·博士论坛·

面向知识网格的本体学习研究

刘柏嵩[1,2]　高　济[1]
[1](浙江大学计算机学院,杭州 310027)
[2](宁波大学网络中心,宁波 315211)
E-mail:lbs@nbu.edu.cn

摘　要　网格计算正在从单纯的面向大型计算的分布式资源共享发展为一种面向服务的架构,以实现透明而可靠的分布式系统集成。网格智能是指如何获取、预处理、表示和集成不同层次的网格服务(如 HTML/XML/RDF/OWL 文档、服务响应时间和服务质量等)的数据和信息,并最终转换为有用的智能(知识)。因为高层知识将在未来的网格应用起到越来越重要的作用,本体是知识网格实现的关键。文章提出了一种实现从 Web 文档中本体(半)自动构建的本体学习框架 WebOntLearn,并讨论了本体学习中领域概念的抽取、概念之间关系的抽取和分类体系的自动构建等关键技术。

关键词　知识网格　本体　本体学习　Web 挖掘

文章编号 1002-8331-(2005)20-0001-05　**文献标识码** A　**中图分类号** TP18

A Study on Ontology Learning for the Knowledge Grid

Liu Baisong[1,2]　Gao Ji[1]
[1](Institute of Computer Science,Zhejiang University,Hangzhou 310027)
[2](Network Center,Ningbo University,Ningbo 315211)

Abstract: Grid Computing is evolving from the merely sharing of distributed resources for large computational tasks to the developing of Grid as a service-oriented architecture for transparent and reliable distributed system integration.The paradigm of Grid computing complements the current approach of Semantic Web and Web Services by providing an infrastructure to handle large-scale distributed enterprise information systems.Grid intelligence refers to an emerging research field that addresses on how the data and information available at different levels of Grid services(e.g.,HTML/XML/RDF/OWL etc)can be effectively acquired,preprocessed,represented,interchanged,integrated and eventually converted into useful intelligences(knowledge).As higher-level knowledge is going to play a more important role in the future Grid applications,ontology is the key of the Knowledge Grid.This paper first proposes a framework of ontology learning from web pages.Then key technologies of ontology learning such as domain concepts extraction and semantic relationships between concepts and taxonomy automatic construction are discussed.

Keywords: knowledge grid,ontology,ontology learning,Web mining

1　引言

目前,网格计算的现状和其远景目标之间有一条鸿沟。计算网格(Computing Grid)是指对网络上的各种结点设备计算和处理能力的共享,信息网格(Information Grid)是指对网络信息的共享,而知识网格(Knowledge Grid)则强调对网络知识的共享。知识网格可以说是对现有网格的扩展,其信息和服务可以良好地定义,使得计算机和人更好地协同工作。高层知识在未来的网格应用将起到越来越重要的作用,网格智能(Grid Intelligence)是知识网格实现的关键,它通过对各种网格服务(如服务响应时间和服务质量等)的信息获取、预处理和集成,并最终转换为有用的智能(知识)。在理想的知识网格条件下,用户完全不必了解所需的知识到底是在哪一台计算机上,也不必知道究竟是哪一台计算机在为自己服务—对于用户而言,这些知识是全透明的,知识象水一样流到使用者面前。可以说,未来整个网格就是一个庞大而有序的知识管理系统。本体机制是科学家为客观地解释对象的语义及它们之间的关系而建立的,反映了人们对语义的共识。本体(Ontology)将是实现这一目标的前提。

目前,虽然具有一定应用需求的推动,知识网格距离实际应用尚有一段差距。为实现知识网格需构建大量的本体(尤其是应用本体)来满足其需求。但本体和知识库从何而来?相对于因特网上海量信息而言,目前只有很少手工构建的本体如

基金项目:国家 973 重点基础研究发展规划项目(编号:2003CB317000);浙江省自然科学基金(编号:M603010);宁波市青年博士基金(编号:2003A62002)资助

作者简介:刘柏嵩(1971-),男,博士生,研究方向为人工智能、语义网、本体工程。高济(1946-),男,教授,博士生导师,主要从事网络计算与普适计算、智能软件与 Agent 技术等领域的研究工作。

WordNet 和 Cyc。但是一方面用手工方式构建本体需要耗费大量的人力和时间,另一方面这些通用本体只包含非常少的领域概念。同时,如何维护现有本体,尤其是如何保持更新(Update)?因为新的概念、新的实例和已有概念的属性在不断引入。譬如,某一医学本体中可能并不包含 SARS,更不可能包括其与香港和北京的关联。几年前手机并没有因特网无线访问这一特征,但如今这是重要的特征之一。为了解决本体工程(Ontology Engineering)中"知识瓶颈"问题,我们需要自动化或半自动化工具来构建本体。

在知识网格初期,要构建大量的领域本体(Domain Ontology)以满足知识网格的需要,大量的 Ontology 主要是通过对网络上各专业领域中大量的 HTML 网页进行抽象分析得到。因此,需要一种简单可靠的 Ontology 的提取方法,即一种高效快捷的 Ontology 构造方式。本文首先提出了一种实现本体(半)自动构建的本体学习框架 WebOntLearn,并分别讨论了本体学习中领域概念的抽取、概念之间关系的抽取和分类体系的自动构建等关键技术,最后回顾了相关研究工作。

2 本体学习的概念及目标

本体建造是一个非常复杂的过程,它需要多个领域的专家参与。虽然目前本体工程(Ontology Engineering)工具已经较为成熟,但本体的手工构造仍是一项繁琐而辛苦的任务,并最终导致所谓的知识获取瓶颈。而且,本体具有任务相关(task-dependent)和静态性(static)的特征。

从目前本体工程的实践来看,本体的构建和维护主要存在如下问题:第一,在构建的初期和维护阶段需要花费大量的人力,包括构建实际的分类体系(Taxonomy),以及将某一特定内容与分类体系中的节点关联起来。例如,在 Yahoo 或 DMOZ 开放目录中包括分层目录和与某一目录相关的站点。第二,本体中俘获的知识是流变的(Evolution),它总是在不断地发展和更新。为避免本体成为过期的无用信息,这就意味着本体不能象字典一样以手工方式构造,否则它的发布之日就已过时。第三,本体中的分类体系具有领域相关性,特定学术或商务专业领域有其自身的词汇表和技术术语,因此构造合适的通用本体或分类体系需要大量的修剪和编辑时间。第四,本体反映了客观世界的某一特定观点,它反映了构建者个人或机构的观点。第五,本体作为一种共享概念模型,但通常很难以某种特定的方式来对客观世界分类。

本体学习(Ontology Learning)技术可以说当前的一个热点[4]。其目的旨在开发能够实现本体自动构建的机器学习技术来协助知识工程师来构建本体,基本原理如图 1 所示。本体学习任务主要包括:(1)本体获取:包括本体创建、本体模式(schema)抽取和本体实例(instances)抽取。(2)本体维护(Ontology maintenance):包括本体集成和导航、本体更新以及本体扩充(enrichment)。

本研究的主要目标是:从 Web 文档中自动获取领域术语及其相互关系;采用信息抽取(IE)技术来确定概念对之间的语义关系,在获取的概念及其相互关系的基础上,构建本体。经过系统所获取的 Web 本体目标不仅仅局限于逻辑学的学术范畴。语义描述要能够为计算机方便利用,因此,并不追求语义的完整和深入,只求语义表达的可扩展性。它的任务是把共同约定、共同享用的知识(词语的语义规范),用计算机容易处理的形式表达出来。

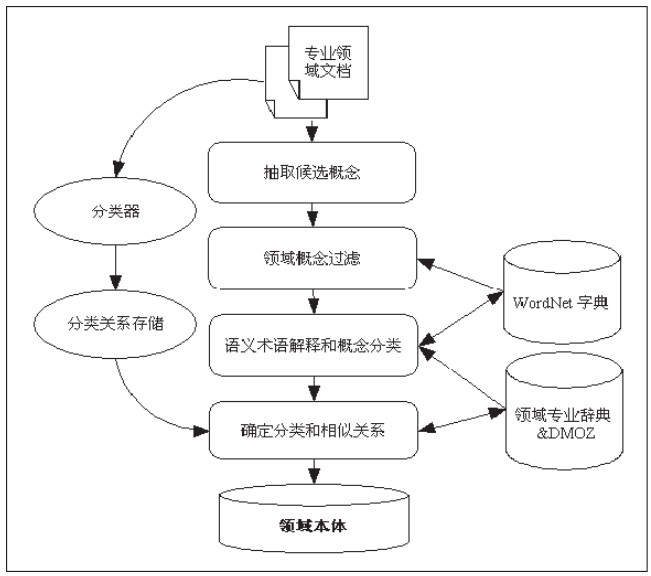

图 1 本体学习的基本原理

3 本体学习的一种实现

3.1 基本框架

本文旨在实现从 Web 页面中自动抽取本体(Web based Ontology Learning,简称 WebOntLearn),从 Web 页面数据中找出本体语义概念的模式及其关系。它通过分析同一应用领域 Web 页面集来半自动化地抽取 Web 本体。其基本步骤包括:

(1)语料库(corpus)和 Web 文档集的收集、选择和预处理;

(2)生成候选关键词集;

(3)抽取领域术语(term);

(4)本体概念选择;

(5)确定语义关系并构建分类层次体系;

(6)创建形式化表示。

3.2 文档的预处理

为了简单起见,本文只处理正文信息。输入一个 Web 页,系统扫描该 Web 页面,进行文档的特征抽取、进行句子锚定,并找出 HTML 文本中可能分类为本体实例的词组或短语(phrase)。

步骤 1 文档的特征抽取:一个 Web 页面中包含了图像、动画、音频、超链接等丰富的信息表达方式,但最主要的信息还是正文的文字信息。

根据 Bing Liu[5]等提出方法,我们并没有应用自然语言处理 NLP 技术,采用如下步骤:

(1)过滤掉那些不包含子主题或关键概念的"噪音"(noisy)文档,如论坛讨论页面以及根本不包含相关术语项的页面。过滤启发式规则是基于在这些"噪音"文档经常出现的线索短语。

(2)对每一页面中的重要短语作出标记。这些标签包括:$<h_1>,\cdots,<h_4><b><big><i>$等。另外,还确定一些规则让某一标记文本能被安全忽略:如包含礼节性称谓(如某某先生、博士或教授)、包含一个 URL 或电邮地址等。

(3)对于 Web 链接,假设网页 p 中有链接指向网页 q,记为

$p->q$，对于主题 t，网页 p，q 对此主题的原始权值分别为 $p(t)$ 和 $q(t)$。设 $w(i,t)$为网页 i 关于 t 主题的权值，则 $w'(p,t)=w(p,t)+f*w(q,t)$，$w'(q,t)=w(q,t)+f*w(p,t)$。其中，$f(0<f<1)$为消退因子，其表示相链接的网页对于此网页属于某主题的影响能力。式中，前一项表示基于网页文档内容分析而获得的主题权值，而后一项可以看作为基于 Web 网络结构信息分析而产生对其主题权值的影响。

(4)采用上述规则，分析所有页面并抽取符合要求的文本段。然后移除停用词并 word stemming。

步骤 2 句子锚定：句子锚定过程选择那此可能包含本体关系的句子(S)。如果 S 中包含了是 C 的成员的提示词，则某一句子 S 被锚定。更准确地说，某一锚定句子 S 可以重写为字符串 $T_{L_{1\cdots m}}.c.T_{R_{1\cdots n}}$，其中 $T_{L_{1\cdots m}}$和 $T_{R_{1\cdots n}}$为名词短语或术语序列。

步骤 3 每一候选规范词的查询结果进行汇总。重复每一候选规范词，将所有候选规范词和候选本体概念导入语言模式以提取假设短语。例如，候选规范词“朝鲜”和概念“国家及公司”合成为一种模式，如假设短语为“朝鲜是一个国家”和“朝鲜是一个公司”。该步骤的结果为假设短语集。然后，Google 通过其 Web 服务 API 查询假设短语。该 API 程序返回每一假设短语的命中数。

经过文档预处理步骤，已经产生一系列的候选规范词。

3.3 术语项生成

术语是专业领域中概念的语言指称。通过提高准确率和召回率，由计算机尽可能准确、全面地抽取候选术语项，是本体学习的关键。本步骤的目标是从上一步的候选规范词中抽取领域术语(Term)。术语表示为某一指定领域内简单或复杂含义的词组或字符串。从某种意义上讲，术语是一种领域知识的文本形式的浅层表示。因其低二义性和高专指性，这些词对于领域知识的概念化尤其有效，可支持领域本体的创建。

步骤 1 候选术语生成：首先采用词组块(Phrase chunking)来确定句子中浅层短语边界。在该过程中，本文采用浅层解析技术以及启发信息，如表示重点句子和段落的提示词。浅层解析器模块可分为两个过程：句子锚定，候选术语生成和本体术语选择。所有锚定句子被分块以形成名词短语、动词短语和从句。该步骤的输出是一组没有结构消歧的候选名词短语。

步骤 2 相关度计算：Roberto Navigli 提出了一种新型的方法筛选“真正”的术语[8]，该方法基于称作领域相关性和领域一致性的两种测度形式。类 D_k 中术语 t 的领域相关性采用如下公式计算：

$$DR_{t,k}=\frac{\mathrm{P}(t|D_k)}{\max\limits_{1\le j\le n} P(t|D_j)} \tag{1}$$

其中条件概率 $P(t|D_j)$采用下式估计：

$$E(P(t|D_k))=\frac{f_{t,k}}{\sum\limits_{i'\in D_k} f_{t',k}}$$

步骤 3 应用互信息(Mutual Information)方法来抽取共生(co-occurrence)词：对上一步中产生的结果中错误名词短语进行修剪(pruning)。在该步骤中，通过应用句法结构和统计技术来分析名词短语，解决名词短语生成过度或不及的问题。从句法标注的 corpus 中，创建了相同名词短语的概率模型，它通过从文档中抽取信息并采用下式计算：

$$P_{NP_i}(w_i,w_j)=P_f(w_i)*P_b(w_j)$$

其中 $P_{NP_i}(w_i,w_j)$为名词短语或复合名词，w_i 和 w_j 可关联到一个新词；$P_f(w_i)$为 i 在名词短语跟在其它后面中的出现频度/i 在所在文档出现的频度；$P_b(w_j)$为 i 在名词短语在其它前面中的出现频度/i 在所在文档出现的频度。

这种概率模型可用于修剪候选名词短语中的错误名词短语。如果前面的名词短语的概率大于阈值，该名词短语则可能为一个合适名称(proper name)。

步骤 4 确定术语项的词序：对选择的术语集根据相关度进行排序，形成术语项列表。

3.4 本体概念选择

概念是知识的基本单位也是思维的最小单位。术语和概念之间应一一对应，即一个术语只表示一个概念(单义性)；一个概念只有一个指称，即只由一个术语来表示(单名性)。在相关学科或至少在一个专业领域内应做到这一点，否则会出现异义、多义和同义现象。在通过浅层句法分析抽取术语项后，可应用多种方法来抽取本体概念。本体中包括叶子概念(leaf concept)和非叶子概念(non-leaf concept)两种，其中叶子概念是指没有子概念的结点。

术语要成为本体概念，必须同时满足两个条件：(1)有明确的含义(specific)；(2)有重要的作用(significant)。判断术语是否有明确的含义，主要是考察其稳定性与完整性。我们说一个字符串是稳定的，是指它包含的各个字符紧密地结合在一起，经常固定地在一起出现。例如，“比尔盖茨”(Bill Gates)是一个概念，它是稳定的，这几个字符往往固定地在一起出现。

根据香农的信息论，术语(字符串)的稳定性可以通过其内部的互信息(mutual information)来度量，并选择互信息值最高的作为候选概念。

定义 1 设文档 T 的一个字符串 S 由 P 个字符组成($P\ge 2$)即“$c_1c_2\mathrm{L}\ c_p$”，则 S 的互信息为：

$$MI(S)=\frac{f(S)}{f(S_L)+f(S_R)-f(S)} \tag{2}$$

其中 S_L 是将 S 去掉最右边一个字符而得到的左段子字符串，即“$c_1c_2\mathrm{L}\ c_{p-1}$”，$S_R$ 是将 S 去掉最左边一个字符而得到的右段子字符串，即“$c_1c_2\ \mathrm{L}\ c_p$”，$f(S)$，$f(S_L)$，$f(S_R)$是字符串 S，S_L，S_R 各自的出现频率。

上述字符串的互信息定义是由一般的两个随机变量之间的互信息定义简化而来。若以 $X(S)$表示字符串 S 的所有出现的集合，以 $X(S_L)$和 $X(S_R)$分别表示字符串和的所有出现的集合。

如果一个字符串的互信息高于某个阈值，那么就可以认为这个字符串是稳定的。例如，字符串“计算机”的左段“计算”与右段“算机”之间联系比较紧密，其互信息的值比较高，因此字符串“计算机”是稳定的，字符串“计算机软件”也是稳定的。

我们说一个字符串是完整的，是指它能够独立地表达完整的含义，因此可以独立地出现在不同的上下文之中。是短语的字符串都应该是完整的。例如，“数据挖掘”是一个短语，它是完整的，可以在独立地出现在不同的上下文之中(象“教授数据挖掘课程”、“购买数据挖掘软件”等)；而“数据挖”只是短语“数据

挖掘”的一个部分，它虽是稳定的却又是残缺的(partial)，它只有后面加上“掘”字才能表达完整的含义，故它只有后面加上“掘”字才会出现在文档之中。

3.5 确定语义关系

在 WebOntLearn 中，考虑两类关系：一类是处于不同逻辑层次上的概念之间的关系，包括种属关系(IS-A relation)和实例关系(Instance-of Relation)；另一类是反映对象组成结构的关系，它是部分和整体之间的关系(Part-whole relation)。然后，根据这些关系构造不同的本体分类体系。本文主要考虑本体的一些基本关系，包括 Instantiation、Membership、Parthood、Connection、Location、Extension、Dependence 关系。以下为两类元关系(primitive relations)即实例关系和部分关系的定义：

定义 2 概念实例：设概念集为 S_C，对于概念集 S_C 中的任意概念 C，概念 C 的外延集 $E(C)=\{x|x\subset C\}$；对于 $E(C)$ 中的任一元素 $C_i\in E(C)$，如果 C_i 的外延集 $E(C_i)=\{C_i\}$，则称 C_i 为概念 C 的实例。

定义 3 (Instance-of)对于概念(或类)C 及其实例集 S_{ic}，则实例集 S_{ic} 中的元素 $e(e\in S_{ic})$和概念 C 之间的关系称为实例关系。记作 Inst(e,C)，可表达为 e is an instance of C。该关系存在于实例(或个体实例)和概念之间。

实例关系没有自反性、对称性和传递性。但是从概念的内涵式可知，实例和概念之间具有很好的性质和属性的继承性。

定义 4 (parthood)部分关系是指对象个体之间的部分关系。将 x 是 y 的一部分记作 Part(x,y)。部分关系没有自反性和对称性，但具有传递性。

如 WordNet 一样，在此我们考虑三种 part-whole 关系：(1) component-composite 关系，如 Wheel->Car；(2)Member-collection 关系，如 tree->forest；(3)Stuff-object 关系，如 aluminum->car。

定义 5 (part)A part-for $B=_{def}\forall x(inst(x,A)\rightarrow\exists y(inst(y,B)\&\ part(x,y))$。

定义 6 B has_part $A=_{def}\forall y(inst(x,B)\rightarrow\exists x(inst(x,A)\&\ part(x,y))$。

由定义 4 和 5，可以得出：

定义 7 (part_of)A 与 B 是部分关系，当且仅当：对 A 的任一实例 x，存在 B 的某些实例 y 在实例级与 x 为部分关系；反之亦然。即定义为 A part_of $B=_{def}$ A part_for B & B has_part A。Part_of 关系不具自反性和对称性，但具传递性。

定义 8 (Is-a)对于概念集 S_C 中的概念 $C_1,C_2\in S_C$，如果有：(1)概念 C_1 的内涵包含 C_2 的内涵，即 $I(C_1)\supset I(C_2)$；(2)概念 C_1 的外延包含于 C_2 的外延，即 $I(C_1)\subset I(C_2)$。该关系存在于种概念和类概念之间。A Is-a $B=_{def}\forall x(inst(x,A)\rightarrow inst(x,B))$，其中→为 if …then 的缩写。则将概念和 C_2 之间的关系称为种属关系(Is-a relation)。记作 Is-a(C_1,C_2)。

种属关系不满足对称性，但有自反性、反对称性和传递性，因此它为偏序关系。

在参考了文献[7,8]的基础上，借助 WordNet 和 HowNet 这两个通用本体库，我们采取监督学习和无监督学习相结合的方法获取概念之间上述语义关系。其基本思想是：根据人工总结的关系抽取模板，抽取出所有可能的关系对，并记录关系在检索集中出现的次数，组成候选关系集。每一种关系对应一组抽取模板。然后计算语义关系的支持度和置信度，并应用集合运算对候选关系集进行优化，逐一删除错误关系。最后进行角色转换，获得各种类型的语义关系，最终建立语义关系库。

3.6 分类体系的构建

根据概念(或类)之间的语义关系，可以构建出概念分类层次(Taxonomy)关系。如图 2 所示。

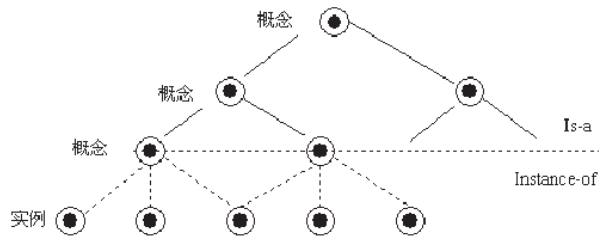

图 2 概念分类层次关系

定义 9 (概念层次)一个概念层次 H 是一个偏序集(c,p)，其中 c 是一个有限的概念集，p 是 c 上的一个偏序。

定义 10 (父结点)在概念层次 H，$H=(c,p)$中，如果 $x,y\in c$，$x\ p\ y$ 且不存在概念$(z\neq x,y)$，满足 $x\ p\ z$ 且 $z\ p\ y$，则概念 y 称为概念 x 的父结点。

$root(A)=_{def}\forall B(Bis\_aA)$

定义 11 (规则概念层次)若概念层次 $H=(c,p)$满足以下两点，则称 H 是规则的：

(1)c 存在一最大概念 y，对任意的 $x\in c$，有 $x\ p\ y$；

(2)存在概念集 $c_i(i=0,1,L,n-1)$，使得 $c=\bigcup_{i=0}^{n-1}c_i$ 且 $c_i\ \mathrm{I}\ c_j=\emptyset(i\neq j)$；

(3)若 c_i 中一个概念的父结点在 c_j 中，则 h_i 中所有概念的父结点都在 c_j 中$(i\neq j)$。

我们可以定义分类层次的根概念和结点概念如下：

定义 12 $leaf(A)=_{def}\forall B(Bis\_aA\rightarrow A=B)$。

定义 13 $top(A)=_{def}\forall B(A=B\ \text{ or }\ Bpart\_of\ A)\&not-\exists B(Apart\text{-}ofB)$。

定义 14 $bottom(A)=_{def}\ not-\exists B(B\ part\_of\ A)$。

因此，对于包含叶子结点的类(或概念)而言：$\forall A\ \exists B(leaf(B)\&\&Bis\_a\ A)$

对于每一包含实例的类而言，则有：

$\forall A\ \forall x(inst(x,A)\rightarrow\exists B(leaf(B)\&inst(x,B)))$

分类体系 T 的构建遵循自顶向下的抽象方式，将对象 D 分成两个大类，类型 C 和实体 I，类型又分为两种：虚拟类 VC 和实体类 IC，虚拟类不能实例化，只能参与类之间的抽象演算，如包含、等价与无关性关系；实体类不但能抽象演算，而且具有实例 Instance-of()，即实体 IO。虚拟类与实体类能够互相包含，我们定义被另一个实体类包含的实体类的实例是第一个实体类的间接实例，记作 Instance-of()。

类演算关系有包含、等价和无关性三种：

$Inclusion(c,d)\Leftrightarrow\{c,d\in C|c\subseteq d\}$

$Equivalent(c,d)\Leftrightarrow\{c,d\in C|c\equiv d\}$

$Disjoint(c,d)\Leftrightarrow\{c,d\in C|(c\subseteq\neg\ d)\wedge(d\subseteq\neg\ c)$

本步骤是自顶向下地进行。首先是获取特定领域内抽象程

度较高、也就是所指的涵义比较宽的那些概念，比如：计算机领域中的“软件”、“硬件”、“网络”等，这些词可以作为该领域内不同分支的主题词，而“办公软件”、“显示器”等词的概念就比较小，抽象程度就相对低，它们作为再下一级的分支。当获取了位上层的词之后，就开始寻找分别位于这些词下方的那些较为具体的词（概念）。不断重复就可不断构造出下层的分支，形成一个越来越大的网络。在领域概念树（Domain Concept Tree）的基础上，最终形成概念的分类体系。

4 相关研究

在本体建立方面，目前存在的绝大多数本体都是手工生成的，该方法费时费力还容易出错，更难以维护和更新[1,4]。由于网格上的信息量巨大、主题繁多，研究如何自动化、半自动化生成本体具有重大的意义。为此，研究者提出了本体学习这一涉及人工智能中信息获取、机器学习、自然语言处理等多领域交叉的研究课题。Maedche 等首先正式提出了本体学习的概念，并给出一个半自动化的需人工干预的 Ontology 学习框架[9]，采用平衡的协作建模方式来构造语义 Web 中的本体，这个框架用半自动化的本体构造工具对典型的本体工程环境进行扩展，在这个框架中本体的建模周期由 5 个步骤组成：Ontology 引用、抽取、剪枝、精炼和评估，这个框架将能够为本体工程师提供丰富的本体协作建模工具。相关的研究项目主要有：Text-To-Onto、OntoLT、OntoLearn、ECAI2000、Inductive logic programming、Library Science and Ontology 等等[4,6,8,11]，这些项目对本体构建中的不同方面均有所研究，包括概念抽取和关系抽取的方法、本体的重用、本体的表示等，形成了一些技术成果，例如 POS、word sense disambiguation、tokenize、pattern matching 等等[2,4,6,8]；还实现了一些实用工具，如 OntoEdit、AeroDAML 和 OilEd 等[4,6,8]。其中 Text-To-Onto、OntoLT、OntoLearn 是面向通用本体的学习工具。这些研究的共同之处是：处理的源数据或多或少都是半结构化的，并且由领域专家提供种子词汇。由于它们针对非结构的源数据以及种子词汇的自动获取缺乏支持，所以在分布式网络信息处理上尚未获得成功应用。在本体建造方面，仍然有不少问题需要解决，如从纯文本和异类数据源中学习本体还停留在实验室阶段，实际应用仍然存在很多困难[2,3,4]；关系抽取也是一个非常复杂和难以解决的问题，已经成为本体学习和应用的主要障碍。因此，在本体体现其在信息组织、管理和理解方面的优越性之前，还有大量的工作需要做。本研究采用 Text-to-Onto 工具包实现本体概念选择（见 3.4 节）和概念的语义关系学习（见 3.5 节），这也是目前唯一可从互联网上获取的工具包。

5 结论

知识网格的目标是提供比现有 Web 信息服务更好的、智能的、高性能的协同问题解决支持平台。在网格环境中，其完全的分布式特性和高度的自治性为在知识网格中获取和使用 Ontology 带来了很大的挑战。本文提出了一种从 Web 文档中（半）自动化构建本体的本体学习方法 WebOntLearn，该方法还处于探索阶段，将来还有大量的工作要做。由于篇幅的限制，本文没有讨论学习所获本体的编码（Coding），即以什么形式来表示结果本体（如 RDF 或 OWL）。由于本体表示了某一领域的共享概念模型，因此本体的构建要求对人类语言的深层次理解。现有的方法主要依靠浅层语言处理，因而很难发现其深层次的关系。另外，由于人类语言的歧义性和数据的稀疏性，从海量文档集中进行领域概念识别以及抽取概念间的关系仍存在困难。本文只考虑了本体的层次分类结构（Hierarchy），在今后的工作中还需考虑概念之间的其它语义关系，以及本体实例的扩充。

本体学习的研究与发展必将从根本上改变网格环境下知识系统的构建方式，对计算机网格和语义网的向前发展并最终普及应用起很大的推动作用。本文的研究仅仅是一个起点，关于本体的自动构建，后续要进行的工作还有很多，主要包括本体集成、本体映射和本体评价等。（收稿日期：2005 年 4 月）

参考文献

1.Maria Teresa Pazienza et al.Modelling semantic grid knowledge embedded in documents[C].In:Proceedings of the 13th IEEE International Workshops on enabling technologies:Infrastructure for collaborative enterprises(WET ICE'04),2004

2.Philipp Cimiano et al.Towards the self-annotating web[C].In:WWW2004,New York,USA,2004-05:17~22

3.Zheng Chen et al.Building a web thesaurus from web link structure[C].In:SIGIR'03,Tornoto,Canada,2003-07

4.A Gomez-Perez,Manzano-Macho.A survey of ontology learning methods and techniques.OntoWeb Deliverable D1.5,2003

5.Bing Liu et al.Mining topic-specific concepts and definitions on the web[C].In:WWW2003,Budapest,Hungary,2003-05

6.P Buitelaar et al.A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis[C].In:ESWS,2004

7.Ana G Maguitman,Filippo Menczer.Algorithm detection of semantic similarity[C].In:WWW2005,Chiba,Japan,2005-05

8.Roberto Navigli,Paola Velardi.Learning domain ontologies from document warehouses and dedicated web site[M].Computational Linguistics(30-2),MIT Press,2004-06

9.Maedche A,Staab S.Ontology learning for the semantic web[J].IEEE Intelligent Systems,2001;16(2)

10.Marta Sabou.Extracting ontologies from software documentation:a semi-automatic method and its evalution[C].In:ECAI 2004,2004

11.Thanh Tho Quan et al.Automatic generation of ontology for scholarly semantic web[C].In:ISWC2004,LNCS 3298,2004:726~740