。一条术语在不同的专业领域可能有不同的意义。在第一步将语义与特有的专业吻合并使用本体论连接知识结构可以进行更有针对性的检索。第三,文献特异性表达不再影响检索结果。在多语种表达的情况下,这是特别重要的。由于几种语言的关键词都用相同的本体论概念连接,因此,提供了相同的意义。无论使用什么语言进行检索,都要创建多语种检索入口,以产生相同的结果。

知识管理的另一个重要问题是文献的分类。至今,这需要主题专家进行长时间的工当今网络上有大量有用信息,需要自动化支持来有效地管理这项工作。本体论在支持自动化机读语义方面起到关键的作用。

512　本体论:模型和表达

在农业本体论服务的前提下,本体论是一个术语、术语定义和这些术语关系的规定的系统。它延伸了典型叙词表的研究,提供了创建无数不同语义关系的机会。在模型前提下,一个概念术语的表达称为词典入口。这些词典入口是无穷的,拥有标签、同义词或词族的特点。每个词典入口至少有二个属性:一个是它所指的概念,另一个是其语种。最后,可以建立概念之间的关系,并用同样的语义入口予以解释。这种研究方法可以描述为一个两个层次的模型。本体论的语义层次与表达层是完全独立的,因此,可以实现文献内容变化后检索的查全率。

513　方法论

本体论方法论集中在文献的实际索取和开发,描述完全的可以重复使用的半自动化的框架。这个框架可以包括在其它生命周期模型中。学科本体论可以使用两种不同的知识索取方法来建立:(1) 创建某一专业的核心本体论;(2) 用叙词表衍生出某一专业本体论。然后,进行对这两种方法而获得的本体论进行合并,最后进行本体论的精选和延伸。

514　创建核心本体论

以联合国粮食与农业组织建立食物安全专业本体论为例,他们使用了 3 位主题专家作为智囊团使用本体论编辑器(SOEP)对欧洲食品宝典和食物安全协议进行了筛选,用网络搜索器对 257 个食物安全领域的网站进行了搜索,这些网站包括政府食物安全信息的网关、共计提取出 67 个概念和 91 个关系,用此作为食物安全的核心本体论。

创建核心本体论主要选择了美国政府食物安全信息网关、美国食物安全与应用营养中心、加拿大食品检疫机构、美国衣阿华州立大学的食物安全项目、美国农业部食品安全与检疫局、食物传播疾病教育信息中心、世界卫生组织等网站共计 257 个食品安全网站的网页。

文献的选择:联合国粮食与农业组织这个项目的主题集是用手工进行选择而获得的 11 篇文献。而采用网络搜索器获得了 5 片文献。此外还采用手工选择方法选择了 8 篇普通文献。

51411　用叙词表抽取本体论知识

联合国粮食与农业组织出版的 AGROVOC 多语种叙词表共有 27 365 个关键词。经过 5 次评估自动选择出 1 632 个常见词汇。所抽取的本体论结构有 504 个概念,共计分为 5 层。

51412　本体论的合并和竞选

从 AGROVOC 叙词表抽取的总共 1632 个术语组成的本体论经过专家筛选得出 12 个新概念 92 个新关系。再与拥有 67 个概念和 91 种关系的核心本体论合并便组成了食物安全本体论的样板,共计有 102 个概念和 183 种关系。

51413 最终本体论样板

核心本体论拥有 67 个概念和 91 种关系,平均每个概念有 1.36 个关系。而食物安全本体论样板有 102 个概念和 183 种关系,平均每个概念有 1.79 个关系。

在经过对本体论词汇进行精选后共计筛选出 3 000个概念。从 100 篇生物安全专业的文献选择出一系列常用词汇。食物安全本体论的创建使用了本体论编辑器(OilModeler)。这个编辑器可以合并和精选所选择出的本体论词汇。共计获得了 102 个概念和 183 个关系。

6 本体论的将来应用

使用本体论来扩展当前的关键词检索是最为直接的应用。以本体论为基础的检索可以将现有的检索步骤分为两个阶段:在实际检索之前或/和获得检索结果之后。使用知识图从语义上安排和组织信息网站。在不使用复杂的布尔逻辑检索式的情况下,指导用户发现知识和比较容易地检索到信息。还可以使我们在网站上用机器进行文本处理。这包括:在网络上进行文本挖掘、自动标引和文字解释工具和可以在网络上建立有意义的分类方法的全文检索引擎。可以用机读媒体建立动态目录的方法,方便网站的智力检索。最后,本体论可以使自然语言机器翻译变得更加简单,还可以使用自然语言进行检索。

在农业科技信息方面,从信息检索角度研究本体论在农业科技数据库中的应用原理和前景。在数据量较小情况下,全文检索在数据库方面也存在明显的问题,表现在用关键词检索时,检索出的文献往往与用户的需求有很大差异。随着数据量的增大,数据库全文检索的弊端将更为突出。而基于本体论构建的数据库检索工具,从理论上可以解决此类问题。本体论是知识的一种表达形式,它将学科知识表示成挖掘算法能够理解的形式,以本体论为出发点去引导数据挖掘过程,从而加快数据挖掘的进程,提高获取知识的效率和质量。本体论将成为数据库框架集成、知识获取和表示的核心。

在农业数字图书馆领域,本体论的应用可以涉及数据资源的自动标引,提供类似与书目的分类功能,提供语义层次的检索,提高检索的查全率和查准率。本体论的构建将是数字图书馆管理的一个重要的研究领域。

本体论在农业机器翻译中也具有巨大的应用前景。开发出的本体论,被证明几乎在自然语言加工,语言知识获得等阶段都具有潜在的价值。

总之,本体论的构建将最终极大地促进农业科技信息数据库、数字图书馆、农业机器翻译、网络信息检索、知识的组织等重要的信息管理环节。

参考文献:

[1] Boris Lauser, Johannes Keizer. A Comprehensive Framework for building Multilingual Domain Ontologies: Creating a proto2type Biosecurity Ontology[C]. Asian Agricultural Information Technology and Management. 2002. 31 - 41.

[2] 常春. 联合国粮食与农业组织 AOS 项目[J]. 农业图书情报学刊,2003(2):14 - 15.

[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] 贺纯佩,李思经. 农业叙词表在中国的发展和农业本体论展望[J]. 农业图书情报学刊,2003(4):16 - 19.

[5] 廖明宏. 本体论与信息检索[J]. 计算机工程,2000,26(2):56 - 58.

[6] 刘柏嵩. 一种面向语义 Web 的数字图书馆框架[J]. 大学图书馆学报,2003(1):13 - 16.

[7] 王泰森. 一个基于本体论全文自动标引方案[J]. 情报科学,2003,21(9):950 - 952.

[8] 武成岗,焦文品,田启家,等. 基于本体论和多主体的信息检索服务器[J]. 计算机研究与发展,2001,38(6):641 - 647.

[9] 杨秋芬,陈跃新. Ontologya 方法学综述[J]. 计算机应用研究,2002(4):5 - 7.

[10] 杨晓青,陈家训. 一种利用 RDF(S)建立本体论的方法[J]. 计算机应用研究,2002(4):54 - 57.

[11] 张晓林. Semantic Web 与基于语义的网络信息检索,情报学报,2002,21(4):413 - 420.