农业本体及本体学习研究

徐济成,李绍稳,张友华
(安徽农业大学 信息与计算机学院,安徽 合肥 230036)

摘 要:目前国际上关于本体学习的研究非常活跃。利用本体学习技术来实现本体的半自动或自动构建就成为克服手工构建本体的困难和大规模开发本体的有效途经。介绍了本体理论和本体学习,综述了国内外农业本体的研究现状,特别介绍了农业本体学习的过程,给出了农业本体学习的关键理论和技术,采用基于统计、隐含语义检索和关联规则的算法提取概念;采用模式匹配和聚类算法提取概念间关系,列举了目前常用的本体学习工具,分析了本体学习结果的评价方法。

关键词:本体;农业本体学习;概念;关系;结果评价
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)08-0212-04

Research of Agricultural Ontology and Ontology Learning

XU Ji-cheng, LI Shao-wen, ZHANG You-hua
(School of Information and Computer Science, Anhui Agricultural University, Hefei 230036, China)

Abstract: At present the research of ontology learning is a trend in the world. The manual construction of ontology exists fatal demerits, so it is an effective way of large scale ontology development to construct the ontology semiautomatically or automatically by ontology learning. In this paper the key theories of ontology and ontology learning are presented, status quo about the research of agriculture ontology in China and abroad is surveyed. The main process and key technologies of agriculture ontology learning are given. Extracting concept based on statistics, latent semantic retrieval and association rules algorithm. Extracting concept based on pattern matching and clustering algorithm. Some popular ontology learning tools are especially introduced. Evaluation methods about the result of ontology learning are simply analyzed.

Key words: ontology; agriculture ontology learning; concept; relation; result evaluation

0 引 言

自20世纪90年代以来,随着知识共享、信息集成、语义Web和Web服务等技术的快速发展,本体研究在计算机领域倍受关注,逐渐成为研究的前沿和热点[1]。本体作为表达知识的共享概念模型,已日渐成为知识工程、知识管理、智能信息集成、信息检索和语义网等多个领域的重要组成部分,特别是自语义Web提出以来,本体正在成为人工智能和知识工程中一种重要的工具,在知识的获取、表示、分析和应用等方面具有重要的意义。

农业科学作为一门复杂的系统科学,其知识系统包含了大量的概念和关系,要从复杂的农业知识系统中抽象出易于重用的领域知识,需要进行有效的概念建模,以便更好地支持基于知识的分布式检索、类比推理和数据挖掘。本体作为一种精粹的知识理论,能够帮助解决这些知识表达问题。

随着对本体的不断研究,出现了许多本体的构建工具,比如从最早的Ontolingua[2], WebOnto[3], 到protégé[4],以及KAON[5]等,本体构建工具日趋成熟完善。但这些工具支持的只是手工构建本体的方式,存在诸多弊端。因此,研究如何从现有信息源,包括文本、词典、遗留知识库、WWW文档等,获取领域知识,以自动方式构建或扩充本体,是开发本体的理想和有效途径。本体学习(Ontology Learning)正是自动或半自动构建本体的一系列方法和技术。目前,把本体学习技术与理论运用于农业本体的建模研究中已逐渐成一种新的研究方向,这对于加快农业的发展将有着积极而深远的意义。

收稿日期:2008-11-24;**修回日期**:2009-03-05
基金项目:国家863计划项目(2006AA10Z249)
作者简介:徐济成(1985-),男,硕士研究生,研究方向为人工智能、本体学习;李绍稳,博士,教授,研究方向为人工智能。

1 本体理论与本体学习

1.1 本体理论

1.1.1 本体的概念

所谓本体,最著名或最多被引用的定义是由Gru-

ber 提出的"本体是概念模型的明确的规范说明[6]"。通俗地讲,本体是用来描述某个领域甚至更广范围内的概念以及概念之间的关系,使得这些概念和关系在共享的范围内具有大家共同认可的、明确的、唯一的定义。Studer 等学者认为本体有四大特征:

(1) 本体是明确的(Explicit):是指概念所属的上位类与在使用次概念时的限制条件应预先得到明确的定义和说明。

(2) 本体是形式化的(Formal):是指本体应该具有机器可读性。

(3) 本体是共享的(Shared):是指在一个本体中,知识所表达的观念、观点应该抓住知识的共性。

(4) 本体是概念化的:它是一个概念体系,又称概念模型(Conceptualization)。

1.1.2 本体的类型

根据本体不同方面的属性,可以将本体分为不同的类别:

(1) 根据详细程度和领域依赖程度两个指标,本体可分为:顶级本体、领域本体、任务本体和应用本体。

(2) 根据本体的形式化程度,本体分为高度非形式化的、结构非形式化的、半形式化的和严格形式化的。

(3) 根据本体的描述对象不同,本体分为特殊领域本体(如农业、医药、地理等)、一般世界知识本体、问题求解本体和知识表示语言本体等。

1.2 本体学习

Alexander Maedche 把本体的结构定义为一个五元组[7]O: = {C, R, HC, rel, AO}。C 和 R 是两个不相交的集合,其中:C 为概念(Concept)集合;R 为关系(Relation)集合;HC 表示概念层次或分类层次,即概念间的分类关系(Taxonomy Relation);rel 表示概念间的非分类关系(Non - Taxonomy Relation);AO 表示使用某种逻辑语言表达的本体公理(Axiom)集。根据这样的结构,形成了一个本体学习的分层模型,见图 1[8]。

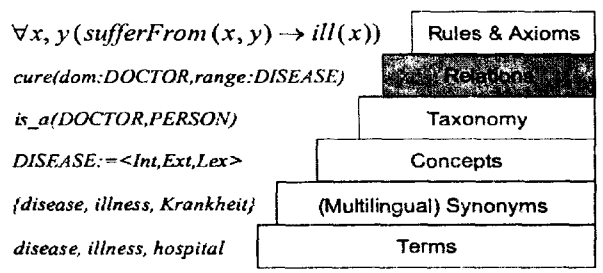

图 1　本体学习分层模型

图中,下面两层 Terms 和 Synonyms 是本体学习的基石,上面三层,即概念(Concepts)、分类关系(Taxonomy)、和非分类关系(Relations)的抽取是本体学习的主要任务[9]。

根据 Alexander Maedche 和 Steffen Staab 对输入的数据类型的区分,本体学习应该包括基于文本的本体学习、基于词典的本体学习、基于知识库的本体学习、基于半结构化数据的本体学习。

(1) 基于文本的本体学习是通过在文本上应用自然语言分析技术来抽取本体,是本体学习的主要方法和基本方法,是本体学习研究中的主要方向。

(2) 基于词典的本体学习,指从机读词典中抽取相关的概念与关系。

(3) 基于知识库的本体学习是从规则的知识库中抽取本体[10,11]。

(4) 基于半结构化数据的本体学习,是指对已经事先定义好结构模式的数据源(如 XML、Schemas、DTD、RDFs 等)进行本体学习。

2 农业本体研究现状

农业本体[12]是一个包含农业术语、定义以及术语间规范关系说明的体系,是农业学科领域内概念、概念与概念间的相互关系的形式化表达。它是一个强有力的农业主题词表,除处理能够提供农业主题词表中内在的基础关系外,还能够创造更多更正式的特殊关系。形式化的农业本体可定义为[13]:

Agri_ Onto = (Onto_ Info, Agri_ Co ncept, Agri_ ConRelation, Axiom)

其中 Onto_ Info 表述对本体基本信息描述,包括本体名称、创建者、设计时间、修改时间、目的和知识来源等本体的元数据信息;Agri_ Concept 就是农业知识概念的集合,Agri_ ConRelation 就是本体中概念的关系集合,包括层次关系和非层次关系;Axiom 包含本体中存在的公理集合。

2.1 国外农业本体研究现状

联合国粮农组织(FAO)自 2001 年起开展农业本体服务(AOS)计划的研究。FAO 建立和维护农业本体分类服务器,其官方网站利用农业本体的结构进行站点内容和结构的组织。最终用户利用 AOS 提供的搜索工具获取自己最终想要得到的信息。

AOS 项目目前构造了三个领域的原始本体[12]:

(1) 渔业本体:渔业本体构建技术主要包括渔业知识本体的设计、存储与查询、推理引擎的构建、知识表示方法、知识录入一致性保障方法、知识关联的方法、知识系统自学习和自修正的方法、批量转换现有数据库的方法、支持多语言数据的方法等内容。目前该本体可以以 OWL 和 RDFS - KAON 文件的格式使用,领域专家主要的工作是修正在本体自动生成后产生的

本体结构错乱问题。

(2) 食品安全领域本体:它的产生依靠半自动化的方法,即首先创建某一领域的核心本体论和根据叙词表衍生出某一领域本体。然后对由这两种方法而获得的本体进行合并,最后进行本体的精选和延伸。食品安全本体的创建使用了本体编辑器,这个编辑器可以合并和精选所选择出的本体论词汇。它从A-GROVOC叙词表中抽取了共计102个概念和91种关系。该本体将被发展为在线词典的原型。

(3) 食物、营养和农业本体:该本体是FAO的营养部门为食物、营养和农业刊物制作的含有简单知识的浏览系统,允许作者以关键词、种类、作者等进行浏览,为用户搜索和浏览FNA的信息提供帮助。该本体的构建包括:将FAO文献数据库和FAO合作文献库的元数据合并成一个XML文件,然后将XML文件转换成RDFS来描述不同元数据对象之间的关系。系统使用KAON工具包作为本体构建和管理的工具,实现对本体的概念、关系、属性和词汇汇编的编辑功能。

2.2 国内农业本体研究现状

中国农业科学院农业信息研究所对农业本体做了一些前瞻性的研究,发表了"构建领域本体的方法"、"构建基于ontology的知识门户"、"叙词表与ontology的不同"等论文。中国科学院李景撰写了"本体理论在文献检索系统中的应用研究",系统比较了本体相关理论、本体主要技术方法,以花卉学为例进行了构建领域本体的探讨和基于领域本体的文献检索研究。

由中国农业科学院农业信息研究所承担的"农村知识本体的研究与知识库构建"子课题,主要开展农村信息元数据标准框架、农村知识本体的构建及本体存储技术、农村知识本体的底层互操作技术、农村知识库构建技术的农村信息智能搜索服务技术研究,为构建农业智能搜索工具提供信息分类、组织和智能推理基础。

由中国科学院地理科学与资源研究所承担的"农业知识语义检索关键技术研究"是国家高科技研究发展计划(863计划)数字农业技术专题中的组成部分。主要开展农业网络信息、科技文献信息和空间信息的融合技术、农业领域本体的构建技术、农业知识语义检索策略与方法、基于主题图的农业知识表达技术、基于本体语义的主题Robot技术、农业本体知识库的自学习和自维护技术等农业知识语义关键技术研究。

3 农业本体学习研究

农业本体属于领域本体的范畴,因此笔者依据领域本体的半自动或自动构建方法,尝试探究基于农业领域的本体学习。

3.1 农业本体学习的过程

首先从各种农业数据、语料库以及Web信息中选取数据源进行预处理。随之确定用于抽取领域相关实体的领域文集,采用基于语言学和统计的方法从其中抽取相关术语,过滤后,作为农业本体概念的候选。通过本体概念和实例学习,进行语义排歧,筛选出与领域相关的概念。然后运用基于符号、统计等方法获取概念之间的分类关系,构建分类体系,初步建立基础农业本体,使用基于模式匹配和关联规则等方法对非分类关系、公理及规则进行学习。最后对目标农业本体进行评估,利用反馈得到的结果,进行重新循环的学习。

3.2 农业本体学习的关键理论技术

3.2.1 农业领域概念的获取

领域概念的获取是领域本体构建的基础。在概念获取阶段不必考虑概念的具体表达形式以及概念间的关系。目前,领域概念获取有以下几个途径:

(1) 农业领域的专业词典;

(2) 农业领域的叙词表;

(3) N元语法抽取等无词典分词算法;

(4) 实体识别、短语切分技术。

3.2.2 农业领域概念的组织

领域概念的组织是本体构建的主要环节。主要是构建本体类层次以及为每个类别填充实例[14]。

(1) 概念关系的获取。

有关对自动分类和信息检索研究积累了大量的概念关系获取技术,主要有以下几种可为本体相关关系的获取提供支持:

a.基于统计的算法。

关于相关度度量的方法有多种,包括互信息、系数方法、Dice系数、Cosine系数、Jaccard系数、开方统计以及极大似然估计等。可以根据不同的语料和应用目的选取不同的计算方法。

b.隐含语义索引(LSI)。

LSI是基于概念空间的文档索引,作为对向量空间的一种改进。LSI利用统计计算导出概念特征词索引文档并进行信息检索,而不再是传统的索引词,从而消减了词和文档之间的语义模糊度,使得词与词之间、词与文档之间的语义关系更为明晰,消除了同义性和多义性所造成的影响,它通过奇异值分解和取近似矩阵来获得概念之间的相关关系。

c.关联规则算法。

设 $I=\{i_1, i_2, \cdots, i_m\}$ 是二进制文字的集合,其中的元素称为项(item)。记 D 为交易(transaction) T 的集合,这里交易 T 是项的集合,并且 $T \in I$ 。

一个关联规则是形如 $X => Y$ 的蕴涵式,这里 $X \in I$, $Y \in I$,并且 $X \cap Y = \varnothing$。规则 $X \Rightarrow Y$ 在交易数据库 D 中的支持度(support)是交易集中包含 X 和 Y 的交易数与所有交易数之比,记为 support($X \Rightarrow Y$),即 support($X \Rightarrow Y$) = $P(X \cup Y)$;规则 $X \Rightarrow Y$ 在交易集中的可信度(confidence)是指包含 X 和 Y 的交易数与包含 X 的交易数之比,记为 confidence($X \Rightarrow Y$),即 confidence($X \Rightarrow Y$) = $P(X \mid Y)$。

(2) 分类关系的获取。

a. 模式匹配。

基于模式匹配的方法是指通过分析领域相关文本,总结出一些频繁出现的语言模式作为规则,然后判断文本中词的序列是否匹配某个模式,如果匹配则可以识别出相应的关系[15]。采用此种方法,语料来源一般采用知识密集型的领域词典(知识词典)、分类表和叙词表。例如领域词典中的"…可分为……""由…组成"等模式;叙词表中的用、代、属分关系模式;分类表中的类名同义词和类目注释等。从上述语料中可以提取同义关系和等级关系。

b. 聚类算法。

聚类算法是本体构建中最常用的算法,这种方法对给定的数据集进行层次似的分解,直到某种条件满足为止。具体又可分为"自底向上"和"自顶向下"两种方案。例如在"自底向上"方案中,初始时每一个数据纪录都组成一个单独的组,在接下来的迭代中,它把那些相互邻近的组合并成一个组,直到所有的记录组成一个分组或者某个条件满足为止。代表算法有:BIRCH 算法、CURE 算法、CHAMELEON 算法等。

3.3 本体学习工具

目前已经有很多的本体学习系统被开发出来。Text2Onto[8]、Hasti[16]和 OntoLearn[17]是常见的三种代表性的本体学习工具,其基本原理都遵循如图 2 所示的框架和处理流程[18]:预处理模块首先对数据源进行预处理;接着,学习模块通过使用算法库中的各种本体学习算法从上一步预处理的结果中获取本体;然后将结果作为候选本体呈现给用户;最后用户在评价/编辑模块的帮助下对该候选结果进行评价和确认,并将最终的结果添加到本体库中。

3.4 本体学习结果评价

目前对本体学习结果的评价还没有统一的标准。常用的评价方法有:

(1) 基于应用的评价:通过选择一些相关的应用,根据这些具体应用的结果来评价本体学习的结果。

(2) 数据驱动的评价:通过与现有的相关领域数据相比较进行评价。例如确定所获本体是否关联特定主题;根据特定领域术语和出现在本体中的术语间的交叉度测定本体和语料库间的匹配度。

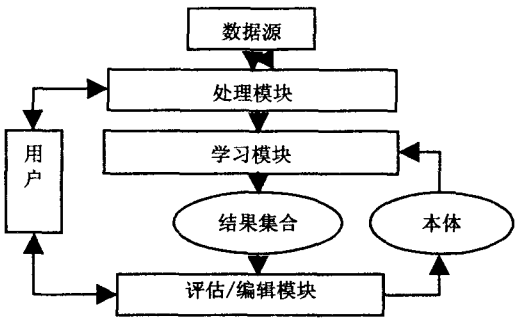

图 2　本体学习工具基本框架

(3) 计算学习模型的查全率和查准率:查全率是指正确概念的数量除以测试数据集中概念的总数,查准率是指正确概念的数量除以所提取概念的总数。

(4) 概念级比较:譬如,对于分类关系的比较,综合考虑被比较概念的父子概念的相似度;对于非分类关系的比较,可以考虑被比较关系的 domain 和 range 的相似度。

4 结束语

虽然本体学习的研究已成为当前的热点,但是目前对本体学习的研究仍有很多问题需要不断探索研究和解决:

(1) 概念间关系的学习多是针对分类关系的,对非分类关系的研究不够深入,大部分方法都停留在判断两个概念之间是否存在关系的层次上,无法进一步为获取的关系赋予相应的语义标签。

(2) 与国外相比,国内在领域概念的自动抽取方面,特别是对中文领域概念的自动抽取的研究较少。

(3) 目前还没有对本体学习结果进行定量评价的统一标准,很大程度上制约了本体学习方法和工具的进一步发展。

鉴于存在的这些问题,对基于深层次语义分析的本体学习方法进行研究;根据中文特点对基于中文的本体学习算法进行研究;开发适用于农业领域的中文本体学习工具以及对本体学习结果的评价等都将成为今后的研究方向。

参考文献:

[1] 李善平,尹奇韡,胡玉杰,等. 本体论研究综述[J]. 计算机研究与发展,2004,41(7):1041-1052.

[2] Farquhar A,Fikes R,Rice J. The Ontolingua server:A tool for collaborative ontology construction[J]. Int'l Journal of Human-Computer Studies,1997,46(6):707-727.

(下转第 219 页)

上和右),则该点的边长权重为$\sqrt{2}$,若交界点在三个方向接触背景区,则该点的边长权重为 $2\sqrt{2}$。这种计算方法在测量斜向的边界时较为准确,使用直径从 10 到 100 的圆形区域做测试,此种方法计算出的圆形度在 0.89～1.01 之间,误差较小。

创建与 P3 图像大小相同的二值图像 P4,用于存放被判定为裂纹的连通区域,P4 各点初值设置为 false,得到一个判定为裂纹的连通区域图像 R 时,P4 与 R 做逻辑或运算,结果返回 P4,即 P4 = P4 | R,将所有连通区域判断完成后即得到原图像中所有裂纹的图像,用于存储和记录。连通区域图像的提取与裂纹判定的流程图如图 3 所示。

文中的处理结果如图 1 中 P4 所示,完成了荧光磁粉探伤图像的裂纹提取及识别。

3　结束语

使用裂纹自动识别系统代替轴承荧光磁粉探伤中的人工观察具有实用价值,笔者对裂纹自动识别的图像处理方法进行了探讨和尝试,使用了一种简便的连通区域提取方法,改进了数字图像区域周长的计算方法。该算法在 Matlab 中完成测试,识别结果具有一定的准确度,证实了图像处理方法的可行性。

参考文献:

[1] 谭任芳.机车轴承检测流水线的应用[J].内燃机车,2006(12):41－43.

[2] 达正雄,朱元高,陆　玮.加强轴承检测,防止机车故障[J].铁道机车车辆,2001(1):35－44.

[3] 吴海滨,郑宏伟,李明琥,等.轮箍表面自动荧光磁粉探伤系统及其图像处理技术[J].无损检测,2007,29(3):128－131.

[4] 王恒迪,尚振东,马　伟.轴承套圈磁粉探伤机的研制[J].轴承,2005(3):32－33.

[5] 许万里,苑惠娟,郑　伟.工业视觉检查系统中的图像处理及模式识别[J].哈尔滨理工大学学报,2001,6(4):22－27.

[6] 苏彦华.Visual C++数字图像识别技术典型案例[M].北京:人民邮电出版社,2004.

[7] 王树文,闰成新,张天序,等.数学形态学在图像处理中的应用[J].计算机工程与应用,2004(32):89－92.

[8] 崔　屹.图像处理与分析——数学形态学方法及应用[M].北京:科学出版社,2000.

[9] 张艳玲,刘桂雄,曹　东,等.数学形态学的基本算法及在图像预处理中应用[J].科学技术与工程,2007(2):356－359.

[10] 高　潮,任　可,郭永彩,等.基于 DSP 和图像识别的拉索表面缺陷检测技术[J].重庆大学学报,2007(9):36－38.

(上接第 215 页)

[3] Duineveld A J. Wonder tools:A comparative study of ontological engineering tools[J]. Int'l Journal of Human－Computer-Studies, 2000,49(6):1111－1133.

[4] Noy N F,Fergerson R W,Musen M A. The knowledge model of protégé－2000: Combining interoperability and flexibility [C]// In Proc of the EKAW. [s.l.]:[s.n.],2000:17－32.

[5] Bozsak E. KAON－Towards a large scale semantic web[C]// In Proc. of the 3rd Int'l Conf. on I－Commerce and Web Technologies. Heidelberg:Springer－Verlag,2002:304－313.

[6] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition,1993,5(2):199－220.

[7] Maedche A. Ontology Learning for the Seman－ tic Web[M]. Boston:Kluwer Academic Publishers, 2002.

[8] Buitelaar P. Ontology Learning from Text[M]. Porto,Portugal:[s.n.],2005.

[9] 刘柏嵩,高　济. 基于 Web 的通用本体学习研究[D]. 杭州:浙江大学,2007.

[10] Murata M,Lee D,Mani M. Taxonomy of XML Schema Languages using Formal Language Theory[M]// In Extreme Markup Languages. Montreal,Canada:[s.n.],2001.

[11] Compton P, Jansen A. A philosophical basis for knowledge acquisition[J]. Knowledge acquisition, 1990,2(6):241－257.

[12] 钱　平,郑业鲁. 农业本体论研究与应用[M]. 北京:中国农业科学技术出版社,2006.

[13] 谢能付,王文生. 农业知识本体构建方法[J]. 农业网络信息,2007(8):13－14.

[14] 何　琳,杜慧平,侯汉清. 领域本体的半自动构建方法研究[J]. 图书馆理论与实践,2007(5):26－27.

[15] 贾秀玲,文敦伟. 一种本体学习中分类关系提取方法的研究[J]. 计算机技术与发展,2007,17(10):33－34.

[16] Johannusson P. A method for transforming relation schemas into conceptual shemas[C]// In: Rusinkiewics M, ed. Proceedings of the Tenth International Conference on Data Engineering. Houston:IEEE Press,1994:115－122.

[17] Velardi P, Fabriani P, Missikoff M. Using text Processing techniques to auto－ matically enrich a domain ontology[C]// In:Proceedings of the International Conference on Formal Ontology in Information System, Ogunquit,2001. New York: ACM Press,2001:270－284.

[18] 孔　敬.本体学习:原理、方法与相关进展[J]. 情报学报,2006,25(6):657－665.