Thesaurus construction through knowledge representation

通过知识表示构建叙词表

Sean Bechhofer, Carole Goble

Information Management Group, Computer Science Department, University of Manchester, Manchester, Oxford Road, Manchester M13 9PL, UK

Data & Knowledge Engineering 37(2001) 25-45

曾新红，张水英译

2006.4

摘要：描述主题内容的语义元数据在支持数字图书馆的标引和检索方面扮演着至关重要的角色。用于传输这种元数据的机制包括关键词集合、叙词表和分类法。然而，构建一部大型的叙词表是一个困难的过程，但可以通过应用知识表示技术来简化，这一技术是发展用来管理和推理概念的。我们将描绘这样一种方式——描述逻辑（DL）——并示例叙述“描述逻辑”如何参与分类法的构建过程，如何帮助建立条理分明的等级结构，并保障反映在叙词表里的关系是合理的。

关键词：描述逻辑；叙词表；基于主题的分类法（subject based classification）

1、 导言

语义元数据，即描述文献内容的信息，是数字图书馆关心的一个主要课题。元数据延伸到各种类型，包含范围很广，包括主题内容描述；作品或目录信息等等[1]。本文感兴趣的是主题分类和使用内容描述支持查询和检索。一般来说分类信息都是人工化的，但也可以通过分析或综合得到。

目前处理内容元数据的方式主要是利用诸如关键词这样的机制。然而，关键词至少存在两个问题：缺乏使用的统一标准，难以进行前后一致的关键词描述。受控的词汇表和叙词表是限制术语和融合标引与查询表述的一种尝试。

1.1 以叙词为基础的检索

当用来进行检索和查询时，叙词表是实用的，可以在标引者提供的元数据和检索者提出的概念之间搭起桥梁[2]。受控的词表限制了可用的词汇数量，增加了查询会使用合适词汇的可能性。如果叙词表具有像上位词或下位词（BT/NT）这样的关系结构，则可以帮助检索者通过元数据导航找到合适的查询表达方式。如果查询过于宽泛，那么就用下位词替代以便精确查找；如果得到的结果太少，就可以通过上位词扩大检索。相关词（RT）也有助于导航和查询构造。搜索系统可以自动（或应用户的要求）包含下位词来帮助用户。

叙词表的构建是门艺术——词汇通常是手工精心选出的——尽管是按照某些规则或方针，还是会引起解释的含义一致性问题。尤其是当叙词表里的各种*关系*像前述建议的那样用来检索的话，就要求关系具有前后一致并易理解的解释或*语义*。如果扩展查询是自动的话这就显得特别重要。

作为一个例子，我们来看看 Fig.1 的层级，这是 ICONCLASS 分类法的一个部分[3]，显示在 Door（门）这个词下最直接的一些词汇。ICONCLASS 本身并不是一部叙词表，因为它缺乏我们在前面讨论的一些结构，但它确实可以作为不易构建分类等级体系的一个例证。尽管术

语“Monumental door”（纪念门）确实是某种形式的 Door ，在等级里却包含了很多其他的关系，包括分元(partonomy)——Door-knocker（门环）是门的一部分——和相关关系——Closing the door（关门）是应用于 Door（门）的一个动作。这种存在于等级关系里的混乱是很普遍的[4]。如果各种关系用来扩展检索的话，就会出现许多不相干的结果。

41A32 Door
- 41A322 Closing the door
- 41A323 Monumental door
- 41A324 Metalwork of a door
  - 41A3241 Door-knocker
- 41A325 Threshold
- 41A327 Doorkeeper, houseguard (inanimate)

表 1. ICONCLASS 等级结构

Svenonius 认为使用“知识表示体现分类结构”是可行的[5]。正如 Svenonius 指出的，这并不能消除智能录入的需要，但来自知识表示的支持却可以减轻构建过程的负担。在同一篇文章里，也讨论了推理关系术语的可能性，他的建议是“按照某种规则推理出来的相关术语的相关关系比从主观的或漫无目的途径建立起来的在检索时更富有成效。”这就是我们提倡的途径，同时我们会提供一种机制来支持相关关系的建立。

在本文中，我们描述一种知识表示（KR）方案，称作描述逻辑（DL），并显示这种方案可以帮助构建有条理的术语集合。特别是，DL 提供一种建模的描述性途径，可以得到清楚的等级和交叉等级关系（cross-hierarchical）。在这里我们尤其对词汇间的*关系*感兴趣，因为这有助于支持诸如导航、扩充检索 (query expansion)和近似检索（similarity-based searching)等活动。我们讨论使用这一基于可靠推理的复合模型，能够帮助建立起有条理的分类法，为链接和关系提供所需的一致的语义。

我们不想说 DLs 是可以解决标引和检索领域所有问题的百灵药，而是认为这一表示可以作为检索服务组合的一部分在某些特定环境中是有益的。用于本体构建的方法显然也是需要的，但由 DL 提供的性能良好的分类法和推理服务对本体开发者的工具箱是个用益的补充。

我们从想要解决的问题表述开始，对已有的解决方法作简单的回顾，接着描述 DLs，讨论 DL 的特性是如何支持叙词表的构建的。然后会给出用该途径建立叙词表的一些实例，并讨论相关的工作以及将来的研究方向。

这个工作组成了《档案结构化术语学》（STARCH）项目的一部分[6]，我们研究了如何应用 KR 技术改进标引和检索。我们的实例是使用服装术语，这些术语已应用于我们与曼切斯特城市艺术馆（City Art Gallery in Manchester）合作的原型应用系统里。

2、分类法和叙词表

我们沿用 Aitchison 和 Gilchrist[7]的说法，认为*叙词表*是术语的集合，术语间有着某种结构或关系。术语间的关系表示有很多种，包括上位/下位词，相关关系词。已有标准规定叙词表应该显示的几种关系[8]。

该领域包含三方面的工作：

- 分类法(classification)

- 术语合成（composition or synthesis）
- 相关关系(associative relationships)

2.1. *分类法（Classification）*

分类法是术语的集合，这些术语分入或组织到子类目，这些子类目可以是也可以不是建立在 kind-of 关系基础上的——Marcella 和 Newton[9]认为分类法是索引款目的系统化排列，这种方式对于那些寻找信息的人是有用的。这样的分类表（比如《杜威十进制分类法》）在传统的图书馆编目中用了许多年。*等级分类法*是术语的集合，并带有一种关系显示分类或 kind-of 等级结构。

请注意这些表述部分内容是重叠的——很多叙词表也是分类法。例如，《艺术和建筑叙词表》（The Art and Architecture thesaurus），简称 AAT[10]就同时是叙词表和分类法，但不是一部真正的等级分类法（尽管有时被当作等级分类法使用）。当用 BT/NT 链接表示其他关系时，它们并不是纯粹的分类——例如，术语“人”（people）就是“人群”（groups of people）的上位词。《杜威十进分类法》（The Dewey Decimal Classification System）就是一部分类法，而 WordNet[11]可以被认为是一部叙词表。

我们认为，如果词表是用以检索，尤其是检索的过程包括扩展查询或导航时，等级分类法是至关重要的。

2.2.*合成（composition）*

除了分类法，术语合成这一理念对于支持用户的查询是有帮助的[7]，有很多支持词汇合成的技术。

2.2.1. 先组式系统（pre-coordinate systems）

在先组式系统里，术语被组合进用作标引的线性字符串里，于是检索词必须将术语按正确的顺序结合起来。我们可以通过看一个标引词串里的首词来实现扩充检索，但这仅仅适用于一个术语。当检索变得更精确时，查全率会因为词序的任何变化抑制了查询而受损。我们可以通过限定词的使用顺序（称作引文顺序）来解决这一问题。其中之一就是记录所有的轮排顺序，但这会使索引篇幅太大。

2.2.2. 后组式系统 (post-coordinate systems)

在后组式系统里，一个文件可以有很多赋予给它的标引词——这些词不是组合在一起而是保持独立。检索时，使用和标引词相匹配的词的组合来查询。后组式系统可以使用布尔运算符（如 AND，OR 和 NOT）、词语片断搜索或类型匹配。*链接*（links）被用来显示哪些术语被结合在同样的文档中——结合体用索引里的一组术语来表示。链接改善了查准率，但可能会影响查全率。*角色*（roles）附加在术语后，说明术语的使用或意义。然而角色却难以一致使用（包括标引和检索两方面）。

因素分析（factoring）是将术语分成几个组成部分[7]，这些部分或者用作先组式标引的组件，或者用作后组式标引的元素。*因素*可以增进一个标引的查全率性能，但也可能影响查准率。另一方面，保持复合词（compound terms）可提高查准率，但也会引发叙词表和索引的维护问题。

2.2.3. 分面分类法（faceted classification）

Vickery[12]指出，分面分类法是基于概念组配（coordination）的观念之上的，里面的主题要素由两个或更多符号来表示，因此一部分面分类法是深深扎根于联合(combination)这一

观念上的。术语分成各个小组，按等级结构组织。分面分类法主要用于先组式系统，使用分面方式可以在几方面有帮助：*分类法编纂者*（或称建模者）先收集一大批需要组织的术语，很多是复合词，需要用上述方法进行因素分析；分面是这一工作的支持手段，指明概念拆分的方式。因为分面明确了复合词的结构和术语应该按何种顺序组合（这在先组式系统里对改进查全率是极其重要的），*分类人员*或标引员可从中得到帮助；最后，*检索人员*也可得益于此，因为分面帮着明确了查询表达式中词汇可能的组合（combination）。

如果术语用在先组式系统里，我们需要提供分面的顺序来确定术语如何在标引中使用——提供有效的途径确保术语的合成。

2.3. 相关关系

除了分类法和词汇合成，还要采用第三种方式——*相关关系*和 RT 链接。叙词表标准[8]描述了多种需要显示 RT 链接的情形，Lancaster[13]提出了许多相关关系的种类，包括：职业和从事这一职业的人员，如会计师和会计学；事情或行为和它的反对物，如害虫和杀虫剂；行为和结果，如建造公路和公路。这些 RT 链接随后可用来进行导航、扩展或构造查询。

2.4. 本体

一部词表，也可以理解为分类法，术语表或叙词表，应具有底层的本体——按 Guarino[14]的说法，本体就是“一种有意的语义结构，它对约束一种现实的结构的隐式规则进行编码。”*形式化*本体具有一些底层的逻辑结构，它允许我们推理本体里的概念。我们正是要应用这种推理来增加要构建的等级的条理性。

2.5. 叙词表存在的问题

由于对术语间的关系没有一致的解释以及在词表之外术语合成的有效产生，由此引起了许多的问题。即使一些方式（如链接，角色或特殊的次序）试图强加某种解释于合成之上，但也只是临时性的。组合术语的合成独立于分类法之外——术语合成和等级构建是极其松散地耦合在一起的，如果存在这种耦合的话。

在多数情形下，分类法自然是*图表*（graph）结构而非树状结构，使用单一的继承会引发许多问题，必须选择把术语放在等级结构的哪个地方，而维护一部多继承的分类法，如果没有相应的支持是非常困难的。

对增量式变化和发展的支持是需要的，尤其当术语集合是自底向上建立，分类法结构也是一点点推导出来时。术语集合的维护或额外术语的增加也可能需要改变分类法。使用严格的墨守成规的分类法结构使这个过程相当困难。

相关关系的*完整性*也是一个问题，比如在 AAT 里，小提琴手和小提琴，吉他手和吉他等之间有 RT 链接，但每一个都必须显式地表达出来，要保证所有的乐器演奏者和相应的乐器都有链接是一件困难、恼人而昂贵的工作。同时，如果这样做的话，会造成层级里的许多重复，就这个意义上来说，音乐家层级和乐器层级可以共享非常相似的结构。

我们建议通过使用一种在某种统一的框架内自然地支持概念合成和分类法的表示法，我们能够提供一部具有一致解释的词表。这样的表示法是由 DL 来提供的。

3. 描述逻辑

描述逻辑（DLs）是一组基于类（class）的知识表示语言，源于 KL-ONE（要全面了解 KL-ONE 请参阅[15]），它允许概念模型的表示和构建。一个 DL 模型是建立在下列概念之上的：*概念*（concept）——表示具有相似特性的一类对象；*个体*（individual）——概念的实

例；*角色*（role）——个体之间的关系。DL 的核心是包含（subsumption）和分类(classification)的观念——当一个概念的所有实例必然是包含者的实例时，一个概念可以被另一个概念（包含者）纳入其内。包含允许分类层级的建立，概念定义按照从一般到特殊的顺序排列。

3.1. 分类（classification）和合成(composition)

一个 DL 模型是建立在许多原始的概念定义上的，并包括关于那些原子定义间的包含关系的断言（assertation），有点近似于传统的等级分类结构。DL 和叙词表或分类法的区别在于，它提供很多概念形成运算符以组合原子概念和角色以形成新的概念定义，并提供推理服务帮助我们推断复合概念的解释。特别是，如果我们建立一个新的合成，它是可以被*分类的*，也就是说，它在包含层级里的位置是自动决定的。此外，这种推理是建立在可靠的逻辑原则之上的。出现在核心 DL 里的运算符如表 1 所示（具体和抽象的形式都基于 Baader 等[16]给出）。运算符的语义也显示出来，以对象集合——学科领域——和一个解释函数（•）^{I} 为基础，解释函数将概念名称映射到领域子集，并按照所示规则描绘复合表示。

3.2. 推理服务

术语知识库Σ包含一系列术语学公理（见表 2），这些公理定义基本概念之间的关系和详细说明术语合成的附加信息，如表所示，它们在可能的解释集上规定条件。给定一个术语知识库Σ，DL 提供许多*推理服务*(reasoning service)，即用户可以从显式给出的知识推理出隐式知识的推理机制。

包含（ subsumption），给出两个概念描述 C 和 D，当所有 D 的实例都一定是 C 的实例时，称 C 包含 D，即$D \sqsubseteq C$。

分类（classification）,使用包含，我们可以建立概念定义的分类点阵，类别（classification）就包含关系而言是最小的，所以当 A 包含 B 和 B 包含 C，A 和 C 之间将没有直接链接。

可满足性(satisfiability)，给定一个概念描述，我们可以检验出该描述是可满足的，即可以找到一种模型，该描述在其中有一个非空的解释。例如，概念描述（C ┐C）是不可满足的，因为我们不可能有某样东西既是 C 又不是 C。

Operator	Concrete	Abstract	Semantics
Top	TOP	$\top$	Δ^I
Bottom	BOTTOM	$\bot$	$\emptyset$
Conjunction	(and C D)	$C \sqcap D$	$\{c \mid c \in C^I \cap D^I\}$
Disjunction	(or C D)	$C \sqcup D$	$\{c \mid c \in C^I \cup D^I\}$
Negation	(not C)	$\neg C$	$\{c \mid c \notin C^I\}$
Existential	(some R C)	$\exists R : C$	$\{c \mid R^I(c) \cap C^I \neq \emptyset\}$
Universal	(all R C)	$\forall R : C$	$\{c \mid R^I(c) \subseteq C^I\}$
Atmost	(atmost n R C)	$\leq nR : C$	$\{c \mid R^I(c) \cap C^I \leq n\}$
Atleast	(atleast n R C)	$\geq nR : C$	$\{c \mid R^I(c) \cap C^I \geq n\}$
Exact	(exact n R C)	$= nR : C$	$\{c \mid R^I(c) \cap C^I = n\}$
Inverse Role	(inv R)	R^{-1}	$(R^{-1})^I(c) = \{d \mid c \in R^I(d)\}$

Axiom	Concrete	Abstract	Semantics
Introduction	(defconcept CN)	CN	$CN^I \subseteq \Delta^I$
Introduction	(defrole RN)	RN	$RN^I : \Delta^I \longrightarrow \mathcal{P}(\Delta^I)$
Functional	(functional R)	func R	$\forall c \in \Delta^I, \lvert R^I(c)\rvert \leq 1$
Transitive	(transitive R)	trans R	$\forall c,d,e \in \Delta^I, d \in R^I(c) \wedge e \in R^I(d) \Rightarrow e \in R^I(c)$
Inclusion	(impliesrole R S)	$RN \sqsubseteq RS$	$RN^I \subseteq RS^I$
Inclusion	(implies C D)	$C \sqsubseteq D$	$C^I \subseteq D^I$
Equivalence	(equivalent C D)	$C \doteq D$	$C^I = D^I$

表 2　DL 语法和语义

除了上述的术语或 T-Box 推理服务外，DL 实现也可以提供个体之上的推理服务（称作

A-Box 推理），知识库于是有了许多断言公理来表达有关领域里的个体的事实。*检索*(retrieval)函数使我们可以检索到特定概念的所有实例，*实例查询*（instance checking）让我们判断某个个体是不是一个特定概念的实例，*实现*（realization）决定该个体是其实例的最具体的概念。在本文中，我们侧重于 T-Box 推理，因为它是我们构建概念模型所采用的功能。

为描述这些观点，在表 3 我们给出了一个范例模型，有基本的层级，以及两个关系 madeFrom 与 wornOn，我们建立起一组描述如下：

1. (and Item (some madeFrom NaturalMaterial));
2. (and Item (some wornOn Arm));
3. (and Item (some madeFrom Wool));
4. (and Item (some madeFrom Silk) (some wornOn Leg)).

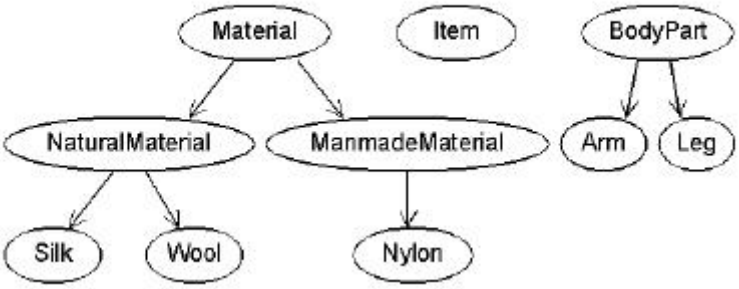

表 3　基本层级

通过对照表 2 给出的描述的语义，我们可以推断出(Item (some madeFrom Silk)and(some wornOn Leg) 的任何实例同样是 Item (some madeFrom NaturalMaterial)的实例。这是包含查证过程能够进行的一种推论。上述例子中的描述能够被分类，如表 4 所示。这里需要指出的重点是这一分类是自动产生的——建模者不必显式安置一个诸如 and（Item (some madeFrom Silk)(some wornOn Leg)这样的复合概念。此外，分类是动态的——在分类被使用前没有必要引入所有需要的合成。描述新的合成时，它们会被置于层级的适当位置。多重遗传在分类中是得到支持的，在上述的例子里，(and Item (some madeFrom Silk) (some wornOn Leg))同时是一种(and Item (some madeFrom Silk)和(and Item(some wornOn Leg))。

3.3. *用作索引的分类法*（The classification as index）

在 A-Box 推理面前，概念层级可以被认为是对个体空间的索引。检索考虑了分类的层级，于是分类法封装了查询内容的层级——如果我们进行一个基于某一高级抽象的查询，那么将返回被包含概念的所有实例。有趣的是我们可以很容易地形成较高级或抽象的概念描述，如概念 Item made from NaturalMaterial 就包含了两个概念(Item made from silk)和（Item made from Wool)。使用 DL 我们可以建立起一致而且清晰的分类层级，可以很容易地建立多轴等级结构（multi-axial hierarchies）（即每个概念在分类法里可以拥有一个以上的父类），而如果用手工则难以做到。当我们进行宽泛检索时多轴分类是十分有用的。

在 Tambis 和 GALEN[17，18]项目里 DLs 已被建议用作本体和词表的传送机制，同时在数字化图书馆也已被建议用作描述数据源的机制[19]。对 DLs 及其使用更详细的叙述请参阅 Borgida 的综述[20，21]。

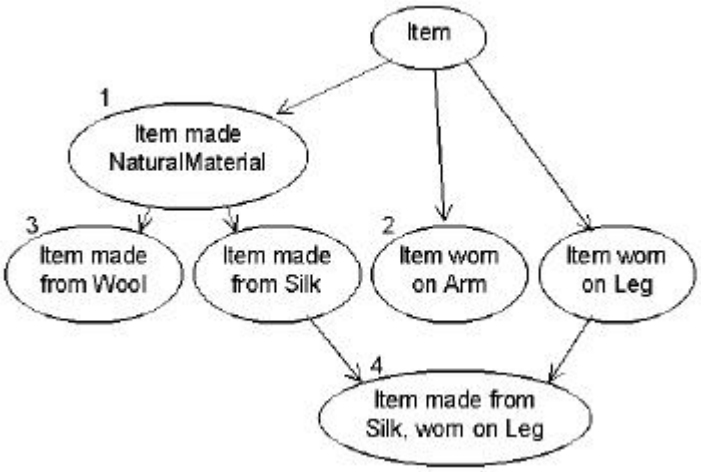

表 4　描述的分类

3.4. FaCT 描述逻辑

以前，因为 DLs 推理的难以驾驭，曾使它被屏除在实际应用之外。尽管通常都知道 DL 语言在最坏的情形下是难以驾驭的，但近期在 DL 界还是有很大的兴趣放在优化推理引擎的实现上，它在实际应用中可以有实在的表现。其中之一是 FaCT，一个在曼彻斯特大学开发的 DL, 它使用基于 tableaux 的推理器（reasoner），加上复杂的优化技术为富有表达力的语言提供可靠而*完全*的推理。FaCT 上的研究工作还在继续，最近的成果是增加了对该语言的受限数字限制（qualified number restriction）[23]以及 A-Box 原型实现[24]。FaCT 的当前实现提供了 SHIQ 语言的表达性[23]，它包含了表 2 中显示的所有运算符号。

在我们计划的体系结构中，正如在文章[25]的作者讨论的那样，DL 被看作是术语学资源。从曼彻斯特大学计算机系我们可以下载到带一个基于 CORBA 的包装器的 FaCT 实现[26]。FaCT 也被用作包括 Tambis[28]和 DWQ[29]在内的许多不同研究项目和研究计划的表示语言，Tambis 是一个系统，提供对多种信息源的访问，在 DWQ 中，DL 被用来进行模式整合和验证。FaCT 提供基本的推理服务，这些服务构成了本体推理层（Ontology Inference Layer ，OIL）的基础，在第 5 部分我们还会进一步论述。

4. 叙词表建模

服装博物馆和收藏国际委员会（ICOM）制作了一部用以描述服装部件的基础术语词表[31]，词表分为三大部分：男装，女装和婴儿服装，每个分部下（即男装/女装/婴儿服装）又分为主要服装，外衣和内衣等等。事实上，词表有许多重复的地方，重复的部分便于合成处理，正如我们在这儿讨论的。AAT 服装层级也是类似的，根据形式和功能组成分支。

我们重新将 ICOM 词表建成 DL 模型，另外我们还合并了 AAT 的部分术语和许多已在 Platt Hall gallery of Costume collection（Manchester City Art Galleries 的一部分）中使用的关键词标引词，该服装廊是一个在检索中使用基于 DL 模型的系统原型实现的素材来源，作为 STARCH 计划[6]的一部分被调查研究。

词表是按自底向上的方式建起来的，使用了一组基本的关系和概念来定义合成方式的概念，迄今在词表中共定义了 300 个概念。

将词表表示为 DL 模型需要使用 DL 推理引擎，也许推理器并不总是传送术语的合适方

式，在很多情形下，传统的静态叙词表可能更好用。既然这样，我们就“输出”叙词表的静态表示，使用关系和 DL 的推理服务决定适当的叙词表链接。这种针对叙词表构建的方法已经在 GALEN-IN-USE 项目中使用[32]，该项目中使用了一个用 DL GRAIL 表示的模型来帮助建立编码方案表示临床术语。尽管最终的叙词表不再用 DL 来表示，但 DL 的使用已经帮助了词汇集合的构造，确保了诸如 BT/NT 这样的关系的条理分明。

4.1. 分类和合成

我们来图解合成的使用，Corset 是一种既可穿在腰以上也可穿在腰以下具有某种支撑作用的衣饰，Shirt 是穿在腰以上的主要服装，Bracelet 是戴在臂上的装饰品，其中每一个都可以用更多的原子概念来定义，即：

Corset $\sqsubseteq$ (and Item (some purpose Support)
WornAboveWaist WornBelowWaist)
Shirt $\sqsubseteq$ (and MainGarment WornAboveWaist))
Bracelet $\sqsubseteq$ (and Item (some purpose Decoration)
(some wornOn Arm))

另外，可以定义更多的抽象概念，例如：

SupportGarment $\sqsubseteq$ (and Item (some purpose Support))

这样，分类器可以照顾到概念 SupportGarment 和 Corset 之间的包含（kind-of）关系，观念 SupportGarment 不必在 Corset 之前引入，但是可以在后面给出定义——这对建立模型是很有用的，我们采用自底向上的工作模式，增量式地引入层级结构的各方面并使推理器能够处理维护层级中一致性问题的负担。当分类层级发生变化时，建模者不需要重新组织层级，因为分类器知道如何处理。

4.2. *通过描述的表达和定义*

在 FaCT 里，包括析取（“or”）在内的概念形成运算符使我们能够以自然的方式为概念间的联合建模。比如说男西装，ICON 给出了对西装的许多不同的描述，按照它们的构成部件进行刻画。例如，西装（suit）可以是三件套的——外套(coat)、马甲(waistcoat)和裤子(trousers)——或者可能只是简单的外套和马甲。

我们可以用“or”运算符和数值限定来给出表 5 里的概念定义。为解释这些描述，我们说 Suit1 是由 Coat, Trousers,Waistcoat 三样东西构成的，而 Suit2 是由 Coat, Trousers 构成的，等等。一般而言，Suit 是已经给出的联合体的任何一个。析取符让我们可以将此表示得很清晰——而在叙词表里，并不能总是很清楚地表达术语联合体是合取还是析取。这种任意的语义会引致术语的混乱和曲解。通过使用 DL 我们可以*精确地*知道每个定义的含义。在 ICOM 分类表里，由 Coat, Trousers,Waistcoat 构成的 Suit1 可以通过使用范畴注释（scope note）或描述来处理。在这种情况下，通过描述，我们显式地表示出透过范畴注释隐式表达的信息。

```
Suit1 ≐ (and Combination (exact 3 composed_of Item)
        (some composed_of Coat)
        (some composed_of Trousers)
        (some composed_of Waistcoat))

Suit2 ≐ (and Combination (exact 2 composed_of Item)
        (some composed_of Coat)
        (some composed_of Trousers))

Suit3 ≐ (and Combination (exact 2 composed_of Item)
        (some composed_of Coat)
        (some composed_of Waistcoat))

Suit4 ≐ (and Combination
        (exact 2 composed_of Item)
        (some composed_of Jacket)
        (some composed_of Trousers))

Suit ≐ (or Suit1 Suit2 Suit3 Suit4)
```

表 5　suits 的定义

分类器现在要注意概念间的包含关系（BT/NT），推断出 Suit1 是一种 Suit2，更有趣的是 Suit2 现在是(some composed _of Trousers)的一种，所以如果我们想引入这个概念去表示包括 Trousers 的联合体的话，Suit2 将被包含在这个概念里。

AAT 含有术语 Tuxedo，被表述为（通过范畴注释）是 Dinner_Jacket 和 Trousers 的联合体，而 Dinner_Jacket 是一种 Coat（但令人惊讶的是，它不是一种 Jacket）。如果我们增加如下定义到知识库：

```
Dinner_Jacket ⊑ Coat
Tuxedo ⊑ (and Combination
        (exact 2 composed_of Item)
        (some composed_of Dinner_Jacket)
        (some composed_of Trousers))
```

我们便能发现 Tuxedo 现在被分类为一种 Suit2。表 6 显示了最终的层级。

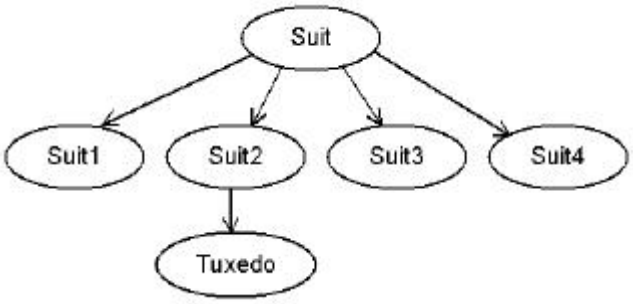

表 6 suit 的层级

P36

这个例子表明复合体的构成部分之间的关系是如何影响复合体和描述之间的关系的——Dinner_Jacket 和 Coat 间的关系导致了 Tuxedo 和 Suit2 间的关系。如果我们改变我们的看法，将 Dinner_Jacket 看作是一种 Jacket 而不是一种 Coat，我们只需要将公理

$$\text{Dinner\_Jacket} \sqsubseteq \text{Coat}$$

改为

$$\text{Dinner\_Jacket} \sqsubseteq \text{Jacket}$$

然后分类器需要注意 Tuxedo 在层级中的位置（在这种情况下它应当是一种 Suit4，而不是一种 Suit2）。层级的组织是由分类器控制和管理的——这个例证表明分类器是如何帮助层级的变化管理和一致性维护的。

4.3. *方法*（Methodologies）

当然，采用 DL 作为底层的表示方式并不能立刻解决本体构建的所有问题！还需要一整套方法来指导建模者准确决定如何分解现实世界（比如，我们是引入基本概念 WornAboveWaist, WornBelowWaist，还是引入带有合适填充符 AboveWaist 和 BelowWaist 的角色 worn？）。使用 DL 并不是让建模者放弃选择合适的表示方式的责任。

然而，使用分类器和推理器，却可以平滑过程，使得建模者能够显式地*描述*世界并能通过那些描述推断出一些结果性的结构。

4.4. *多重分类*（Multiple classification）

简单的叙词表的其中一个问题是多重分类，在许多情况下，允许概念有多重视角是有用的。例如，在 AAT，词汇“雨衣”（Raincoat） 出现在“按形式分类的服装”<costume by form>：outwear : overcoats : raincoats 中。有人可能说雨衣应该同样被认为是一种保护性的衣着。在保护性衣着范畴注释里，有这样的效用说明，对付天气的保护性衣着，应该使用外衣（outerwear）或者它的下位词 。这样的机制似乎并不适合计算机检索。同样地，术语 running shoes 出现在 <shoes by function>之下，而<shoes by function>本身出现在<footwear by form>之下。 人们也许希望 running shoes 出现在<footwear by function>之下的层级里，但事实上并不是这样。

也许那样会更好，将雨衣(Raincoat)同时置于外衣（outwear）和保护性服装（protective garment）之下以及允许跑鞋可以出现在层级里的许多地方。然而，如果缺乏 KR 方案提供的分类支持，这样的多轴层级是很难维护的。

4.5. *相关术语*(related terms)

在 ICOM 模型里，包含关系与 BT/NT 关系是对应的，对关系术语链而言，我们可以检查用于特定定义的描述。例如，头盔（helmet）被定义为 safety 目的而戴在头上的：

Helmet (and Item (some purpose Safety)
(some wornOn Head)

所以头盔（helmet）就被同时分在 SafetyItem 和戴在头上的物品（Headwear）两个地方，同样地，分类器可以推断出 Head 和 Safety 是与 Helmet 相关的，它们被用在 Helmet 的定义中。

现在我们可以引入行为或场合的概念，来描述与行为或场合有关系的穿着，比如，如果我们引入骑摩托车(Motorcycling)是一种行为(Activity)的话，那么可以定义：

MotorcycleHelmet (and Helmet (some worn_during Motorcycling))
MotorcycleGlove (and Glove (some worn_during Motorcycling))

这样，Motorcycling 被推断出是 MotorcycleHelmet 和 MotorcycleGlove 的关系术语(RT)。必要时，以这种方式想出来的关系术语可以沿着 BT/NT 链接继承。在上面的例子中，术语 Glove 还可以被定义为戴在手（Hand）上的物品，从而引出 Hand 和 Glove 间的关系。而这种关系又可以被 MotorcycleGlove 继承并引出 MotorcycleGlove 和 Hand 以及 MotorcycleGlove 和 Motorcycling 间的关系。

在上述的 Suit 例子里，通过在 Suit2 中的描述中使用 Jacket，我们可以推理出 Jacket 与 Suit2 是相关的。Svenonius[5]论述了用这种方式推理出相关术语的益处。

正如 Lancaster[13]及其之前的研究论述过的，可能会出现许多不同种类的相关关系。通过使用关系和底层知识表示中的角色，可以得到更清晰的特定的相关关系。在上述关于 Helmet 的例子中，我们可以看到 Safety 是一个相关术语，因为它是佩戴 Helmet 的目的。

4.6. *折叠的语义*（Collapsing semantics）

我们在 2.3 节曾讨论，可能有许多对 RT 链的不同解释在 DL 中可以显式地表示出来，例如部分（partitive）关系（a Suit is composed Of a Jacket）或用法关系（a Glove is wornOn on the Hand）。然而，当把关系的语义转译进叙词表时，它们会被折叠为一个概念“related to”(RT 链接)，所以要特别小心这种转化。

但是，由于叙词表只提供单一的 RT 概念，而成为叙词表结构与生俱来的表达性问题。且不论产生关系的方法和途径，叙词表结构里极少量的关系也会引致这一难题。尽管有关于叙词表使用的注释和指南——如术语的范畴注释——但如果叙词表是用在检索系统里，这些注释可能并不能被系统访问或解释。因此，编撰者需要提供更丰富的表示以便捕捉不同关系的更精确的含义并帮助它们在系统中的使用。本文讨论的途径只是朝向这一目标的第一步。

正如前面讨论过的，原则性的方法是需要的，因为我们可能不希望每一个关系都导致一个相关术语，而且很明显在这一领域方法论是其核心。如果存在一个角色层级，那么会有些高层的角色我们并不希望以这种方式来使用（如某些高层的分元关系）。再者，底层的 DL 表示和显式的角色间的关系表示给我们提供了“概念衣帽架”来放置我们的需求。

4.7. *自底向上的结构*(bottom up construction)

这里的演习就是一种“自底向上”的叙词表结构，与构建分面分类法的方法近似[12]。我们从 ICOM 提供的一组术语开始建立起叙词表的层级结构。通过使用 DL 提供的分类功能，大体上由描述和复合概念的定义*引导*出这个结构。

这种技术产生的词汇表可以通过 Java applets 程序在网页上浏览，这些程序可以访问用上述的方式产生的静态词表[34]。表 7 显示了这个简单的叙词表阅读器的屏幕拷贝。

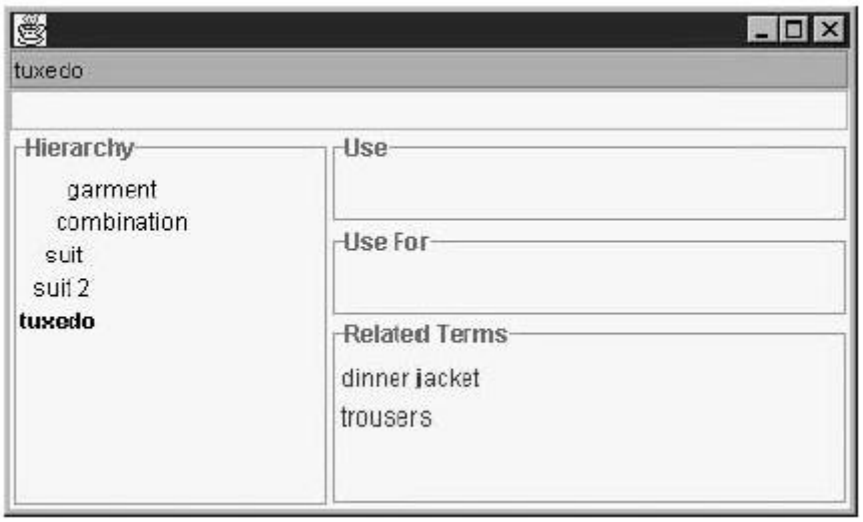

表 7. 叙词表术语阅读器

4.8. *先组和后组*（pre-and post coordination）

这里讲述的方法令合成(composition)理念得以充分运用，并且有自动分类来控制层级的建设，这与术语结合起来产生单个的标引词这种先组理念是一致的。但是，因为合成的顺序无关紧要（在 DL 里诸如合取之类的符号是可以互换的），所以这里仍然存在后组式途径的元素。此外，我们可以用层级去表达抽象的或一般的概念。

在 DL 模型里，由于结合体存在于角色的上下文里，所以可以为并置的概念提供更多含义。同样地，DL 中更丰富的概念合成运作，可以更好地表示概念如何及为何结合在一个复合的表示中，从而提供一个更有原则性的机制，该机制由（例如）链接和角色给出。

合成的 DL 模型可以被看作是分解成了个体分面，相当于顶层抽象概念带有被结合起来提供复合概念的概念。当建立 DL 模型时，需要多思考如何识别和构建合适的模型高级分支[18]。约束的系统，如许可（sanction）[6]有助于确定概念应该如何进行结合。

5. 相关研究

近期许多元数据工作似乎集中在四个方面以使数据易于共享和互操作：

1、描述电子或物理对象的标准标签集或元素集，例如 Cathro 探讨的 Dublin Core[35]或国会图书馆的 Machine-Readable Cataloging[36]；

2、产生和修订标签的规则（如 XML）；

3、确认标准词汇的方式，这些词汇为用于描述对象的元素提供术语；

4、容器体系结构[37]。

这些工作对我们在这儿讨论的是个补充，我们相信可以为内容的一致性描述提供框架——这些术语随后会被用作编目记录中特定字段的值。

Glamorgan 的语义超媒体体系（Semantic Hypermedia Architecture）[38]使用二元关系语义数据模型作为超媒体系统的语义索引空间。该模型以静态的分类法作基础。然而，他们的方式不支持术语推理或动态分类——在传统的叙词表理念里，任何概念的合成都是沿着先组或后组这条线来进行的。

在 Glamorgan 工作的背景中，Alani[39]讨论了由于在结构中缺乏相关术语关系，当使用传统的叙词表检索时会出现的问题。在诸如 AAT 等叙词表的构建中，可能会有编辑信息表明某些术语为什么是相关的。Alani 呼吁更丰富的语义和 RT 关系的子关系显式表示——这正是一个富有表达力的 DL 可以提供的支持。

Ontobroker 项目计划[40]用形式化的本体来表示元数据，但这种 Frame-Logic 表示是静态的，并不能提供 DL 支持的迭代分类。

Meghini[41]采用 DL 用于信息检索，但与我们的方法是不同的，他们企图在一个框架下同时建立形式的（句法元数据）和内容的（语义元数据）两种模型 。这要求扩展 DL 的形式体系去解决特殊具体领域的问题。与此相反，我们提倡用“简单的”DL 来帮助建立有条理的主题内容分类。

Weinstein 描述了一种使用 DL 表示元数据的方法[42]，奇怪的是，他声称不要用 DL 的分类推理服务（下面将讲述）去生成知识库，并进一步指出“……[自动分类]对支持最终用户查询也不是那么重要”。我们认为，推理服务是 DL 表示的最大财富，对支持建立清晰的模型至关重要，可以架起标引者和检索者之间的桥梁。

在参考文献[43]中，Jannink 和 Wiedergikd 阐述了一种依靠在线词典自动分析的叙词表构建途径。当某个术语出现在定义里时就被认为是相关的，然后用一个算法去排列这些关系，找出最强烈的相关关系。该方法并不能揭示出关系的类型，而是简单地衡量它的重要程度。这和我们的方法有几分相似，因为我们建议将术语描述的组成部分用作指向术语的指针。DL 描述的显式结构可为关系的类型提供更多的信息（当然也增加了叙词表建设的难度）。正如 Jannink 和 Wiederhold 所说的，他们的方法不是试图取代手工编撰叙词表，而是可以被考虑作为它们的补充。

Frank 等[44]讨论了利用 CIA World Fact Book 作为来源建立知识基础库。其中一部分是自动的过程，使用初始源数据的结构抽取知识，同时用手工将术语组织进层级。文章作者讨论了是使用几种正交的（多重的）分类法还是用属性来描述术语。允许引入公理（等同于描述）的知识表示可以支持这种方式。作者同时简单提到在建模过程中描述词间关系的方法——比如，“hydropower potential as natural resource”被”energy industry”所利用——但没有详细描述这是怎样进行的。

通过服务传输本体或词表一直是若干领域感兴趣的焦点，已有诸如 OKBC[45]，OMG LQS[46]等服务的规约。 [25]讨论了术语学服务器（Terminology service）的体系结构。

语义元数据语言最新的发展是 OIL（Ontology Inference Layer）[30，47]，它是对面向 Web 元数据语言和标准的一项建议。OIL 利用了框架建模原语，并在此之上建立了定义明确的语义。此种语义是由具有表达力的 DL（带有具体域附加物的 SHIQ）来定义的。这样就可以提供可帮助建模者建立清晰模型的推理支持。这种在“友好面纱”下利用 DL 推理服务的哲学与我们在这里推崇的方法很相近。

文章[48]的作者 Welty 论述了将 DL 用于编目并分析了数种途径，尤其分析了将编目款目表示为属于某些特定类的个体，以及将诸如“aboutness”这样的属性表示为这些个体之间的关系时会遇到的问题。与此相反，我们的建议是与其将 DL 用作本体或编目的基本传输机制，不如将推理服务用于支持模型建设。前面已经提到，在 GALEN-IN-USE 项目中已使用了这种成功的方法[32]。

6. 讨论

我们提倡使用知识表示来支持结构化受控词表的建设。

6.1. 优势

已经证明受控词表如叙词表和分类表(classification scheme)对于编目和检索过程是极具价值的工具。如果词表具有某种结构将会增加标引的能力并有助于检索者进行检索。然而，在很多情况下这种结构仅仅是临时的——如果缺乏对关系的一致解释，检索会得出许多不可

预知的结果。

DL 提供了一种知识表示方案，以一种有原则的方式传递词表，以确保概念表达的一致性。对词汇合成和（自动）分类的支持使 DL 成为受控词表表示的理想选择。

动态的分类允许层级可以以递增的方式建立。因为分类是在合成概念描述的基础上推理出来的，所以不需要对层级进行整体重组就可支持变更和演化。多轴层级很容易被 DL 分类器支持，而多轴层级对建立概念模型是很有用的手段。

正如 Bates[2]所讨论过的，如果能够通过导航改进查询，可以使检索者进行检索尝试并“摸索前进”，并在此过程中学到更多的关于此模型的知识。Bullock 和 Globe[49]也证实了这一点，他们发现，人种学研究分离出一种对信息系统的自发性的需求——随着揭示的信息越多用户的需求也不断变化。利用清晰和严格组织的分类法作为组织模型及其内在关系组织的主干使得这样的导航和浏览更容易。

除了支持分类法的构建，DL 推理器还能够通过使用可满足性测试帮助鉴别不一致的概念，使建模者能够识别出矛盾的概念定义——这种对不一致的识别鉴定同样是重要的，特别是在使用具有表达力的语言的时候。

DL 推理器也可以用来控制与底层概念定义的交互，支持基于表格的界面（form based interfaces）的定义[50，51]。在叙词表里，这有助于提供另一可选的术语集合导航机制（除了通过遍历层级或从检选目录里选择之外）。

6.2. 局限性

简单地使用 DL 并不能解决我们所有的问题，因为使用知识表示并不能排除对智能化输入的需求。仍然必须建立概念模型——GALEN[18]和 TAMBIS[17]研究项目表明建模是一个困难的过程，要求建模者在知识领域和表示语言两方面都是内行。支持工具对建模过程是至关重要的。

期望建模者直接用 DL 形式方法进行工作也是不现实的。GALEN In Use 项目已在运用中间表示法的工作中取得了一些进展[32]，这种中间表示法可以让建模者不必接触 DL 表示的底层句法。尽管如此，通过使用分类和一致性检测推理服务，还是利用了 DL 的强大力量。Data Warehousing Quality 计划使用了一个类似的方法，即用一种工具将 ER 模式转译为 DL 模型，以便对中间模式约束进行定义和一直性检测。

有必要将已经存在的分类法（如 AAT）重新表示为一种复合模型的可能性及其所需要的工具进行调查研究。

本文中提到的例子中，主题领域涉及的对象比较容易被分解或分面为目的、材料、形式等等。这样的领域就很适合使用合成的方法，并可与已成功应用 DL 的主题领域（如医学术语学，软件工程，结构管理）共享属性。而在这样一种分解并不如此清晰的其它领域——如《ASIS Thesaurus of Information Science>[52], 就很难采用 DL 建模。

DL 推理的计算复杂度有时也引为缺点。不过，经验表明在*现实*知识库中极少出现最坏复杂度[22]。开发 DL 表示的有效推理算法是当前的一个研究课题[53]。如果我们扩大使用 DL 的 A-box 功能，会对查询种类有些限制。

我们并不是说 DL 是解决标引和检索的万能良方，而是说在检索服务中使用这样的表示方式在某些情形下是很有效的。

6.3. 基于知识的检索

DL 可以帮助构建受控的词表，当然这只是利用了 DL 功能的一部分。在前述的例子中，DL 被用来建立静态的术语集合（带有相关层级）；如果更进一步把 DL 用作存取机制，无须静态的叙词表，还会有更多的好处，使我们能利用 DL 引擎的动态特性来进行快速的查询表

示设计与分类。

如前讨论的，DL 能支持任意的表达，包括诸如包含裤子(“trousers”)的合成概念这样的抽象。如果我们直接使用 DL，这些抽象的表示可以被用做查询（词汇），而不必事先插入“叙词表”里。

基于主题的导航在诸如 Nanard 和 Nanard’s Mac Web[54]这样的概念超媒体系统里可以起重要的作用，在这些系统里，文档是由带有术语间关系的概念术语标引的，从而导引出文档间的联系。文献[6]进一步探索了基于导航的检索思想。

Bates[2]则建议，用户可以有更有效的检索，只要最初的主题或术语能产生更多可能的相关主题或分类——DL 模型可以支持这一功能，它通过使用约束机制来描述应该如何形成复合概念。在文献[50]和[55]中，作者描述了一种基于表格的模型驱动用户界面，使用户可以用这种方式增量式地细化查询，必要时利用表格为用户指引查询方向，从而可以通过索引空间推进导航[6]。此刻，用户不需与用来标引的生疏的分类法打交道——尽管也存在在特定情形下明确揭示分类法是有益的。缺乏使用分类法经验的用户可以在进行导航和查询构造时得到帮助——信息系统更像中介而不仅仅是简单的查询回答机器。

进一步使用 A-Box 推理器，DL 可以真正地产生强大而灵活的查询引擎，此时 DL 模型提供了一个有关用户可导航内容的空间，将浏览活动与查询结合起来，如[6]所述。不过，真正的 A-Box 对检索来说并不是强制的（可能会引起某些问题，文献[48]进行了讨论）。二者择一，这种检索功能可以用选择或查找服务取代——在 TAMBIS[17]研究项目里采用了这一方法。

致谢

本文得到 EPSRC 资助项目 GR/L71216 和 GR/M75426 支持。作者需要感谢 Richard Giordano (Center for Innovation in Product Development, MIT) 的指导和建议，并要感谢审阅者为我们提供了宝贵意见。