●李 景 钱 平

叙词表与本体的区别与联系*

摘 要 本体是一个关于一些主题的清晰规范的说明,它提供了一个用来表达和交流某些主题知识的词表,还包括一个关系集,关系集把握着词表中这些术语间的联系。它与叙词表的区别表现在逻辑表达形式和组织结构等方面。图 5。参考文献 10。

关键词 叙词表 本体 语义关系 体系结构
分类号 G254.21

ABSTRACT In this paper, the author explains the meaning of ontology and its distinctions with thesaurus in logical expressions and organizational structures. 5 figs. 10 refs.

KEY WORDS Thesaurus. Ontology. Semantic relationship. Architecture.
CLASS NUMBER G254.21

1 叙词表的概念与应用特点

叙词表又称为主题词表,它是一种语义词典,由术语及术语之间的各种关系组成,能反映某学科领域的语义相关概念。ANSI Thesaurus 标准(Z39.19-1980)规定有 13 种词汇间关系。这 13 种关系完全包括了中国《汉语主题词表》的"用、代、属、分、参"结构。

叙词表主要用于检索时的后控制和标引时的自动或辅助选择索引词,是提高查全率和查准率、实现多语种检索和智能化概念检索的重要途径。在 TRS 系统中,叙词表作为一种特殊的数据库,其规模能够达到一般数据库所能达到的规模。

叙词表的使用要点主要有二:

(1)数据录入时,利用主题词可进行正确性校验或选择规范化的主题词进行标引,或进行上位词的自动录入。

(2)检索过程中,可根据主题词表中词汇间的关系实施交互式的导航检索过程,或选择相关的主题词进行检索。利用主题词典函数或自动扩展功能进行多语种和智能化概念检索。

2 本体的概念和应用特点

本体(以下统称 ontology)是一个关于一些主题的清晰规范的说明。它是一个规范的、已经得到公认的描述,它包含词表(或称名称表、术语表),词表中的术语全是与某一学科领域相关的,词表中的逻辑声明全部是用来描述那些术语的含义和术语间关系的(它们是怎样和其他术语相关联的)。因此,ontology 提供了一个用来表达和交流某些主题知识的词表,还包括一个关系集,关系集把握着词表中这些术语间的联系。

构建一个 ontology,可以解决以下 5 个问题。

(1)在用户间或软件代理间达成对于信息组织结构的共同理解和认识[1~2]。假设有若干包含医药信息或提供医药电子商务服务的 web 站点。如果这些 web 站点共享相同的底层 ontology,那么计算机代理就可以抽提和集成这些来自不同站点的信息。代理软件可以利用这些集成的信息来回答用户的检索式或向用户提供数据。

(2)可以复用专业领域知识。譬如,许多不同专业领域的模型均需要描述有关时间的概念。

这些描述中包含时间间隔、时间点、相关的时间测算等概念。如果某个研究组织开发出这样一个详细的 ontology,其他研究组织就可以轻而易举地将它复用到各自的专业领域。而且,如果要构建一个大型的 ontology,也可以将几个现成的 ontology 进行集成。还可以复用一个通用的 ontology 框架,如 UNSPSC ontology[3],将这个框架(skeleton)进行扩展和填充,来描述人们感兴趣的领域。

(3)使专业领域内的假设变得更加明确。对专业领域知识进行明确地规范说明,对于那些必须理解该领域术语含义的新用户来说很有帮助。

(4)将专业领域的知识从运筹学、知识管理的环境中剥离出来。我们可以按照一种必须的规范说明

* 本文的研究得到国家十五科技攻关计划"农业信息智能检索、发布与传播技术研究与开发"专题(2001BA513B01-03)的支持。

和执行程序来完成配置一项产品的任务[4]。

(5)分析专业领域的知识[5]。在进行复用现有ontology和扩展这些ontology的尝试中,对术语进行规范地分析是极有价值的。

3 叙词表与ontology的区别联系

(1)叙词表中的术语均是规范的科学语言,而ontology中的概念、术语可以用自然语言和半自然语言来表达。这是二者在逻辑表达形式上的区别。

(2)在组织结构上,叙词表中知识点的分布是线性的、一维的。而ontology中的知识点、概念分布是网状的,它可以不单纯是一张平面的网格,而是一个在四维空间中伸缩的网状结构(见图1、图2)。

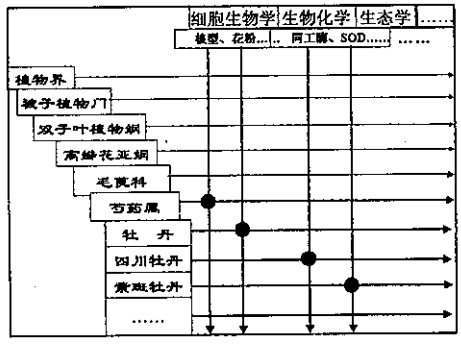

图1 花卉学本体的构建设想,
学科树与植物分类树的交叉

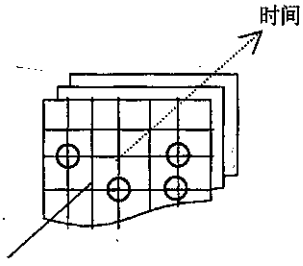

图2 花卉学本体中的知识分布的设想:
动态的、跨越时空的四维网状结构

首先,图1是知识分布的一个瞬时的反映,反映了一段时间里,较为稳定的知识分布情况,但随着时间流逝,学科的发展变化,学科分类(图1中的上部)和植物学分类(图1中的左部)都会产生变化,其中的概念集(图中交叉点处的球体)的内容和分布也会产生变化。交叉点处的概念集是三维的立体结构,它的三个维度分别是:逻辑表达(可以有多种表达形式)、属性(可以有多个属性,还可以再添加和更新)、子类和实例(可以有多个,还可以不断添加)。用历史的、发展的眼光看待图1,它只是图2中的一个页面而已。

(3)Ontology是一个开放集成的体系,它的底层知识库与概念集可以随着学科领域的更新和发展随时进行修正和更新,在这一点上,叙词表则望尘莫及。利用ontology动态更新的特点,可以找出学科发展的规律。在图3a中叙词表代表某一专业领域的理论域,而基于特定目的构建的ontology,它的概念集会与叙词表中的术语有部分的交叉重合,但不会完全一致。随着这一学科的演化发展,ontology中知识体系的不断更新,二者在概念集上会有更进一步地重合,但是永远不会完全重叠。或者可以这样理解,ontology与叙词表的分类等级体系可以是完全一致的,或者说二者包含的概念集完全相同,但图3a中重合的部分代表在ontology中,不仅有这个概念,底层知识库中还必须包含其科学研究进展的相关信息,而不是只有一个"空"的概念。新增加的重合部分,代表特定时间段内的研究热点。而未重合的叙词表中的概念(主题词),则很可能成为未来的研究热点(但至少目前不是)。例如,在2000~2001年,微生物学叙词表中会有"冠状病毒OC43(上位词),新冠状病毒(下位词:冠状病毒的变种)"的概念,但是gene ontology或virus ontology的知识库中很可能没有任何这方面研究的信息,或者即使有,数量也少到可以忽略不计的水平;那么,在2000~2001年可以认为"冠状病毒-新冠状病毒"的研究属于图3b中的"无人问津"领域。2003年春夏,"非典"肆虐,ontology的知识库中可能会爆发式地增加若干"冠状病毒-新冠状病毒"的研究信息,此时"冠状病毒-新冠状病毒"研究就演变为图3b中的"新增的研究领域"。

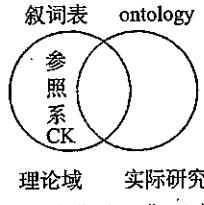

图3a 概念集(术语集)互有重合的叙词表和本体

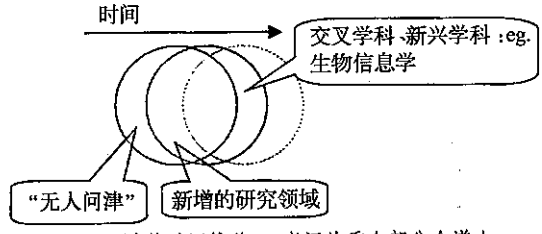

图3b 随着时间推移,二者间的重合部分会增大

图3 利用文献计量学的手段推测出新的研究热点

本体ontology的知识库和概念集中,始终会有些内容是叙词表中没有提及的。这些内容是具有创新意义的前沿领域,如近一两年兴起的以分子遗传学、分子生

物学和 IT 技术为基础的生物信息学、计算机神经网络等新兴概念,在传统的叙词表中就不会存在。这些内容在图 3b 中表示为"交叉学科、新兴学科"。

如果以关系型文献数据库作为基于 ontology 检索系统的底层,那么从图 4 就可以看出学科发展走向和规律。从新的概念(知识点、科研信息)被加入到在 ontology 底层知识库之始,这个时间点可以看做是包含这些概念的学科的增长点,从这一时间点开始,这门学科可能步入了研究热点或主流的行列。从主要利用某些概念进行标引的文献,其数量不再增加之时起,可以看做是一门包含这些概念的学科发展的萎缩点(例如"图书馆学的历史")。如果将次第新增的概念术语按照数量、频次、学科分布和时间顺序进行统计分析,可以推测出下一次新增的概念是什么,大约在什么时段出现,即新的学科增长点是什么? 利用概念之间的彼此组配,可以看到新的研究热点[6]。例如 20 世纪末,IT 技术和分子遗传学、分子生物学都是科学研究的热点主流,21 世纪之初生物信息学的研究逐渐兴起,生物信息学中的很多概念都是前两个学科中概念组配的结果(例如基因组数据库、核酸数据库、蛋白质数据库、序列数据库等)。

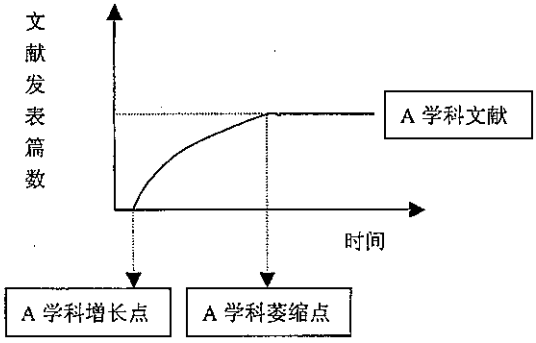

图 4 从文献的发表数量看学科的兴衰规律

(4)叙词表中只包含"用、代、属、分、参、族"这样的简单的语义关系[7~9]。而 ontology 中概念间的关系,则被描述得更为广泛、深入、细致和全面。这是二者的最主要区别,也是为什么基于 ontology 的系统可以实现语义检索和半自然语言、乃至自然语言检索功能的奥妙所在。基本说来,ontology 中概念间的关系分为以下 6 种:(1)A synonym B: A、B 同义;(2)A hypernym B: A 包含 B;(3)A hyponym B: B 包含 A;(4)A overlaps B: A、B 相互交叉重合;(5)A disjoint B: A、B 互不相关;(6)B, C, D cover A: B、C、D……的全集包含 A。

但是通过对概念添加属性,对属性添加逆反属性,属性与属性之间再添加映射关系,ontology 可以体现一些在叙词表中无法描述的关系。例如,在葡萄酒本体 Wine ontology 中,定义 class Wine 和 class Grape,再定义 class Winery;为 class Wine 添加属性 maker 和 staple(原料),maker 的实例就是 class Winery 中的子类,staple 的实例就是 class Grape 的子类。而 maker 的逆反属性 produce 被添加到 class Winery 上,produce 的实例就是 class Wine 中的子类。原料 staple 的逆反属性 brewing(酿造)被添加到 class Grape 中,brewing 的实例就是 class Wine 中的子类。那么这个 ontology 就可以回答"哪种葡萄酿造了哪种葡萄酒,哪家葡萄酒酿造厂生产了哪种葡萄酒,哪种葡萄酒是由哪家酿造厂生成的,哪种葡萄酒是由哪种葡萄酿造而成的"等问题(见图 5)。

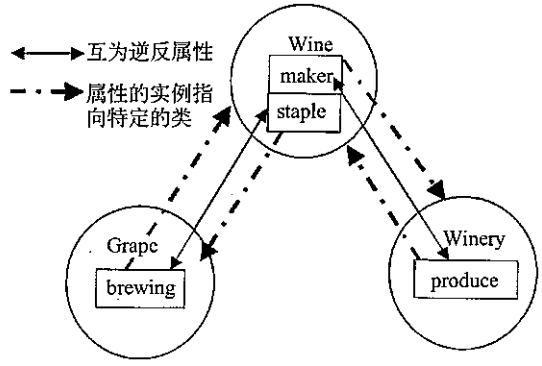

图 5 Wine ontology 中的 3 个类及其属性关系的示意

(5)叙词表是一个词汇库(语料库)但不是知识库,而 ontology 不仅仅是概念集、语料库,还可以是一个知识库。如果说一部叙词表只是一个原始的 ontology,或者可以称之为粗糙的 ontology,或者叫"Light-weight ontology(轻量级的本体)",那么一个实际存在的 ontology 还可以是一个自备知识库或知识底层的完备的知识组织体系,具有智能查询、回答用户问题、机器翻译和预测知识增长点等等传统的基于叙词表的检索系统不具备的功能。

(6)叙词表相对稳定,结构保守而单一,不可能经常修订。而 ontology 的结构应该是一个基于 XML 的开放体系[10],可以被复用,可以增加或减少枝节,可以对每个枝节处进行修改和校验,可以对原有的模式进行保存,还可以"温故知新"。

而且,将概念内容掏空后的框架 skeleton 又可以用作其他领域本体(Domain ontology)的框架(Framework)。譬如,花卉学本体一旦构建完成,如果将其中的木本植物分类部分去掉,就变成为"草本花卉本体";如果将植物分类树的其他不属于观赏植物的部分添加完整,就可以成为一个完整的植物学本体;再将植物学本体推而广之,添加上分类体系框架相同

的微生物学和动物学的知识与术语词汇,就会成为完整的宏观生物学本体。

4 结语

由于目前大量的标引工作必须依赖手工,机器标引和自动编目存在很大的局限性,其精确性尚无法和手工标引媲美。低效率的手工标引成为制约ontology系统构建和完善更新的最大瓶颈。而且研究ontology构建的技术体系与研究检索的技术体系之间存在脱节问题,使得ontology对检索系统的嵌入(登录)成为难点。这还导致了ontology工程的生命周期不能顺利进行。

Ontology的应用有着非常广阔的前景,而针对ontology工程的生命周期,进行自动标引、自动的信息抽提、信息更新和数据挖掘等智能代理技术已然成为ontology研究的焦点。

参考文献

1 Musen, M. A. (1992). Dimensions of knowledge sharing and reuse. *Computers and Biomedical Research* 25: 435-467

2 Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition* 5: 199-220

3 United Nations Standard Products and Services Code. http://www. unspsc. org/

4 McGuinness, D. L. and Wright, J. (1998). Conceptual Modeling for Configuration: A escription Logic-based Approach. *Artificial Intelligence for Engineering Design, Analysis, and Manufacturing-special issue on Configuration*.

5 McGuinness, D. L., Fikes, R., Rice, J. and Wilder, S. (2000). An Environment for Merging and Testing Large Ontologies. *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference* (*KR*2000). A. G. Cohn, F. Giunchiglia and B. Selman, editors. San Francisco, CA, Morgan Kaufmann Publishers.

6 丁学东.文献计量学基础.北京:北京大学出版社,1993

7 丘峰.情报检索与主题词表.北京:书目文献出版社,1988

8 邱明今.主题检索语言.成都:四川大学出版社,1990

9 张琪玉.情报检索语言.武汉:武汉大学出版社,1982

10 Rick Jelliffe著;滴湘工作室译.XML & SGML参考手册.北京:人民邮电出版社,2000

李 景 中科院文献情报中心2001级博士研究生。通讯地址:北京市中关村北四环西路33号。邮编100080。

钱 平 中国农业科学院科技文献信息中心博士、研究员、博士生导师。通讯地址:北京市中关村南大街12号。邮编100081。

(来稿时间:2003-06-03)

(上接第8页)力量是无法满足社会各种需求的,同时我们正在面对入世之后信息咨询市场激烈竞争的压力,面对互联网上中文信息资源极度匮乏的现实,从长远看,只有联合才有出路。

这一点已经成为图书馆界的共识,并在许多方面开展了密切合作。国家图书馆牵头建设的中国数字图书馆工程以及所承担的中华再造善本工程、全图文化信息资源共享工程、送书下乡工程,就是在全国各兄弟馆的大力支持下进行的。借此机会,我代表国家图书馆,对积极参与这些工程的兄弟馆表示衷心的感谢。

最后,我想简单介绍一下国家图书馆文化体制改革试点工作的情况,这也是图书馆界近来比较关注的话题。按照中央部署,国家将在2005年全面展开文化体制改革工作,国家图书馆作为中央确定的25个试点单位之一,先期进行改革的探索,为大家提供借鉴,同时为中央制定总体方案的内容提供依据。这次改革试点,我们立足于1998年以来改革的基础,着眼国际图书馆发展趋势,综合考虑我国图书馆的实际,围绕遵循中央确定的对"公益性文化事业单位要深化劳动人事、收入分配和社会保障制度改革,增加投入,增强活力,改善服务"的要求,从8月初至9月,在馆内外开展了全面调研。在调研过程中,我们开始思考一些问题,如国家图书馆的职能定位问题、如何做足公益事业改善服务的问题、公益事业与有偿服务的关系问题、国家增加投入与自创收入的关系问题、人员岗位设定问题、人员出口问题、如何按生产要素参与分配问题等等。在思考的基础上,我们正在制定《国家图书馆改革试点工作实施方案》,明年初全面展开国家图书馆改革试点工作,为全国图书馆的改革积累经验。

在改革试点过程中,我们一定要认真吃透中央精神,结合实际,紧紧围绕中央确定的关于公益性文化事业单位改革的要求,争取更大更多的政策支持,改善国家图书馆的服务工作,积极推动图书馆法的立法进程,不辜负大家的期望。同时也希望大家多提宝贵意见和建议,献计献策,协助我们做好这项工作。

最后,预祝大会圆满成功!

杨炳延 国家图书馆党委书记、副馆长。通讯地址:北京中关村南大街33号。邮编100081。

(来稿时间:2003-10-25)