●张云秋　冷伏海

领域本体整合的问题及对策研究

摘　要　领域本体整合是在某一特定领域中已有本体的基础上生成新的本体。根据原有本体的改变程度,领域本体整合可进一步分为浅层整合与深层整合。目前,在领域本体整合中存在概念分类体系问题、构建技术问题、本体进化问题和实用性问题。针对这些问题,应采用重构领域本体高层分类、构建基本类间关系本体以及语言分层转换等方法进行领域本体的整合。图 1。参考文献 10。

关键词　领域本体　本体　本体整合　对策研究

分类号　G253

ABSTRACT　Domain ontology integration is to generate new ontologies on the basis of existing ontologies in a specific domain. According to the changes of the existing ontologies, domain ontology integration can be divided into low-level integration and deep-level integration. At present, there are problems concerning concept classification systems, ontology construction technologies, ontology evolution and practicality. To solve these problems, we should reconstruct domain ontology high-level classification, construct basic class relationship ontology and other methods to integrate domain ontologies. 1 fig. 10 refs.

KEY WORDS　Domain ontology.　Ontology.　Ontology integration.　Strategic studies.

CLASS NUMBER　G253

领域本体是通过定义类、实例、属性、关系、公理等元素,刻画出某一领域中的类和实例及其之间的层次关系,对领域知识进行归纳和抽象[1]。近年来,各个领域的本体建设发展迅速,尤其是国外相继建成了很多领域本体,并在实践中得以应用。但这些本体一般覆盖同一领域的不同方面,内容上常交叉重复,而用户在解决某一问题时往往需要一个特定领域的多方面知识。随着用户需求的发展,信息技术的逐渐完善,信息服务将朝着一站式、智能化的方向发展,这就不但需要将分布的、异构的资源进行整合,领域本体的整合也是必然的发展趋势。

1　领域本体整合的含义及类型

目前,对于本体整合的含义并没有一个统一的说法,H. Pinto 等人曾撰文对本体整合的概念进行剖析[2]。笔者认为,所谓本体整合(Ontology Integration),是在两个或多个已有本体的基础上生成新的本体。领域本体整合,是在某一特定领域中已有本体的基础上,生成新的本体。新的本体能促进基于原有本体的计算机系统之间的互操作,可以替代原有本体,也可以作为基于原有本体系统的中介。被整合的本体从概念体系、构建语言到建模方法等方面可能相同,也可能不同。整合后的本体可以是实体,也可以是虚拟的。

根据原有本体的改变程度,领域本体整合可进一步分为浅层整合和深层整合。浅层整合是指两个或多个本体间的映射,也可称为虚拟整合。近年来这方面的研究比较多,但由于单一本体本身存在的一些问题,使得本体间的映射并没有实质性的进展。并且映射仅能支持简单的互操作,并没有太大的应用意义。而深层整合,是指本体之间的融合。目前来看,一个领域内诸多本体的完全融合很难做到。而且一个全封闭的、庞大的、自成体系的本体是没有意义的,因为这样的体系既无法让人验证其逻辑准确性,又无法成为不同系统间互操作的基础,因而没有被复用的价值[3]。而在复用、扩展、修改原有本体基础上的部分融合则更有可行性和现实意义。本文正是在这样理解的基础上对领域本体整合中存在的问题进行剖析,并有针对性地提出相关建议。

2　领域本体整合中的问题分析

2.1　概念分类体系问题

领域本体整合中存在的主要问题是本体间的差异,而概念体系的不兼容是其中的难题。领域本体实际上是对特定领域的概念及概念间关系的精确描述,它是通过分类体系,即类和类间关系来反映的。因此,这一问题可从类和类间关系两方面进行分析。

2.1.1　类的不兼容

领域本体整合中类的问题主要表现在同一概念的不同表达上。以生物医学领域常见的概念"血液"为例,它在 4 个有代表性的生物医学领域本体 GALEN、UMLS、SNOMED 和 FMA 中的类的表达情况如图 1。"血液"同时具有两个不同的上位类:"组织"和"身体物质"。GALEN 和 UMLS 两个本体将血液

直接归入“组织”下，而 FMA 将血液直接归为“身体物质”。SNOMED 介于两者之间，将血液归为“体液”，进而归为“身体物质”，而 GALEN 又将“组织”归入“身体物质”。

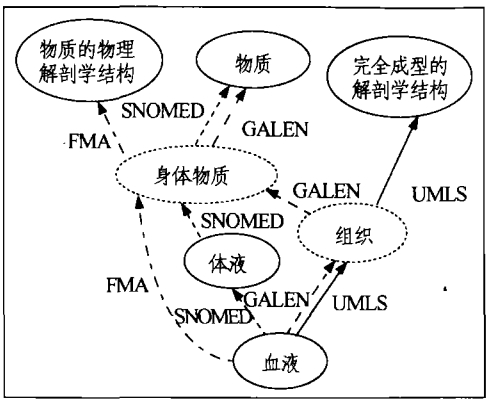

图 1　概念“血液”在 4 个生物医学领域本体中类的表达

表面看来，血液的“组织”和“身体物质”的双重表示并没有揭示出本体间不兼容的问题，如循环出现的等级关系。然而血液作为“组织”和“身体物质”的共同下位类的表示将有悖于同位类的相对原则。同一概念的类的不同表达使得本体之间无法进行概念的正确映射和直接整合。分析这一问题产生的原因，主要在于各领域本体对类的定义上存在差别。即一个领域内对同一类概念尚没有一个统一、完善的定义。

2.1.2　类间关系的混乱

类间关系的混乱主要表现在 7 个方面。

(1)关系定义不清。如 GALEN 中 part of 关系的定义同时描述为：两个概念间的包含关系；生物学实体间可能存在的部分关系；生物学实体间存在的必要的部分关系。这种互相矛盾的定义将导致依据类间关系进行的概念逻辑推理的错误[4]。

(2)不同关系之间没有明确区分。ls_a 和 part of 是领域本体中普遍存在的关系，但在一些本体中却没有明确区分。如在 UMLS 中存在 plant leaves is_a plant[5]。

(3)用基础关系替代特殊关系导致的类间关系混乱。有些本体并没有直接设立诸如位置的相关关系，因此在表达“位于”、“在……内部”、“在外部”等关系时，用 is_a 和 part_of 来表达，如 extracellular region is_a cellular component，而这种表达是不正确的。

(4)相似的关系类型没有详细描述和区分。例如 FMA 中的 derives_from 和 develops_from[6] 两个语义类型在实际应用中很难区分清楚。

(5)缺乏必要的关系说明，例如 replication fork is a part of the nucleoplasm，然而 replication fork 不总是 nucleoplasm 的组成部分，它只是在特殊时期，即细胞周期时是 nucleoplasm 的一部分，而这一点在本体当中并没有必要说明。

(6)关系在实际应用中形成的人工错误。如在 SNOMED 中，存在 diagnostic endoscopic examination of mediastinum NOS is_a mediastinoscope[7]，在这个关系中，前者为过程，后者为工具，将过程归为工具的子类，逻辑上是讲不通的。

(7)技术上导致的错误关系，如 structure of labial vein is_a vulval vein 与 structure of labial vein is_a structure of vein of head 同时存在，这可能是单词“labia”的含义不清晰所造成的，而技术上尚不能消除该词的歧义理解。

2.2　构建技术问题

目前存在着多种构建领域本体的技术与方法，每种方法都有自己的特点。领域本体要进行整合，构建技术上的问题是一个很大的障碍。这个问题可以分解成两个层面，即语言层面和建模层面。前者是指构建领域本体所使用的语言上的差异；后者是指领域本体建模方法上的不同。

2.2.1　语言层面存在的问题

(1)语法不同。这是最简单的一种语言层面的问题。不同本体语言经常使用不同的语法，例如在 RDF 框架下定义类“disease”，在一个本体中的表达是 < rdfs:Class ID = “disease” >，而在 LOOM 中，对这个类的表达却是(defconcept disease)。另外，同一种本体语言也可能具有多种语法表示方法。这类问题的解决可利用重写机制。

(2)逻辑表示法不同。这是语言层面较为复杂的问题。例如，一些语言能够直接将两个不相关的类表述清楚，如 disjoint A B；而另一些语言却是利用对子类进行否定的形式来表达，如 A subclass-of(NOT B)，B subclass-of(NOT A)。这个问题的关键不在于是否事物能够被表达成逻辑上等价的陈述，而是哪一种语言结构应该用于表达某一类事物。这类问题较容易解决，如可利用转换规则实现不同逻辑表示法间的转换。

(3)原语的语义不同。这是语言层面更为复杂的问题。尽管有时同一命名在两种语言中用同一种语言结构来表达，但它的语义还是有区别的。例如“A equalTo B”会有多种解释。而且即使两个本体使用同一种语法，它们的语义也会有区别。例如 OIL RDF Schema 对 < rdfs:domain > 的解释为交集，而 RDF Schema 解释为并集。

(4)语言表现力不同。这是语言层面上最大的问题。指一种语言能够表达一些事物，而另一种语言不能。例如一些语言能通过一些结构表达否定意思，而另一些则不能。衡量一种语言表示能力的最好标准是对公理的定义。

2.2.2　建模层面存在的问题

(1)概念模型的覆盖度和粒度不同。覆盖度是指领域本体对领域知识的覆盖程度，粒度是指领域本体对领域知识的细化程度。例如一个关于 cars 的本体只是对 cars 进行建模，而没有对 trucks 建模；而另外一个本体可能对 trucks 建模，但只是将它们分为几个大类；而第三个本体可能针对一般物理结构、载重量和用途等将 trucks 进行细粒度的区分建模。

(2)概念建模形式不同。有的领域本体通过列举实例来对概念建模,列举实例的方法也会有不同。例如一个时间概念模型可能利用基于间隔逻辑的时间表示法,而另一时间概念模型可能利用基于点的时间表示法。有的领域本体也通过概念描述建模,例如通过建立 is_a 等级来区分不同类的特点。

(3)概念表示形式不同。即表示概念的词汇不同。目前领域本体的概念表示有自然语言、术语、叙词等形式。叙词是规范性语言,与概念成一一对应关系。但对于用自然语言或术语描述概念的领域本体之间,则可能存在词的错误匹配问题。主要有两种情况:一是同义词,如"肿瘤"在一个本体中用词"cancer",而在另一个本体中用"neoplasm"来表达。同义问题相对来说比较容易解决,如可以借助某种工具,实现自然语言到叙词的转换,但需要大量的人工参与。另一问题是多义词,即一个词在不同的上下文语境中有不同的意思,如"conductor"在音乐领域与电子工程领域有不同的含义。多义问题较难解决,需要利用领域知识,同时结合语境进行判断。

(4)编码方法不同。领域中的值会以不同形式来编码。例如,日期可以表示为"dd/mm/yyyy"或"mm-dd-yy",距离可以用米或千米来表示。这导致很多类型的错误组配。但是这类问题很容易解决,可增加转换步骤或用打包的方法去掉这些不同。

2.3 本体进化问题

本体进化问题是本体整合的一个后续问题。该问题的关键是如何在新环境下复用现有本体。目前越来越多的本体都可以通过网络免费获得。在这样一个开放的网络环境中,一成不变的本体已无法反映知识世界的新状态,本体必须随着外部世界的变化而不断进化。领域本体整合本身是一个动态的过程,是随着要整合的领域本体的变化而不断更新的。即整合后的领域本体也存在着进化问题。及时跟踪领域本体的改变非常重要。一般来讲,一个本体在变化后并不能同步地反映到基于该本体的应用和数据当中。因此,建立一种高效地处理本体进化的方法是非常必要的。

本体进化中主要存在两个问题:一是原本体与其进化后本体之间的关系;一是本体与其附属部分的关系,附属部分包括该本体的实例、在该本体基础上构建的其他本体、以该本体为基础的应用本体等。表现在3个方面:

(1)本体进化可能导致整个系统产生语义不一致。例如,新增加一个概念,这个概念可能会与其他的概念之间存在语义冲突。

(2)本体进化可能会导致实例产生数据错误。例如,本体进化时某一实例的原属性被替换,而替换后的属性并不能保证该实例数据的唯一性。

(3)本体进化可能导致基于该本体的应用产生功能错误。对于与本体内部相关的应用,在其内部要用到与本体相关的数据,如果本体被修改,该应用将出现内部运行错误。

2.4 实用性问题

领域本体整合是一个复杂的过程。到目前为止,在国外所进行的一些本体整合的初步尝试中,都需要大量人工参与。也就是说,在短时间内,领域本体整合在各个层面实现完全自动化是不现实的。但是,自动化是本体建设的目标,从这个角度讲,领域本体整合在实际应用中还存在以下问题:能否自动地确定需要整合的概念?是否对整合后的结果进行判断与评价?整合后的领域本体能否被有效地复用?

3 领域本体整合的对策

3.1 针对概念体系问题的对策

(1)重构领域本体高层分类。独立的领域本体对同一个概念的表达并不相同,使得本体之间的映射或直接整合较难进行,解决的办法是重新构建领域本体的高层分类框架。因为各独立领域本体的高层分类存在差别,所以可利用现有通用本体的高层分类,如 BFO 等,根据所整合的领域本体进行改造,并进行合理外推,从而形成粗粒度的松散结构,同时具有可扩充性。该方法需要遵循的原则,第一是尽可能地复用现有本体,也就是从最省力原则出发,总结各领域本体的类目共性,保留大多数本体所采用的一致类目,对少数本体的类目进行修改。第二要对主要类目给出统一、完善的定义,定义要尽量宽泛,覆盖度大。第三要预留可扩展类目,为本体的进化打下基础。该方法也能在一定程度上解决各领域本体的覆盖度及粒度问题。

(2)针对领域本体中类间关系存在的问题,提出构建基本类间关系本体。首先,构建基本类间关系本体的顶层分类,根据对特定领域本体的分析,确定基本类间关系大类,将每一个大类进一步划分为一级类、二级类、三级类等,细化程度可根据整合本体的具体情况而定。同时对每一个基本关系类型列举实例。然后对各关系类型进行清晰的定义及描述,对关系所连接的类的性质进行限定说明,并对关系类型的逻辑属性,即继承性、对称性、反转性和反对称性进行分析,形成基本类间关系顶层分类结构[8]。对该基本结构进行形式化表示,并揭示关系之间的关系,从而构成基本类间关系本体。

该本体能在一定程度上解决目前类间关系所存在的问题,是因为对关系进行清晰定义是根本,是关系本体的质量保证,能从根本上解决关系不一致和含义不清的问题;对关系所连接的类的限定,能在一定程度上使关系类型更确切,并能区别相似的关系类型;对关系的逻辑属性的分析,可使基于本体的逻辑推理更正确。需要注意的是,具体应用时应根据应用对象不同添加或删除关系类型,对关系本体进行修改,以适应更广泛或更专业的应用目的。

3.2 针对构建技术问题的对策

(1)语言分层转换。针对语言层面存在的问题,可通过

以下4种方法来解决。

调整原模型：在通用模型中正式定义一种语言结构。

分层互操作：将语言层面的各个问题，如语法、语义、逻辑等分解人各定义清楚的层中，互操作问题可逐层解决。

转换规则：将不同本体语言中的两个特定结构之间的关系通过一个规则来描述，该规则对两个本体间的语言结构转换进行详细说明。

映射至通用知识模型：将多种本体语言映射到一个通用知识模型，如OKBC，在该通用知识模型中实现不同语言的映射，在映射的过程中要利用转换规则。

(2)用自然语言和一般术语表示概念所存在的问题，可以用交互启发式术语匹配方法解决。该方法提供一种用户参与的机制。首先自动地识别需要整合的词汇，生成词汇列表，提供给用户，并同时生成一个分类列表。该分类列表给用户提供整合后本体的分类范围，采用启发式的策略帮助用户根据词表和分类表，将他认为应该整合的词进行匹配、合并和归类。启发式策略可利用基于语言学的匹配和基于结构模式相似性匹配两种模式。前者是指对某一词的各种变化形式，如缩写、同义词，曲折变化和派生词缀变化等加以汇总。后者是根据词汇的结构和模式，利用相似性匹配算法，发现相似的概念表达，进行整合。

(3)在要整合的领域本体中，如果多数本体是以叙词作为概念的表达形式，可考虑借助现有工具或自行建立工具将自然语言映射至叙词。如NLM建立的MetaMap是一种将文本映射至UMLS叙词的工具，在生物医学领域中已有很广泛应用。MetaMap映射过程主要包括文本解析、变量产生、候选叙词、候选叙词评价和映射构造5个步骤[9]，通过这样5个步骤，就能将自由文本中的词汇及其各种变形词映射到相应的概念上。如果领域本体整合的规模较小，可考虑自行建立映射工具。词语形式的统一将使概念整合更容易。但映射工具并不能保证完全准确地映射，最主要的问题是由于一些词含义的模糊性导致映射错误，需要在映射工具中加入消歧规则，即定义上下文中有意义的同现词。该方法也能部分地解决多义词问题。

3.3 针对本体进化问题的对策

构建一个领域本体进化框架，该框架不但能实现本体自身的进化，还能应用变化扩散的思想，对与该本体相关的所有对象实施进化，以达到整个系统的完整性和一致性。本体整合进化框架的功能包括对每一个概念及其关系消除歧义，即鉴定功能；能确定一个概念的不同版本之间的关系，即跟踪改变功能；能自动地从一个版本到另一版本执行规则，即透明翻译功能。整合本体的进化过程是复杂的，主要包括进化对象的捕获、潜在冲突的剖析、进化对象的更新等处理环节[10]。其中关键是潜在冲突的剖析。剖析内容包括：本体概念间是否有语义冲突，本体与实例间是否有数据冲突，应用与待修改本体间是否有内部相关。若存在不可解决的冲突，系统将中止进化过程并给出相关的说明。如果本体的进化是可行的，系统开始实施实际的本体进化，对与本体进化相关的概念、实例和应用分别进行适当的更新，即实施“变化扩散”，同时应具备进化失败恢复功能。

3.4 针对实用性问题的对策

在领域本体整合的过程中应该能够自动地确定需要整合的概念，其实质是概念相关性的分析评价。而概念之间关联是可继承的，因此就决定了分析评价相关概念的复杂性。解决该问题的方法，可以利用前面建立的高层分类框架和基本类间关系本体。首先，利用概念分类体系，对自动获取的新概念进行归类，然后利用基本类间关系本体分析这一概念和已有概念可能存在的关系，进一步根据这种关系进行整合，最后对分类体系中类的冗余进行修剪。对整合的结果可通过推理机制进行验证，并结合专家知识进行评价。

关于构建语言的表现力及整合后本体的可复用性等问题，目前还没有一个好的解决办法。上述的相关建议，其实还只是针对各类型问题的一个基本的解决思路。它们是否可以整合起来应对一个具体的整合过程中存在的问题，还有待于实践来验证。验证可在两个层面展开：是否能解决上述分析中所存在的问题；是否有助于基于整合本体应用系统间的互操作，进而改善基于领域现象的逻辑推理。

参考文献

1 王海涛，曹存根，高颖. 基于领域本体的半结构化文本知识自动获取方法的设计和实现. 计算机学报，2005，28(12)

2 H. Pinto, A. Prez, J. Martins. Some issues on ontology integration. Proceedings of the IJCAI-99 Workshop on ontologies and Problems-Solving Methods (KRR5), 1999

3 李景. 本体理论在文献检索系统中的应用研究. 北京：北京图书馆出版社，2005

4 GALEN. [2006-05-01]. http://www.opengalen.org/

5 UMLS. [2006-05-01]. http://umlsks.nlm.nih.gov/

6 FMA. [2006-05-01]. http://fma.biostr.washington.edu/

7 SNOMED. [2006-05-01]. http://www.snomed.org/

9 Browne AC, Divita G, Aronson AR, McCray AT. UMLS language and vocabular tools. AMIA Annu Symp Proc, 2003

8,10 周明建，高济，李飞，面向OML的本体进化框架. 计算机辅助设计与图形学学报，2005(3)

张云秋　吉林大学医药信息学系讲师，中国科学学院文献情报中心情报学专业博士研究生。通信地址：吉林省长春市。邮编130021。

冷伏海　中国科学院文献情报中心教授，博士生导师。通信地址：北京。邮编100080。　(来稿时间：2006-07-14)