

# 中文本体构建及可视化研究

王晓盈<sup>1</sup>, 王晓璇<sup>2</sup>, 刘 鹏<sup>1</sup>

(1. 解放军理工大学 指挥自动化学院 网格技术研究中心, 江苏 南京 210007;

2. 同济大学 测量与国土信息工程系, 上海 200092)

**摘 要:**本体是解决语义层次上 Web 信息共享和交换的基础,随着本体研究的发展,中文本体应用需求不断扩大。对于本体工程的第一步——本体构建,现已出现很多本体构建工具,如:Protégé-2000、WebODE、OilEd、OntoEdit 以及 KAON 等。然而,却缺乏一款功能完整的中文本体构建工具。文中对现有本体构建工具进行了比较分析,并对 Protégé 的中文本体构建能力进行评估,深入探讨了其构建中文本体过程中遇到的可视化问题,提出了可行性解决方案。全面讨论了中文本体构建工具现存的问题,指出了未来的研究方向。

**关键词:**中文本体;本体构建;本体可视化

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2010)02-0121-04

## Research on Chinese Ontology Construction and Visualization

WANG Xiao-ying<sup>1</sup>, WANG Xiao-xuan<sup>2</sup>, LIU Peng<sup>1</sup>

(1. Research Center for Military Grid Technology, Institute of Command Automation, PLA

University of Science & Technology, Nanjing 210007, China;

2. Department of Surveying and Geoinformatics, Tongji University, Shanghai 200092, China)

**Abstract:** Ontology plays a key role in the web information share and exchange at semantic level. With the development of ontology, the need of application of Chinese ontology is increasing. Talking about the first step of ontology engineering - ontology construction, now many web ontology editing prototypes, tools or environments appear, for instance: Protégé - 2000, WebODE, OilEd, OntoEdit, KAON, etc. However, there is no an effective tool with integrated functions. This paper introduces and compares existing ontology construction tools, and evaluates the capability of Protégé on Chinese ontology construction. The problems of Chinese ontology visualization met during the process of Chinese ontology construction are investigated deeply, and available solutions are given. Based on that, it discusses the present problems and future research directions in this field.

**Key words:** Chinese ontology; ontology construction; ontology visualization

## 0 引言

本体(Ontology)最著名的定义是由 Gruber 提出的“本体是概念模型的明确的规范说明”<sup>[1]</sup>。本体的重要性已在许多方面表现出来并得到广泛认同。目前,本体已经应用于语义 Web、智能信息检索、信息集成、数字图书馆等领域<sup>[2]</sup>。本体构建属于本体工程的一部分,由于知识具有无限性、领域性,人对知识的认知具有主观性,使得本体的构建需要大量的领域背景知识,建模过程复杂,是一个比较庞大的系统工程。好的工具可以帮助快速而有效地建立本体,从而使语义 Web

具有广泛应用的基础。

随着本体研究的发展,本体构建工具也层出不穷,完整的本体构建工具能够完成对本体的创建、解析、存储和重用等工作。目前已有许多本体构建工具,功能也在日趋完善,然而,却没有一款支持中文本体构建的工具。构建环境可读性差、功能不完善已成为中文本体研究进一步开展的瓶颈所在。当前本体构建工具在对中文进行可视化、推理、解析等方面存在诸多问题。例如,可视化是反映本体的直观而简洁的方式,可以从图示中清晰地看到概念的层次结构、属性关系、实例等内容,并可以直观、简便地进行正确性检查,现有的本体构建工具无法提供对中文本体的可视化功能,在处理中文问题时会出现乱码或无法显示等现象,更无法与推理功能相结合,为用户提供直观的推理结果。中文本体构建环境的不完善为中文本体的应用带来极大

收稿日期:2009-06-07;修回日期:2009-09-10

作者简介:王晓盈(1984-),女,辽宁沈阳人,硕士研究生,研究方向为语义网;刘 鹏,博士,教授,研究方向为分布/并行体系结构、语义网络。

的不便,亟需一款能够支持中文本体的构建工具。

文中将对中文本体在构建和可视化过程中遇到的问题进行深入的分析 and 探讨,笔者就本体构建环境可读性差、中文本体无法图形化显示等问题,开发了支持中文本体构建的工具并就其中的相关技术问题进行了深入研究。

## 1 中文本体构建环境

### 1.1 本体构建工具概述

当前国外许多大学和研究机构正在研究与开发的本体构建工具很多,到目前为止,本体构建工具的总数超过了 90 个<sup>[3]</sup>。在过去的 10 年里,已经出现了许多本体构建工具,从最早的 Ontolingua<sup>[4]</sup>、OntoSaurus<sup>[5]</sup>、WebOnto<sup>[6]</sup>到 Protégé-2000<sup>[7]</sup>、WebODE<sup>[8]</sup>、OilEd<sup>[9]</sup>、OntoEdit<sup>[10]</sup>以及 KAON<sup>[11]</sup>等,本体构建工具也日趋成熟。这些工具总的来说具有以下优点:

①提供了较为友好的图形化界面和一致性检查机制;

②这些工具独立于语言,即用户不必了解本体描述语言的细节,只需把精力集中在本体内容的组织上,避免了很多错误的发生,方便了本体的构建;

③提供了本体的编辑功能和推理功能,用户可以输入和编辑每个概念的名字、约束、属性、实例等内容,并可以基于这些知识进行推理或获取新的知识。

已经有很多文献对各个工具的用法进行了综述性的介绍,并对当前较为流行的几种本体构建工具进行了比较分析<sup>[12~17]</sup>。其中,文献[13]对中文支持能力较好的两种工具进行了比较分析,认为 KAON 能够支持非拉丁字符集,因此其具备支持中文的能力。但由于其检索功能简单、更新速度慢等原因,已无法满足大量用户日益增长的科研和应用需求,导致 KAON 的用户群远不及 Protégé。而 Protégé 并非如文献[13]所认为的仅支持英文,它可以完成中文本体的编辑工作,但由于其插件结构,大部分扩展功能以插件的形式存在,绝大多数插件是不支持中文的,这即是很多人认为“Protégé 不支持中文、很多功能不可用”的看法产生的原因。

### 1.2 Protégé 中文本体构建能力评估

Protégé 是美国斯坦福大学医学院 (Stanford Medical Informatics) 开发的本体构建工具。第一个 Protégé 系统为 1987 年开发,其后经历了 Protege2000、Protege3.1、Protege3.3、Protege3.4 等版本,目前最新的版本是 4.0 版,与以往版本的最大不同在于 4.0 版支持 OWL2.0。它采用 Java 语言开发,提供开源代码,界面风格与 Windows 操作系统的风格一致,采用插件模

式,具有极强的可扩展性。除了提供本体构建的基本功能外,还包括本体复用工具集 PROMT<sup>[18]</sup>。它支持多种数据输入格式,对各项标准支持较好,文件的输出格式可以定制,包括 XML、RDF(S)、OIL、DAML、DAML+OIL、OWL 等语言。Protégé 也存在一些功能上的缺陷,例如:一次只能打开一个本体,不支持协同开发,运行速度慢等。但是,其源代码共享、更新速度快、界面友好、可扩展性高等优良特性仍使它成为当今国内外最为流行的本体构建工具。

较之其他工具而言,Protégé 在支持中文方面又具有如下的优势:

①基于 UTF-8 编码。虽然对于完整的编辑过程来说,其对中文的支持能力不尽如人意,但其核心部分可以提供中文本体基础的编辑功能,且其支持的编码利于中文支持功能的开发。

②开源的特性。Protégé 是开源项目,可以获得源代码,系统分为界面(ui)、模型(model)和存储(storage)等模块,虽然庞大,但结构清晰,有利于系统的分析和扩展。

③良好的可扩展结构。一方面,Protégé 为用户提供可扩展的 API 接口,可以通过修改和替换数据显示和获取模块来适应新的语言,从这个意义上说,对中文本体构建环境的搭建提供了方便。另一方面,可以通过开发自己的插件来完成满足需要的中文本体可扩展功能。

对于中文本体的构建环境而言,笔者认为 Protégé 存在以下缺陷:

a. 界面易读性差。由于本体领域较新,Protégé 的默认构建环境是英文环境,虽然对研究机构来讲不会形成障碍,但是不可否认,它在很大程度上妨碍了本体在应用领域的普及;

b. 辅助功能不完善。Protégé 虽然可以提供对中文本体的基础编辑能力(例如:类、属性、实例的中文编辑功能等),但是作为必需的辅助扩展功能的诸多插件均不支持中文,例如:图形化显示是本体构建过程中很重要的可视化环节,简洁而直观,Protégé 却无法提供中文本体的图形化显示功能。

由上面的分析可知,Protégé 具备诸多无与伦比的优势,但在中文本体构建能力上存在一定的缺陷,不能满足当前的应用需求,笔者针对这些问题,开发了支持中文本体构建的工具并就其中的相关技术问题进行了深入研究。

### 1.3 关于开发中文本体构建工具的思考

笔者在应用研究过程中遇到了中文本体构建的诸多问题,认为当前亟需一款支持中文本体构建的工具,

解决中文本体的支持问题迫在眉睫。虽然 Protégé 的上述诸多优点十分诱人,但是,由于 Protégé 是插件构造,很多插件是用户根据自身需求而开发的,庞杂而缺乏统一规划,有的功能之间存在重复开发的现象,且大多数插件不支持中文,同时,Protégé 中很多扩展功能不常用,日常的应用不需要过于复杂的扩展功能,完全汉化 Protégé 工作量比较大,且不适合应用需求。其次,笔者认为,仅仅汉化是不够的,中文有其自身的特点,应该借鉴 Protégé 的开发结构,从中文的思维方式出发,设计一套中文本体开发工具。因此,考虑选择 Protégé 的核心功能进行汉化作为工具的基本集,再根据中文的特点,确定需求,以开发一个真正意义上的中文本体构建工具。

#### 1.4 本体构建环境的中文可视化

Protégé 提供了两种本体建模方式,基于框架的本体建模和基于 OWL 的本体建模。这两种建模方式可以很好地满足基本构建需求,笔者将这两种建模方式继承下来,并将其构建环境进行汉化,汉化结果如图 1 所示。

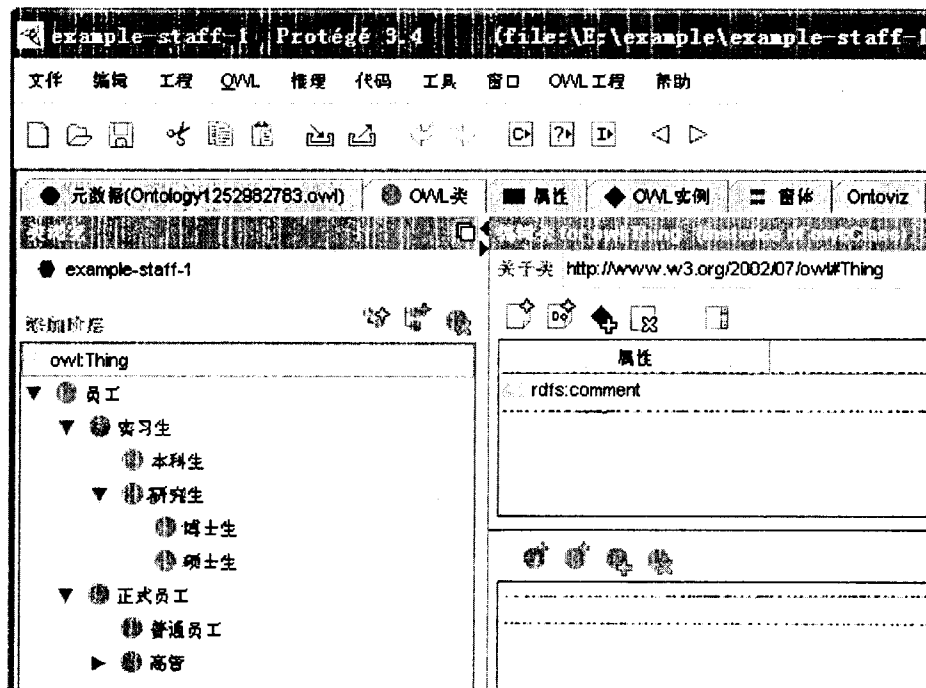


图 1 中文构建环境汉化结果示例

## 2 中文本体可视化

### 2.1 中文本体可视化问题及其表现形式

图形化显示作为可视化的一个重要环节,能够提供对事物更为简洁直观的描述,采用图形化显示可以为本体的构建工作带来极大的方便。Protégé 提供了三种可视化标签: OWLVizTab, TGVizTab 和 On-

toVizTab。OWLVizTab 显示了各个类的层次结构间的上下位关系; TGVizTab 显示本体的树结构; OntoVizTab 展示了本体中包括类、属性、实例等全部元素的图示关系,能够不同侧面、不同深度展示本体。

但这三种可视化标签在处理中文本体以实现中文本体的可视化的过程中遇到了不同现象的显示问题:

① OWLVizTab 无法对中文本体图示进行布局,生成的图形堆叠在左上角;

② TGVizTab 可以正确显示中文类,无法正确显示中文属性,显示为乱码;

③ OntoVizTab 无法生成中文图示。究其本质原因,在于 Java 语言对中文处理存在的编码问题。笔者通过修改和替换接口模块解决了中文本体可视化问题。

### 2.2 出现中文问题的原因

Java 中的字符数据是 16 位无符号型数据,它表示 Unicode 集,而不仅仅是 ASCII 集。这种方式有着不言而喻的优势,然而,由于所处理的信息绝大部分都是英文,7 位的 ASCII 码已足够,使用 16 位的 Unicode 无疑

浪费了大量的存储资源,降低了效率。UTF-8 的出现解决了这一问题。UTF-8 是一种中间格式,是变长内码,8bit 编码,ASCII 不作变换,其他字符做变长编码,每个字符 1-3 byte,通常作为外码。Protégé 是支持 UTF-8 编码的,则其中文本体可视化问题的实质在于 Java 对 I/O 流的编码处理。

Java 中 I/O 流分为字节流和字符流两种,分别由四种抽象类表示: InputStream、OutputStream、Reader 和 Writer。具体实现时字节流的输入输出分

别使用 FileInputStream 和 FileOutputStream,字符流使用 FileReader 和 FileWriter。字节流转换成字符流可以用 InputStreamReader 和 OutputStreamWriter。通过对输入输出流的转换,以及指定输入输出流的编码,解决中文可视化问题。

### 2.3 典型案例

Protégé 所提供的三种中文本体图形化显示插件中,OntoVizTab 所展示的信息最为清晰、全面,它可以



因素对指纹图像质量的影响,但对指纹库FVC2004

表 1 各评测指标统计表

	均值	方差	有效面积 $Q_1$	干湿度 $Q_2$	奇异点	偏移量		
						d	x/H	y/W
图 a	241.56	1.80e+003	0.0506	0.51	有	151.87	0.18	0.47
图 b	246.08	1.28e+003	0.0367	0.59	有	59.89	0.39	0.45
图 c	241.30	1.66e+003	0.0449	0.72	有	99.68	0.46	0.35
图 d	228.81	4.39e+003	0.1177	0.78	无	12.55	0.52	0.49
图 e	250.60	410.2800	0.0285	0.86	有	28.68	0.46	0.47
图 f	223.00	3.77e+003	0.1322	0.43	有	7.47	0.51	0.49

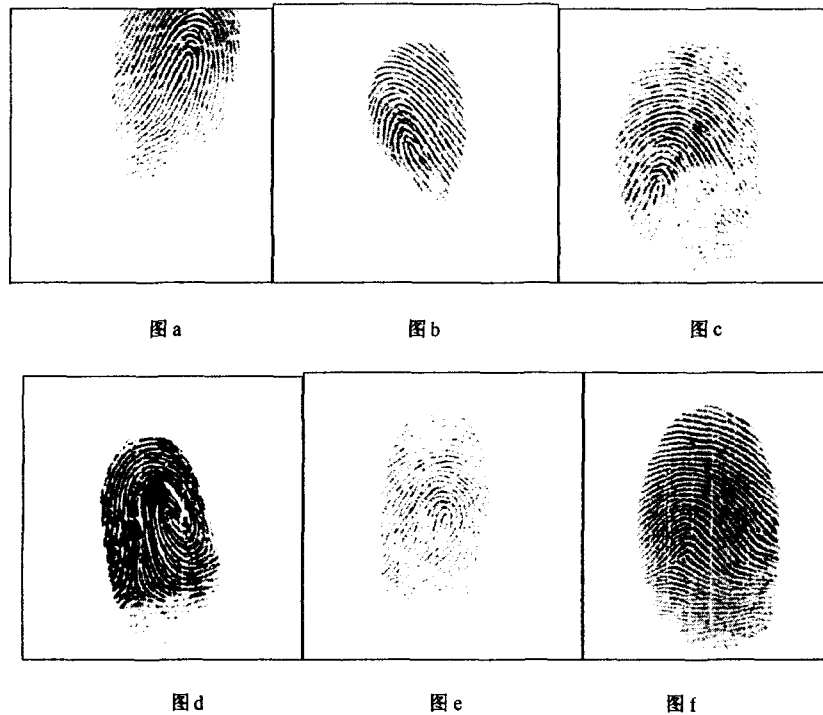


图 1 指纹图像

DB1\_B 中 80 幅指纹图像进行测试的结果证明,该算法可以有效、快速地确定指纹图像的质量,其质量评测的准确性可以达到 95% 以上,具有一定的准确性。

参考文献:

[1] Jain A K, Hong L, Pankanti S. Biometrics Identification[J]. Comm. ACM, 2002(2): 91-98.

[2] Ling Hong, Wan Yifei, Jain A K. Fingerprint image enhancement: Algorithm and performance evaluation[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1998, 20(8): 777-789.

[3] 林国清. 指纹识别中的图像处理研究[D]. 重庆: 重庆大学, 2003.

[4] 曾京文, 汪庆宝, 胡健. 指纹自动识别中心的搜索和特征分块抽取方法[J]. 北京工业大学学报, 1996, 22(4): 115-121.

[5] 谭台哲, 宁新宝, 尹义龙, 等. 一种指纹图像奇异点检测的方法[J]. 软件学报, 2003, 14(6): 1082-1088.

[6] Srinivasan V S, Murthy N N. Detection of singular points in fingerprint images[J]. Pattern Recognition, 1992, 25(2): 139-153.

[7] 沈伟, 陈霞. 指纹图像奇异点提取的一种鲁棒方法[J]. 计算机工程, 2003, 29(2): 45-48.

[8] 田捷, 杨鑫. 生物特征识别技术理论与应用[M]. 北京: 电子工业出版社, 2005.

(上接第 124 页)

trian Conf. on AI. Heidelberg: Springer - Verlag, 2001: 396-408.

[10] Sure Y, Angele J, Erdmann M, et al. OntoEdit: Collaborative ontology engineering for the semantic Web[C]//In: Horrocks I, Hendler J A. Proc. of the ISWC 2002. Heidelberg: Springer - Verlag, 2002: 221-235.

[11] Bozsak E, Ehrig M, Handschuh S, et al. KAON - Towards a large scale semantic web[C]//In: Bauknecht K, Mintjoa A, Quirchmayr G. Proc. of the 3rd Int'l Conf. on E-Commerce and Web Technologies. Heidelberg: Springer - Verlag, 2002: 304-313.

[12] 孙瑾. 本体编辑工具的分析与研究—Protégé2000 对中文本体编辑的适用性探析[J]. 图书情报工作, 2006, 50(12): 26-29.

[13] 范轶, 牟冬梅. 本体构建工具 Protégé 和 KAON 的比较研究[J]. 现代图书情报技术, 2007(8): 19-21.

[14] 杜文华, 董慧. 本体建设工具比较研究[J]. 情报杂志, 2005(2): 5-7.

[15] 陶皖, 廖述梅. 当前本体编辑工具的分析与研究[J]. 计算机工程与设计, 2005(3): 761-763.

[16] 李景. 主要本体构建工具比较研究—上[J]. 情报理论与实践, 2006(1): 109-111.

[17] 李景. 主要本体构建工具比较研究—下[J]. 情报理论与实践, 2006(2): 222-226.

[18] Noy N F, Musen M. PROMPT: Algorithm and tool for automated ontology merging and alignment[C]//In: Proceedings of the 17th National Conference on Artificial Intelligence (AAAI2000). Austin, Texas, USA: [s. n.], 2000.