国内外主要本体库比较分析研究*

白如江 于晓繁 王效岳
(山东理工大学科技信息研究所 淄博 255049)

【摘要】介绍 4 种国内外主要的通用本体库 WordNet、DBpedia、Cyc、HowNet 和两个比较成功的专业领域本体库生物医学和企业领域本体库,从描述语言、存储方式、查询语言、构建平台和应用领域 5 个方面分别对 4 种通用本体库和领域本体库进行比较分析,为国内外学者在本体库及其应用研究方面提供帮助。

【关键词】本体库 WordNet DBpedia Cyc HowNet 生物医学本体 企业管理本体
【分类号】G250.76 G354.4

The Comparative Analysis of Major Domestic and Foreign Ontology Library

Bai Rujiang Yu Xiaofan Wang Xiaoyue
(Institute of Scientific & Technical Information, Shandong University of Technology, Zibo 255049, China)

【Abstract】The paper introduces the major general Ontology libraries in domestic and foreign: WordNet、DBpedia、Cyc and HowNet, and the successful professional domain Ontology libraries: Biomedical Ontology and Enterprise Ontology. Then it separately compares and analyzes them from five aspects as the description language, storage mode, query language, platform building and application to provide assistance for the study in Ontology library and its application.

【Keywords】Ontology library WordNet DBpedia Cyc HowNet Biomedical Ontology Enterprise Ontology

1 背 景

本体(Ontology)的概念最早起源于哲学领域[1],作为语义基础被广泛应用于信息检索、人工智能、语义网络、软件工程、自然语言处理、电子商务和知识管理等领域。为满足企业界和学术界的需求,现已开发出了多种通用的常识性本体库系统(如 WordNet、DBpedia、Cyc、HowNet、Frame Ontology、DublinCore 等)和大量的领域本体库系统。

领域本体库系统方面存在两个问题:

(1)不同的领域积极开发自己领域的本体,如生物医药本体、金融本体、法律知识本体、电子政务本体、新闻本体、旅游本体、生物基因本体等。

(2)同一领域也存在两种情况:由于地域的差异,同一知识范畴出现了不同版本的本体和本体模型;由于领域的概念结构庞大,逻辑结构复杂,产生多个相互关联的本体,这些本体组合起来,共同表示某一领域的知识范畴。

本体如此广泛应用的原因是:它提供了对特定领域知识的共享和共同认识,以便实现人机应用系统中的通信。利用本体技术构建的领域知识库不仅可以清晰地描述领域中的概念及其关系,还可以实现领域知识的共享和重用,且有利于领域知识库的管理和维护。

收稿日期:2010-11-02
收修改稿日期:2010-12-08
*本文系国家社会科学基金一般项目"海量网络学术文献自动分类研究"(项目编号:10BTQ047)和山东理工大学项目"山东理工大学青年教师发展支持计划经费资助"的研究成果之一。

国外对本体的研究项目很多,研究成果已十分丰富,并且建成了许多正在使用的开源本体知识库系统。国内对此的研究十分有限,与国外存在很大的差距。通过对文献的搜集发现,目前国内外关于本体库比较分析研究的论文很少。本文选取了目前4个主要的、较为成熟的通用本体库系统:WordNet、DBpedia、Cyc、HowNet和两个专业领域的领域本体,从描述语言、存储方式、查询语言、构建平台和应用领域5个方面分别进行比较分析,希望为自然语言处理等的研究和科研人员在本体库的选取和使用方面提供帮助。

2　国内外主要的本体库

2.1　WordNet

WordNet(http://wordnet.princeton.edu/)是由美国普林斯顿大学的Miller带领的一组心理词汇学家和语言学家于1985年起开发的大型英文词汇数据库,它是传统词典信息与现代计算机技术以及心理语言学研究成果有机结合的一个产物[2]。目前与WordNet相关的研究已经涉及到德语、法语等其他多种语言,被认为是计算语义学、文本分类等相关领域研究者可获取的最为重要的资源[3]。

WordNet以同义词集(Synsets)为单位组织信息,对查询结果的演绎比较符合人类的思维定势。同义词集是在特定的上下文关系中可互换的同义词集合。它与普通词典的最大区别在于它根据词义而不是词形来组织词汇信息。WordNet关心词与词之间的联系,认为词的意义在于词与词之间的区别和联系,而词与词之间的组织方式显示了词概念之间的区别和关联;词性反映了词汇所包含的概念的类别,在组织中将词汇分成5个类:名词、动词、形容词、副词和虚词。实际上,WordNet仅包含名词、动词、形容词和副词,忽略了英语中较小的作为语言句法成分的虚词集。WordNet使用同义词集表示一个语言符号,重点分析名词、动词、形容词和副词的语义关系,构建了如层级系统、N维空间关系、蕴含关系等关系系统,通过这些关系来表征语言的意义。

WordNet的各个版本均可以从普林斯顿大学认知实验室的网站上(http://wordnet.princeton.edu/wordnet/)免费下载。WordNet3.0数据库中所包含的词汇统计数据,如表1所示。

表1　WordNet3.0的数据库统计数字

统计量 \ 词性	名词 Noun	动词 Verb	形容词 Adjective	副词 Adverb	总计 Totals
个数	117 798	11 529	21 479	4 481	155 287
同义词个数 Synsets	82 115	13 767	18 156	3 621	117 659
单词-含义匹配对数 Word-Sense Pairs	146 312	25 047	30 002	5 580	206 941
单义词和单义含义	101 863	6 277	16 503	3 748	128 391
多义词	15 935	5 252	4 976	733	26 896
多义含义	44 449	18 770	14 399	1 832	79 450
包含单义词的多义词平均数	1.24	2.17	1.40	1.25	
排除单义词的多义词的平均数	2.79	3.57	2.71	2.5	

(注:数据来源 http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html)

因WordNet3.0版本对于安装系统的要求较高,可选择WordNet2.1版,使用WordNet浏览器界面,深入了解其用途。图1是笔者在浏览器中输入"mouse",了解与"mouse"相关的信息。可见,单词"mouse"既有名词的词性也有动词的词性,点击"Noun"选项可以查询其同义词"Syonoyms"、并列术语"Coordinate Terms"、上位词"Hypertms"、下位词"Hyponyms"、摘要"Brief"、下位词"Hyponyms"、完整"Full"、组分概念"Holonyms"、规则的部分词"Meronyms"、继承的部分词"Meronyms"、关联格式的变形"Derivationally Related Forms"和歧义参数"Familiarity"。点击"Verb"选项可以查询其以估计频率排列的同义词、以相似性分组的同义词、并列术语、上位词、关联格式的变形、句式框架"Sentence Frames"和歧义参数。

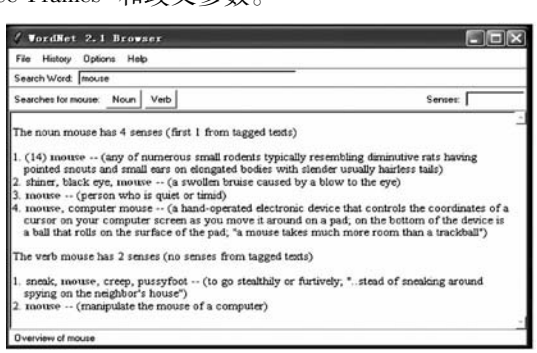

图1　WordNet2.1的浏览界面

如果查询的是形容词,系统可以提供以下信息:同义词和相关名词性概念、反义词、该词的值、关联格式

变形和歧义参数等。如果查询的是副词,可以提供:同义词和以其为词干的形容词、词域和歧义参数。

尽管 WordNet3.0 版本比最初的版本无论是在词汇量还是在用户界面上都有了很大改进,但它的查询范围仍然只限于英文的名词、动词、形容词和副词 4 种词汇。WordNet 将代词归入名词概念中,而定冠词则无法查询。本质上,WordNet 更像一部电子词汇数据库(An Electronic Lexical Database),与真正意义上的本体库相差甚远。由于系统原始条件的缺陷以及词库数据庞大而又无法再进行重新标引等限制因素,WordNet 注定不能成为具有推理功能的系统,而只是"一部基于网络的叙词表检索系统"[4]。

2.2　DBpedia

知识库在提高网络智能和智能搜索方面起着重要的作用,同时也支持信息的集成。由大量人员维护的维基百科(Wikipedia)是人类知识资源的知识库。DBpedia(http://dbpedia.org/About)项目通过从维基百科的词条里抽取结构化数据,以更加有效的方式获得信息来平衡这个巨大的知识资源。基于维基百科数据集,DBpedia 允许用户进行复杂问题的查询,并链接网上其他数据集到维基百科数据集[5]。

目前,DBpedia 描述 340 多万个事件,其中 150 万个一直以本体的方式进行分类,包括 312 000 个人物、413 000 个地方、94 000 个音乐专辑、15 000 个视频游戏、140 000 个组织机构、146 000 个物种和 4 600 种疾病。DBpedia 数据集有特色的标签,它从 90 多种语言中提取 320 万个事件、841 000 种链接图片、5 081 000 个外部网络连接、9 393 000 个外部链接到 RDF 数据集、565 000 个维基百科类别和 75 000 个 YAGO 类别。DBpedia 包含的 10 亿多条信息中有 2.57 亿是从维基百科的英文版本中提取的,7.66 亿是从其他语言的版本中提取的(数据来源 http://wiki.dbpedia.org/Datasets)。图 2 展示了 DBpedia 强大的链接数据。

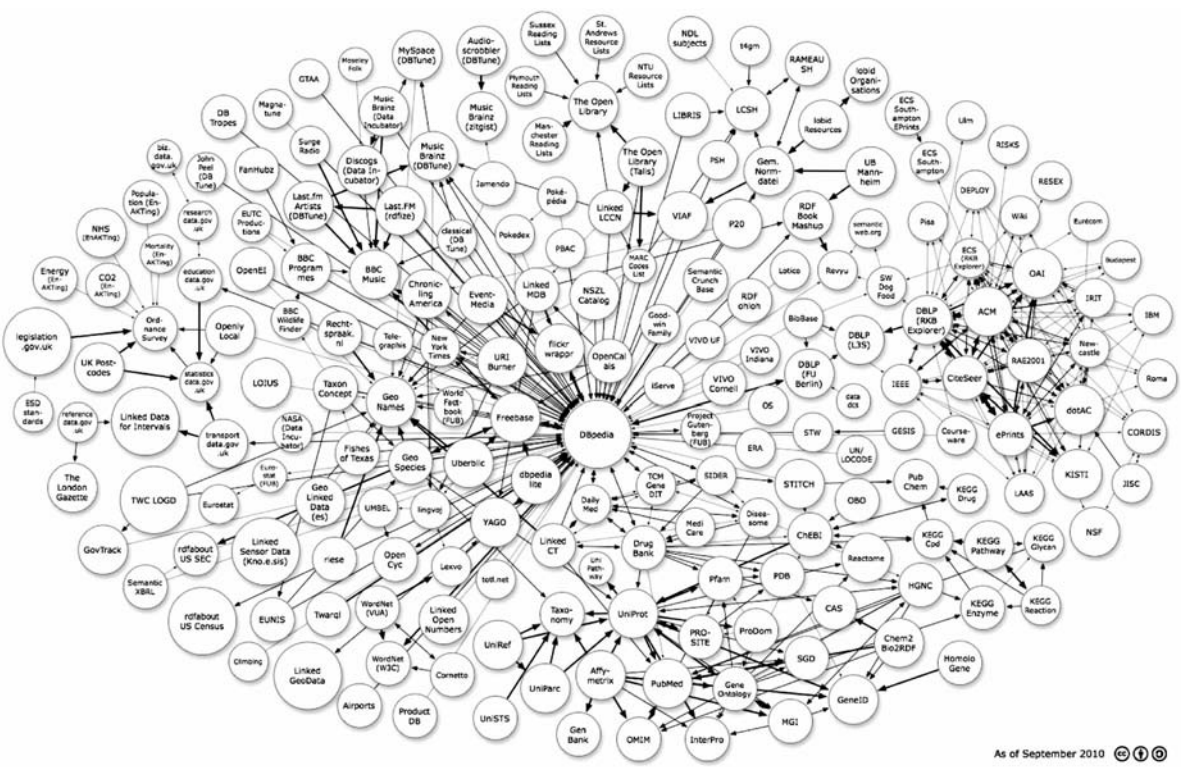

图 2 DBpedia 的链接数据资源

(注:图片来源:http://richard.cyganiak.de/2007/10/lod/,最后更新日期 2010-09-22)

DBpedia 的优势在于:它涵盖很多领域,可代表真实团体的一致意见,能自动地随着维基百科的变化而发展,是真正多语种的。DBpedia 工程展示了一个多种类知识的丰富语料库,这些知识是致力于建立结构化知识库的人们大规模的共同合作的结果。DBpedia 知识库涵盖了一系列的不同领域和这些领域的实体联

系,代表了数以千计的维基百科工作者对概念的一致意见并且随着概念的改变而进化。

2.3　Cyc

Cyc(http://www.cyc.com/opencyc/)提取了单词Encyclopedia(百科全书)中间的三个字母,百科全书并非包括所有的知识,一些显而易见的知识就没有,但正是这些显而易见的知识就是常识性知识,Cyc项目用电脑表示需要了解但百科全书中没有的常识性知识。这个项目始于1984年,由Cycorp集团的总裁和首席执行官、卡耐基梅隆大学和斯坦福大学计算机科学系的教授Dong Lenat[6]发起。Cyc是一个试图综合日常生活常识,建立综合的本体库和数据库的人工智能工程,其目标是使人工智能具有与人相似的推理能力。

1994年度的图灵奖获得者Edward Feigenbaum在2001年1月曾说过:"智能系统的动力是系统所包含领域的知识……Cyc不仅有世界上最大的知识库,也是技术论的最佳代表。"[7] Cyc旨在提供一种可以被其他程序灵活使用的深层次的理解。它的知识库服务器是一个非常庞大的多语境知识库和Cycorp集团自主开发的推理引擎。Cycorp集团的目标是打破"软件开发的瓶颈",构建通用性常识知识基础——集结了术语、规则和关系的语义底层,这一知识库的成功将带来大量的知识密集型产品和服务。Cyc技术包含以下内容,这些技术之间的联系如图3所示[8]:

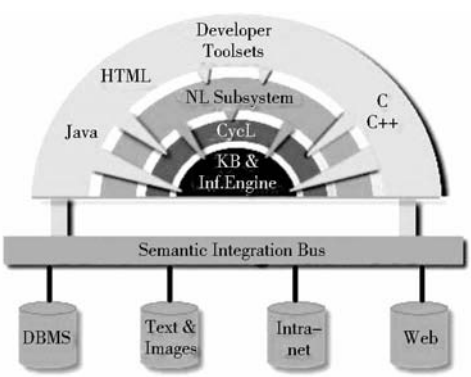

图3　Cyc技术之间的联系

(1)Cyc知识库,利用形式语言CycL,形式化地表达了大量的人类基础知识:事实、规则和用于推理的启发式。知识库的术语构成了庞大的词表和断言集合。Cyc知识库被分成了数以千计的微理论,每个微理论都由一串断言构成。微理论机制允许Cyc独立地维护看起来具有矛盾对立性的断言,并促使Cyc系统提高专注于推理过程的能力。目前,Cyc知识库包括接近50万条术语、1.5万个关系类型和500万条关于这些术语的断言,以及数以万计的手工录入和解释术语的断言。另外,术语合并的功能还可以自动生成数以百万计的非原子化术语。

(2)Cyc推理引擎,可以执行通用的逻辑学推理,还带有AI领域著名的推理机制。Cyc也包括一些特殊目的的推理模型以处理一些特殊类别的推理。

(3)CycL,即Cyc表示语言,是一种非常灵活的知识表示语言。本质上说,CycL是一种增量式的一阶谓语逻辑微积分,它具有易于操作的等式扩展、缺省推理机制和一些二阶谓语逻辑的特征。Cyc用一种定义形式,包括特殊名称假设,能恰当地接近人类的假设。

(4)自然语言处理子系统,由三个部分组成:词典部分、语法分析器和语义注释器。词典部分是自然语言系统的主干,包含英文单词的语法和语义信息。每一个单词都用一个Cyc常量来表示。语法分析器利用松散的基于控制和构建原则的短语结构语法,还利用了大量的与上下文无关的规则为输入的句子构建自底向上的树状结构。语义注释器是Cyc自然语言系统中的语义单元,输出的都是纯CycL语句,一个经过解析的句子可以被立即加入到知识库中。语义单元在解译句子的每一步骤中都会使用知识库中的知识。利用常识来指导解译的程序,可以解决有关自然语言模糊性的任何疑难问题,从而摆脱单纯依靠统计技术的局限。

(5)Cyc语义集成的数据传输总线,如图4所示。信息有很多存储格式,包括结构化、半结构化和非结构化三种。Cyc通常将非结构化信息视为无用信息,保留经过注释的可以为人所获取的信息。Cyc将每一条数据库记录都看作是知识库中隐含的断言,这些断言在进行推理时很有用。

(6)Cyc开发工具包,Cyc系统包含了各种界面工具,允许用户浏览、编辑和扩展Cyc知识库,向推理引擎提出检索式,支持自然语言和数据库集成模块间的互操作。最常用的工具是Cyc的HTML浏览器,允许用户以超文本方式查看知识库,还包含对知识库进行查询和编辑的功能。

OpenCyc是Cyc技术的源代码版本,可通过网络获取(http://sourceforge.net/projects/opencyc/files/),包括Cyc本体的部分内容,以及公理、规则集和推理引

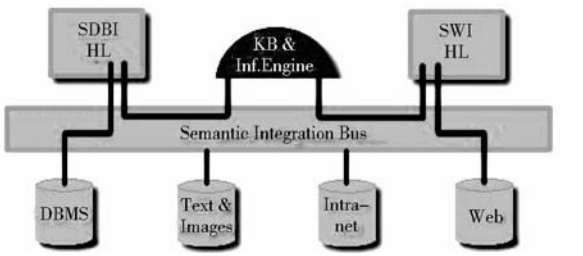

图 4　Cyc 语义集成的数据传输总线

（注：图片来源：http://www.cyc.com/cyc/technology/technology/whatiscyc_dir/whatsincyc）

擎。作为目前世界上最大、最全的综合性知识库和常识推理工具之一，OpenCyc 包含全集的 Cyc 术语和 Cycorp 无偿提供的上百万的本体断言。OpenCyc1.0 版本包含全部的 Cyc 本体，包含成千上万的术语和上百万的术语之间的关系断言，是一个含有客观世界常识知识的顶级本体库；对应于所有概念术语的英语串，帮助研究和显示；Cyc 推理引擎和知识库浏览器的可编辑版本；帮助用户自己把握学习节奏的教学资料；包含知识表示和利用 Cyc 进行软件开发的工具；CycL 的规范说明，还有 CycL—to—Lisp 语言和 CycL—to—C 语言的翻译器；Cyc 应用程序接口（API）的规范说明；Cyc 概念和 WordNet 同义词集之间的链接。OpenCyc1.0 版本还包括一些实验性的开源程序：选择 OWL 输出文件的本体输出；支持 DAML 查询的语义网服务器；推理图程序；Java 版本的 Cyc API[9]。

2.4　HowNet（知网）

HowNet（知网）（http://www.keenage.com/html/e_index.html）是由中国科学院董振东教授开发的一个汉语和英语的常识知识库。德克萨斯大学计算机系知识系统研究小组将知网列为本体项目之一，认为："知网是一个在线的常识知识库，用于自然语言处理。它包含中文词典中概念与概念间的关系，概念的属性与属性之间的关系。同时还包含了与中文对应的英文概念，以及概念的属性之间的关系[6]。"

知网是一个以英汉双语所代表的概念以及概念的特征为基础，以揭示概念与概念之间以及概念所具有的特征之间的关系为基本内容的常识知识库。知网的中文信息结构的描述对象是：由中文词语所表述的、由知网所规定的最基本的运算单元，如万物、部件、属性、属性值、事件、时间和空间等。信息结构的描述内容是：中文词语的各个组成部分之间的、由知网所规定的动态角色关系或属性。通过对信息结构的揭示，可以认识到中文如何描述诸如万物、部件、属性等概念，或如何由简及繁地表达意义，进而揭示中文的语言结构的规律[10]。知网的中文信息结构库数据如表 2 所示：

表 2　中文信息结构库数据（单位：个）

统计项	数量
信息结构模式	271
句法分析式	49
句法结构式	58
词语实例	11 000
总字数	6 万

（注：数据来源 http://www.keenage.com/html/e_index.html）

HowNet 基本组织单位是概念，概念由义原定义。概念与概念之间的关系、概念与义原之间的关系以及义原与义原之间的关系构成了知网的知识体系。义原之间存在复杂的关系，组成了复杂的网状结构。（以下关于 HowNet 的介绍是根据 http://www.keenage.com/zhiwang/e_zhiwang.html 的内容进行的归纳总结。）在知网中共描述了义原之间的 8 种关系：上下位关系、同义关系、反义关系、对义关系、属性－宿主关系、部件－整体关系、材料－成品关系、事件－角色关系。这些关系主要体现在知网的词典和各个特征文件描述中。而在各个特征文件中这些关系体现在特征的层次组织树、必要角色框架和共性特征描述项中。这就使得知网知识库对概念的描述必然是复杂性描述，知网中概念的描述既具有概括性、一般性的描述，又具有针对不同类别的细节性描述，由此而引发了概念描述的一致性和准确性问题。为确保概念描述的一致性和准确性，知网开发出一套知识描述规范体系——知网知识系统描述语言（Knowledge Database Mark－up Language，KDML）。

作为一个知识库，知网的知识结构与其说是知识树不如说是知识图表，它致力于展示概念的一般和特殊属性。例如，对于医生和病人，人是一个一般属性的概念，人的一般属性被记录在概念的主要性能中，作为治病的代理机构对于医生是一个特殊的属性，就像疾病对于病人一样是特有的属性。一个人就是一个一般的属性，但是又享有独有的特性——价值、名字、富有、贫穷、漂亮或者是丑陋。知网不遗余力地反映概念内部关系和属性内部关系的复杂性[11]。

从本质上来看，知网词库中虽然蕴含了大量的概念与概念、属性与属性之间的关系，但是系统仍然以词

汇作为概念的基本单元,不具备本体系统的推理、知识发现等功能,所以知网本身也不是真正的基于本体的系统,它可以作为汉英机器翻译的语料库使用。

2.5　Biomedical Ontology(生物医学领域本体)

现存的生物医学领域的表征足够用于信息检索的目的,但是这些表征的知识组织不适用于计算机推理。计算机推理需要本体提供的有原则的、一致性的组织结构。因此生物医学领域使用各种方法来开发本体,可以从现有的资源中获得本体,也可以通过其他的知识资源获得。

(1)转化医学本体(Translational Medicine Ontology,TMO)

转化医学本体 TMO(http://esw.w3.org/HCLSIG/PharmaOntology)的研究力量来源于 World Wide Web 联盟的医疗保健和生命科学利益集团,并且是生物医学本体国家中心的一部分。TMO 是一个高级的、以患者为中心的本体,它架构了现存的开源领域本体,并为关联和集成全部转化机构以患者为中心的数据提供了框架。转化医学本体为架构转化医学的多个领域提供了术语,这些领域包括假说管理、探索研究、药物开发和规划、临床研究和临床实践。首先从使用案例进行设计,这个本体包含能够映射到其他本体的必要的术语。它作为一个全局的模式服务于数据集成,同时便于异质资源的复杂查询的规范化。

转化医学取决于综合的集成患者的全部数据以评估并促进药物的发展。本体在自动集成患者相关信息数据以促进探索研究、假说管理、规划、临床试验和临床研究方面发挥了重要的作用。语义 Web 技术能够确保使用明确的语义集成异质的数据、对于数据聚集提供丰富和定义明确的表达、在原始数据的基础上获得新知识的逻辑应用。知识表征的 4 个主要的语义 Web 标准是:RDF(Resource Description Framework)、RDFS(RDF Schema)、OWL(Web Ontology Language)和作为查询语言的 SPARQL。开发 OWL 本体支持药物、药物基因和临床试验,并逐渐应用于医疗保健和生命科学中。

TMO 定义了横跨材料实体的 75 种类别(如分子、蛋白质、细胞系、药物制剂)、任务(如项目、目标、有效成分)、进程(如诊断、研究、干预)和信息实体(如剂量、作用机制、迹象\症状、家族史)。TMO 扩展了 Basic Formal Ontology 定义的基本类型和关系本体中的使用关系[12]。TMO 能够使科研人员回答新问题,更快地回答现存的科学问题,也能够帮助制药公司塑造以患者为中心的信息模型,以明确药量和次佳安全的化合物的早期检查。

(2)UMLS Semantic Network(UMLS 语义网络)

美国国家医学图书馆开发了一体化医学语言系统 UMLS(Unified Medical Language System)(http://www.nlm.nih.gov/research/umls/),它的目标是通过获得一体化生物医学资源的词表为大量的生物医学资源的集成提供解决方案。目前 UMLS 连接了生物医学领域的 60 个受控词表。UMLS 覆盖范围十分广泛,不仅包括临床医学的很多概念,还包括大量的生命科学等扩展领域的概念。为提供一个全面的概念框架,UMLS 开发了一个上层本体 UMLS Semantic Network[13]。UMLS Semantic Network 是 UMLS 开发的三个知识库资源之一,这个网络为 UMLS 词表的所有概念提供统一的分类。

UMLS Semantic Network(http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html)包含:

①一套广的主题类别或语义类型,目的是为 UMLS 词表的所有概念提供统一的分类。

②一套有用的重要的存在于语义类型之间的关系或语义关系,这部分文档为语义网络提供概述,并且描述语义网络的文件。

语义类型的主要组群包括:有机体、解剖学结构、生物学功能、化学品、事件、物理对象的概念或观点。这个语义网络有 134 个语义类型,用 UMLS 为所有的概念表示提供了一个一致性的类别。语义类型之间的 54 个链接展示网络的结构并表现了生物医学领域的重要关系。语义类型中的主要链接是"is - a"链接,这种链接确立了网络类型的层次,用于决定有效地分配词表概念的最具体的语义类型。也有一套无层次的关系,主要分为 5 种类型:physically related to spatially related to、temporally related to、functionally related to 和 conceptually related to。UMLS Semantic Network 对于所有请求者的查询都是有效的,并且是免费的[14]。

(3) Gene Ontology(基因本体)

GO(http://www.geneontology.org/)项目是 2000 年由基因本体联盟(The Gene Ontology Consortitum,GOC)研发的。GOC 的目的是要创建一套动态的受控

词表。GO 项目旨在定义出一套结构化的、定义精确的、通用受控词表,可用于描述任何有机生物体中基因和基因产物。GO 项目开发了三个结构控制词表(本体)用来描述基因产物,这三个独立的词表本体是:生物学过程本体(Biological Processes)、分子功能本体(Molecular Functions)和细胞成分本体(Cellular Components)。这项工作有三个独立的方面:开发和维护自身本体;基因产物的注释,确保合作数据库中的本体、基因和基因产物相关联;开发创造、维护和使用本体的工具。

GO 项目是一个合作项目,为解决不同数据库基因产物描述一致性的需要,它合并了三大模式生物数据库,包括:果蝇数据库(FlyBase, Drosophila)、老鼠基因组数据库(Mouse Genome Database, MGD)和酵母基因组数据库(Saccharomyces Genome Database, SGD)。

在 GOC 的官方网站上,对于 GO 有如下定义[15]:

①GO 不是基因序列的数据库,不是基因产物的分类目录。GO 描述的是基因产物如何在细胞环境中发挥作用。

②GO 不是一个指令型标准,不是那种跨系统使用的术语或命名体系。基于参加研究的合作方各自的利益协商以达到一致。

③GO 不是将生物信息数据库进行标准化统一的途径。GO 提供的可共享词表只是迈向标准化的中间步骤,但仅有这一步是远远不够的。

目前 GO 存在如下缺陷:

①知识的变化与更新远远滞后。

②对各种不同的数据,要达成共同的评价或认识很困难。只有在合作方达成共识的基础上,才可以进行基因产物的比较研究,并确定它们之间是否有关联,是否相互作用。

③GO 并没有打算去描述生物学的每一个方面。

2.6　企业领域本体(Enterprise Ontology, EO)

企业本体 EO(http://www.aiai.ed.ac.uk/project/enterprise/enterprise/ontology.html)是与工商企业有关的术语和定义的集合。这个本体是在英国爱丁堡大学的人工智能应用研究学院和它的合作者 IBM、Lloyd's Register、Logica UK Limited 和 Unilever 开发的企业项目(Enterprise Project)的基础上发展起来的,得到英国政府工业与贸易部门的赞助,它是智能系统集成项目的子项目,项目编号是 IED4/1/8032[16]。企业项目目的是通过合作产生一个企业模式化的框架,企业本体为此框架提供基础服务,包括方法和企业模式化的计算机工具箱。

企业本体可以被划分为以下几个主要部分:

(1)行动与过程(Activities and Processes)核心概念是行动。

(2)组织(Organisation)核心概念是法律实体和有组织的单位。

(3)策略(Strategy)核心概念是目标。

(4)营销(Marketing)核心概念是销售。

企业的概念模型必须是连贯的、综合的、一致的、简洁的、必要的。

3　四种本体库的比较分析

3.1　通用本体库比较分析

(1)描述语言

WordNet 词库是一种人机可读的 ASCII 格式,人们可以方便地获得并以自己的方式使用。Grinder 是以 C 语言编辑的多途径编译器,它是一个通用的工具,首要的目的是以词库的格式编译编纂者的文件,能够促进 WordNet 信息的机器检索。它也可作为一个确认工具,当存档系统的还原命令返回时确保编纂者文件的语法完整。

DBpedia 的描述语言是 RDF,目前有两种不同的方法来提取语义关系:把关系型数据库中的关系映射成 RDF;直接从文本和文章的信息盒模板中提取信息。

CycL 是 Cyc 系统的描述语言,CycL 是一种较好的本体表示语言。CycL 的学习与应用都较为便捷,普通用户通过学习可较快掌握其语法结构,而且 CycL 后台有超大容量的 Cyc 知识库,前台有良好的应用界面和推理引擎的支持,这使 CycL 具有优越的应用背景。OpenCyc 项目的目的是要将 CycL 逐渐推广,为用户所接受。它的缺点在于本身不是 Web 的推荐标准,难以作为所有网络资源的标引规范使用。

HowNet 的描述语言 KDML 是知网知识系统语言。这是一套崭新的知识描述规范体系,经过对中英文两种语言各 8 万多概念的描述证明其:有很强的描述能力;便于对意义的计算;直观、有较好的可读性。它包含词汇近 1 500 个特征及动态角色;标识符号和标点;词序[17]。

本体库及其描述语言如表 3 所示:

表 3　本体库描述语言

名称	WordNet	DBpedia	Cyc	HowNet
描述语言	(Grinder)C	RDF	CycL	KDML

(2)存储格式

WordNet 源文件用一种管理多重版本文本文件的 RCS(Unix Revision Control System)存档系统保存。确定使用这个存档系统的原因如下:

①支持 WordNet 词库不同版本的重建;

②保存编纂者文件的所有变化历史;

③防止相同文件制造的冲突变化;

④支持 WordNet 词库版本的不断更新。

存档系统的程序是 Unix Shell Scripts,它能够根据 RCS 的命令控制编纂者源文件,并为编纂者提供一个友好的用户界面。

DBpedia 的存储格式是 RDF 三元组。目前主要的 DBpedia 都用 Virtuoso 和 MySQL 作为存储后台。

Cyc 系统的核心是基于知识的 Cyc 推理程序,它包含两个文档——World 文档和 Cyc 可执行文档。World 文档是知识库知识的副本,已经被翻译成压缩的、可有效下载的二元组格式 CFASL。Cyc 可执行文档包含为推理机和 Cyc 议程编辑的目标代码。推理机允许运行的 Cyc 图像从事实和规则中提取新的结论存储在知识库中。Cyc 议程提取知识库更新操作的流程。Cyc 可执行文档也包含以 Java API 支撑的可编译编码,用于功能界面和网络接口连接,实施基于 CGI 的 Cyc 网络浏览器用户界面的 HTML 产生程序。

HowNet 词典的记录样式是知识词典。知识词典是知网系统的基础文件,在这个文件中每一个词语的概念及其描述形成一个记录;每一种语言的每一个记录都包含 4 项内容;其中每一项都由两部分组成,中间以“=”分隔;每一个“=”的左侧是数据的域名,右侧是数据的值。它们的排列如下:W_X = 词语;G_X = 词语词性;E_X = 词语例子;DEF = 概念定义。

本体库及其存储格式如表 4 所示:

表 4　本体库的存储格式

名称	WordNet	DBpedia	Cyc	HowNet
存储格式	RCS	RDF 三元组	CFASL 和 HTML	概念及描述

(3)查询语言

WordNet 的用户界面有很多种形式。标准的界面是 X Windows 界面,能被移植到一些计算机平台,还应用 Microsoft Windows 和 Macintosh 界面。Shell Scripts 和一些其他程序用于写命令行界面。JAWS(Java API for WordNet Searching)提供了从 WordNet 数据集中检索数据的 Java 应用程序界面。

DBpedia 开发了一系列的界面和存取模块,通过 Web 服务器或者是链接到其他站点就能获得数据集。在 Web 上获取 DBpedia 关联数据集有三种方式:链接数据、SPARQL 协议和可下载的 RDF Dumps,获得链接数据的网络代理包括:

①语义 Web 浏览器如 Disco 和 Tabulator;

②语义 Web 爬虫如 SWSE 和 Swoogle;

③语义 Web 查询代理如 Semantic Web Client Library 和 Semantic Web Client for SWI Prolog。

Cyc 开发了从技术专家到初学者的一系列用户界面。Cyc 浏览器包含大量的动态的 HTML 页面,允许用户进行查询、浏览、编辑或为知识库添加内容。Cyc 系统提供了两个应用程序接口层(APL):SubL 和 Java。SubL APL 提供了允许外部程序访问推理机和据此在知识库上运行的方法。Java APL 建立在 SubL APL 的顶层,并对此进行了改进[18]。

HowNet 的检索以关系为主,无论从概念表、特征表还是直接从关系表入手,都必须通过检索关系表来达到目的。具体来说,就是通过两个关系表的扇入、扇出单头指针联系概念表和特征表,再通过概念表和特征表的扇入、扇出多头指针联系到更多更广的范围,直到相关联的特征、概念、关系都被检索过。

(4)构建平台

WordNet 系统包含 4 个部分:WordNet 词典编纂者的源文件、转换源文件到 WordNet 词库的软件、WordNet 词库、一套使用此词库的软件工具。

WordNet 系统是在 Sun - 4 智能终端的网络开发的。软件程序和工具用 C 程序语言、Unix Utilities 和 Shell Scripts 语言写成。为了更新,WordNet 还被移植到以下计算机系统:Sun - 3、DECstation、NeXT、IBM PC and PC Clones 和 Macintosh。

Cyc 系统包括知识库、词典、推理机、用户界面、副本和副本服务器、分区、语义知识资源集成设备和应用程序接口(APIs)。基于知识的推理程序是 Cyc 系统的核心,它包含两个文档:World 文档和 Cyc 可执行文档。World 文档包含知识库知识的备份,这些备份转换成了简洁的、有效的、能被下载的二元组格式 CFASL。Cyc 可执行文档包含为推理机和 Cyc 议程编码的可编译的目标、为功能界面和联系支撑 Java API 网络接口的可编译编码和实施基于 CGI 的 Cyc Web 浏览器界面的

HTML 产生程序。Cyc 的程序核心是推理机,以表处理机语言 SubL 开发和应用。

HowNet 系统包括下列数据文件和程序:中英双语知识辞典、知网管理工具和知网说明文件。

(5)应用领域

WordNet 的应用包括:图表网络、启发式、语义消歧、自动问答、语义抽取。JWord 是一个有关英语词汇、词间关系信息的 Java 浏览器, 目前使用的 JWord 数据库有三个,WordNet 是其中之一。

DBpedia 的应用包括:

①多面向浏览器,允许通过多面向浏览器检索维基百科全书,以关键字、URI 和标签为入口,为 DBpedia 数据挖掘提供了多种渠道。

②用户应用,例如 DBpedia 手机:用从其他数据库中挖掘的 DBpedia 实体和信息来提供电子地图指导;DBpedia 关系发现器:输入目标就能找到与它有联系的事物;DBpedia 导航:通过 DBpedia 数据进行导航。

③URI 查找服务,通过 DBpedia URI 查找关键字,或者是从关键字、URI 和标签为 DBpedia 数据挖掘提供途径。

④问答系统生成器,基于 DBpedia 和其他数据集,通过拖放式可视化界面建立 SPARQL 问答系统,还可以建立自己的问答系统。

⑤SPARQL 问答系统界面,使用 SPARQL 问答语言来查询 DBpedia。

⑥浏览器增强功能,通过链接相应的 DBpedia 网页增强了 Wikipedia 的网页。

OpenCyc 的应用包括:本体在垂直范围内的快速发展,电子邮件优化、路径选择、摘要和注释、专家系统、游戏开发等;通过扩展 OpenCyc 知识库某一学科领域知识而构建领域本体,可以促进领域本体的快捷开发。利用 OpenCyc 可以解决很多实际的问题,如基于 OpenCyc 开发多种类智力应用程序。

HowNet 的应用包括:语义网络(本体注释、词库、命名实体识别)、语义消歧、汉语极性词词典、基于语义理解的垃圾邮件过滤处理、语义相似度计算、语义关系图的自动构建和多语种研究等。

3.2　专业领域本体库比较分析

由于不同的机构和组织、不同的地域开发自己领域的专业本体库所使用的构建思想和构建方式不同,所以此处的比较分析主要是针对前面介绍的几个科研机构开发的本体库系统进行的,与其他具体机构开发的本体可能不一致。

(1)描述语言

TMO 的描述语言是 OWL。

UMLS Semantic Network 提供了两种格式:关系表格式和单元记录格式。关系表格式是 ASCII 关系格式,它有两个基本表、两个辅助表和两个簿记表。两个基本表中的信息和单元记录文件里的一样,但表示方法却不同。一个包含语义类型和关系的定义信息,另一个包含网络的结构信息。每一个语义类型和每一个关系都被由唯一标识符(4 个字符)指定。辅助表是基本表的扩展,包含网络结构。它们给出了网络中表示的链接的继承集合,第一个表用唯一标识符的三元组格式表达,第二个表用名称的三元组表达。两个簿记表描述了关系文件和它们的字段。单元记录格式也是用 ASCII 表示。单元记录文件中包含语义类型和关系的对立记录。每一个记录都用包含 4 个字符的唯一标识字段开始。每个记录的每个字段都是从新的一行开始,并且持续好几行。有些字段有选择项。

Enterprise Ontology 描述语言是正式的 Ontolingua 编码语言,斯坦福大学知识系统实验室(Knowledge Systems Lab, KSL)的本体编辑工具可生成此编码。

GO 注释基因和基因产物的工具有 Blast2GO、GeneTools、Goanna、GOCat、GOMO 等。

(2)存储格式

TMO 数据以 RDF 形式进行存储。

UMLS Semantic Network 存储格式是用包含 4 个字符的唯一标识符记录语义网络的语义类型、关系和网络结构。

Enterprise Ontology 是语言编码形成的本体被存放在 KSL 的本体库里。

GO 的存储格式分别是文本文件(Flat File,每天更新一次)、XML 文档(每月更新一次)和 MySQL 数据库文档(每月更新一次)。GO 数据库可以免费下载。

(3)查询语言

TMO 的查询语言是 SPARQL。

UMLS Semantic Network 的查询语言是 MS - SQL、MySQL、Protégé 和 MS - Access 2000。

Enterprise Ontology 用 Ontolingua 和 KSL 服务器查询浏览。

AmiGO 由 GOC 开发维护,是 GO 的官方浏览器和

搜索引擎,能够搜索和浏览本体和数据注释。AmiGO 还提供了一个搜索引擎 BLAST,能够搜索 GO 术语中已注释的基因序列和基因产物。AmiGO 可获取 GO 的 MySQL 数据库信息[19]。

(4)构建平台

TMO 用 Protégé 4.0.2 工具构建。

OWL 版本的 UMLS Semantic Network 通过对源文件的语法解析用个性化的 OWL 构造器创建[20],可详细解析源文件的语义类型、关系的基本信息及网络的结构信息。

Enterprise Ontology 领域本体的本体编辑和管理工具有 Tucana、Protégé、OILed、SWOOP。分析工作类似于构建一个概念企业数据模型,并且包括一些技巧如:形成好的抽象的能力、通过谈话从用户中提取信息、通过现存的文档和数据发现信息线索。

OBO – Edit 是由 GOC 开发和维护的开源资源,是一个独立的平台,用于查看和编辑 OBO 格式的本体,它是一个基于图表的工具,重点是为生物学家提供一个友好的基于本体的全局图表架构,能使 OBO – Edit 快速地产生类别相对简单的以关系为重点的大本体[19]。

(5)应用领域

TMO 是一个高级的、以患者为中心的本体,它架构了现存的开源领域本体,并为关联、集成和全部转化机构内以患者为中心的数据提供了框架。

UMLS 词表已成为词典标准在生物医学知识中共享,并被应用于生物医学数据库的信息提取和集成、本体的语义集成等。

Enterprise Ontology 是与工商企业有关的术语和定义的集合。

GO 项目旨在定义一套结构化的、定义精确的、通用的受控词表,可用来描述任何有机生物体中基因和基因产物的作用。

4　结　语

早期的本体研究工作是围绕词典、叙词表等资源展开的,面向的领域是机器翻译和初级的自然语言处理。WordNet 是围绕着西方经典辞书和其他语种与英文的双解词典展开的,知网的词义定义基础也得益于《现代汉语词典》。WordNet 可以被认为是一种表象,这种表象体现了词汇所表达的概念之间的语义关系,而这种语义关系可以通过 HowNet 中有关义原的关系得到解释。也就是说,WordNet 中所描写的各种语义关系能够通过 HowNet 中的义原得到验证、推导。DBpedia 类似于一部百科全书,是一个十分丰富的多种类语料库,但与 OpenCyc、WordNet 和 HowNet 等手工本体相比,DBpedia 的不足之处是:没有形式化的结构,数据质量低并且数据不统一。经过不断发展,Cyc 终于走出了只能成为一部“百科全书”的局限。它具有完备的常识库和经过多年检验和修改才逐渐完善的概念/类的体系结构。系统具有概念与概念间的关系、实例及公理等本体必备元素。它具有自己的标示语言 CycL,利用形式化语言的描述,以断言的方式来定义概念和类,然后再不断添加到数据库中。它利用微理论来定义和区别不同概念出现的语境。通过这种机制,Cyc 知识库系统将越学习越聪明。随着常识的增多,其解决问题的能力也将以几何级数增长,从而有望成为新一代专家系统的原型。

当今的本体研究要解决机器如何理解自然语言的难题,以及多语种问题。WordNet 和知网可以作为早期进行本体系统开发的雏形,DBpedia 是个大型的多种类语料库,Cyc 不仅有完备的开发工具和标示语言,还有大型的自主开发的知识库作为领域本体的概念基础,是具有推理能力的最为完备的本体库系统。

本体的研究虽然起于人工智能领域,但专业领域本体的构建不仅需要本体工程师,更加需要专业领域专家的参与,以实现对知识体系构造、组织和完善,由于专业背景和研究目的的不同,两者统一协作也存在一定的困难。即便是同一领域的专家,对同一问题的看法也未必一致。所以专业领域本体构建的前提是领域专家对专业知识及系统功能达成共识。

无论是通用本体库系统还是专业领域本体库系统,都是在自然语言处理中受到广泛重视和使用的在线知识资源库。它们已应用于自然语言处理的各个领域,如句法歧义消除、语义歧义化解、信息检索、机器翻译等。上述各本体库各具特色且不可替代,都拥有稳定的用户群体,其研究人员或者是开发者都在尽力完善其功能并提供更加友好的界面,以便更好地为用户服务。然而,这些本体库也各有不尽如人意之处。要真正解决这些问题,还有待于开发一种标准化的工具,

这个工具需要满足一定的要求,如具有一定的开放性、提供通用概念体系和常识库、支持符合 Web 标准的统一的输入和输出标示语言、支持多语种并使用 Unicode 字符集、能够广泛应用于 AI 领域和知识表示领域并得到领域专家和 IT 专家的认可。

参考文献:

[1] 张秀兰,蒋玲. 本体概念研究综述[J]. *情报学报*,2007,26(4):527 - 531.

[2] Miller G A, Beckwith R, Fellbaum C, et al. Introduction to WordNet: An On - line Lexical Database [EB/OL]. [2010 - 09 - 01]. http://wordnetcode. princeton. edu/5papers. pdf.

[3] WordNet. A Lexical Database for English [EB/OL]. [2010 - 09 - 01]. http://wordnet. princeton. edu/wordnet/.

[4] 张晓林. 元数据应用与研究[M]. 1 版. 北京:北京图书馆出版社,2002:204 - 205.

[5] Bizera C, Lehmannb J, Kobilarova G, et al. DBpedia - A Crystallization Point for the Web of Data[C]. In: *Proceedings of Web Semantics: Science, Services and Agents on the World Wide Web*. 2009:154 - 165.

[6] 李景. 本体理论在文献检索系统中的应用研究[D]. 北京:中国科学院文献情报中心,2005.

[7] Cycorp. About Cycorp[EB/OL]. [2010 - 09 - 01]. http://www. cyc. com/cyc/company/about.

[8] Cycorp. What is Cyc [EB/OL]. [2010 - 09 - 01]. http://www. cyc. com/cyc/technology.

[9] OpenCyc. Foemalized Common Knowledge [EB/OL]. (2009 - 04 - 08). [2010 - 09 - 01]. http://www. opencyc. org/releases/.

[10] 董振东, 董强. 关于知网—中文信息结构库[EB/OL]. [2010 - 09 - 01]. http://www. keenage. com/html/e_index. html.

[11] Dong Z D, Dong Q. HowNet [EB/OL]. [2010 - 09 - 01]. http://www. keenage. com.

[12] Dumontier M, Andersson B, Batchelor C, et al. The Translational Medicine Ontology: Driving Personalized Medicine by Bridging the Gap from Bedside to Bench[C]. In: *Proceedings of the 13th ISMB'2010 SIG Meeting "Bio - Ontologies"*. 2010:120 - 123.

[13] McCray A T. An Upper - level Ontology for the Biomedical Domain [J]. *Comparative and Functional Genomics*, 2003, 4(1): 80 - 84.

[14] Fact Sheet UMLS Semantic Network. United States National Library of Medicine [EB/OL]. [2010 - 12 - 01]. http://www. nlm. nih. gov/pubs/factsheets/umlssemn. html.

[15] An Introduction to the Gene Ontology. The Gene Ontology [EB/OL]. [2010 - 12 - 01]. http://www. geneontology. org/GO. doc. shtml.

[16] The Enterprise Ontology [EB/OL]. [2010 - 12 - 01]. http://www. aiai. ed. ac. uk/project/enterprise/enterprise/ontology. html.

[17] HowNet Knowledge Database. KDML—知网知识系统描述语言 [EB/OL]. [2010 - 09 - 01]. http://www. keenage. com/html/e_index. html.

[18] Siegel N, Goolsbey K, Kahlert R, et al. The Cyc® System: Notes on Architecture [EB/OL]. [2010 - 09 - 01]. http://www. cyc. com/cyc/technology/pubs.

[19] Gene Ontology Tools. The Gene Ontology [EB/OL]. [2010 - 12 - 01]. http://www. geneontology. org/GO. tools. shtml.

[20] The UMLS Semantic Network in OWL. Temporal Knowledge Bases Group[EB/OL]. [2010 - 12 - 01]. http://krono. act. uji. es/people/Ernesto/UMLS_SN_OWL.

(作者 E - mail:brj@ sdut. edu. cn)