基于范畴论的多 TBox 整合研究

夏战锋[1]，彭志平[2]，胥杜鹃[3]

（1. 江苏科技大学 计算机科学与工程学院，江苏 镇江 212003；
2. 广东石油化工学院 电子信息与计算机学院，广东 茂名 525000；
3. 九江学院 信息科学与技术学院，江西 九江 332005）

摘 要：在语义 Web 环境下，单一和分散的知识库会引发数据的冗余和不一致性，且严重影响知识库智能化查询，降低知识库中数据的重用性和可操作性。文中提出一种基于范畴论的多 TBox 整合方法来整合知识库，旨在解决上述诸类问题。此方法具有高度抽象性和强大表达能力的范畴论是多知识库整合的理想工具。将 TBox 作为操作对象，范畴论中“态射”功能可实现 TBox 的映射，“外推”功能可实现 TBox 的整合，最后通过实验验证方法的可行性和有效性，实验结果基本符合要求。

关键词：知识库；范畴论；TBox 映射；TBox 整合；描述逻辑

中图分类号：TP31　　**文献标识码**：A　　**文章编号**：1673-629X(2011)05-0095-04

Research of Multiple TBox Integration Based on Category Theory

XIA Zhan-feng[1], PENG Zhi-ping[2], XU Du-juan[3]

(1. Department of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China;
2. Department of Computer and Electronic Information, Guangdong University of Petrochemical Technology, Maoming 525000, China;
3. Department of Information Science and Technology, Jiujiang University, Jiujiang 332005, China)

Abstract: In the context of semantic Web, single and dispersed knowledge bases will cause data redundancy and inconsistencies, seriously affect knowledge base intelligent query and reduce data reusing and operability. A method of multiple TBox integration based on category theory is proposed to integrate knowledge bases. Category theory with high abstraction and strong power in expressing is a perfect tool for multiple knowledge bases integration. Taken the TBox as object, the concept of "morphism" achieves TBox mapping and "pushout" achieves TBox integration, at last, the feasibility and effectiveness of the method are verified by experiment, the results basically meet the requirement.

Key words: *knowledge base; category theory; TBox mapping; TBox integration; description logic*

0 引 言

知识库是使语义 Web 具备自主推理和智能效果的关键技术，在语义 Web 实际应用当中，它针对特定领域提供相关的共同认识和知识共享，实现人与机器，机器与机器在语法和语义上准确交流。随着越来越多的知识库被开发利用，研究者渐渐发现，知识库在数据语义化和查询智能化上已经存在极大的瓶颈，从而严重阻碍语义 Web 的发展。

基于描述逻辑[1]（Description Logic，DL）建立起来的知识库由 TBox 和 ABox 两部分构成，其中 TBox 由概念和概念间关系的断言构成；ABox 由特定对象的断言（实例断言）构成。此种知识库虽能有效解决语义 Web 中数据的存储问题，在一定程度上实现了自主推理和智能查询，但是目前单一、分散的知识库并未解决语义 Web 中的“信息孤岛”问题及各知识库数据表示方式和数据结构的不一致性问题。所以知识库整合技术已经成为语义 Web 领域的研究热点。

在基于描述逻辑的知识库整合上，目前国内的研究主要集中在 TBox 的整合，即概念的整合和概念间关系的整合[2]。文中的研究也仅限于对多 TBox 的整合。关于多 TBox 的整合问题，国内研究者绝大部分都是从集合论的角度考虑，然而文中提出的基于范畴论的多 TBox 整合方法比集合论具有更高的抽象性和

收稿日期：2010-09-21；**修回日期**：2010-12-19

基金项目：广东省自然科学基金(8152500002000003)

作者简介：夏战锋(1983-)，男，硕士研究生，研究方向为描述逻辑、语义 Web；彭志平，博士，教授，研究方向为语义 Web、移动 agent 技术、机器学习。

更直观表达力。此法旨在解决当前单一的、分散的知识库给语义 Web 发展带来的阻碍。

1 相关的基本概念

1.1 范畴论

范畴论的主要特点是把每类数学元素的普遍性以及相似性作为研究对象，重点关注各类数学对象之间的内在联系，而不是孤立地单独研究。在计算机学科理论体系研究中，范畴论在程序语言逻辑学、程序语法和语义学以及程序指令等领域有着非常广泛的运用，被看作计算机学科体系中强大的数学工具。

定义 1[3] 范畴(category) C 由以下对象组成：

(1) O：表示一组对象(object)的集合；

(2) M：表示一组态射(morphism)的集合，其中态射 m：$D \to C, D, C \in O$；则称 D 是态射 m 的论域(domain)，C 是态射 m 的余论域(codomain)，记作 $\mathrm{dom}(m) = D, \mathrm{cod}(m) = C$。

定理 1[3] 一个范畴满足以下条件：

1) 复合律：如果 $A, B, C \in O$ 且态射 $m: A \to B; n: B \to C$，那么存在唯一复合态射 $n \circ m: A \to C$，则称 m 与 n 的复合；

2) 结合律：如果 $A, B, C, D \in O$ 且态射 $m: A \to B; n: B \to C; h: C \to D$，那么就有 $(h \circ n) \circ m = h \circ (n \circ m)$；

3) 单位态射：一个对象 A，存有一个单位态射 $IM_A: A \to A$，使对任意态射 $m: A \to B$，有：

$m \circ IM_A = m, IM_B \circ m = m$

范畴是基于图的，显然，一个范畴可看作一个有向图，则定义 1 和定理 1 可形象地表示成图 1 所示。图中的节点表示对象；带箭头的边表示态射；每条边的起始节点和终结点分别表示论域和余论域。

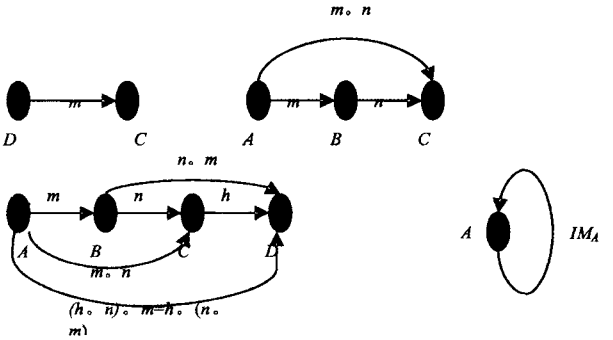

图 1 范畴的图表示

1.2 多 TBox 整合

随着知识库数量的增多并且知识库的构建从来都没有规范的标准，这就导致知识库中数据的冗余和不一致，影响知识库的复用性和可操作性，严重降低语义 Web 的智能化程度以及查询效率。因此，通过整合 TBox 来整合知识库显得非常之必要。TBox 的整合过程主要包括两个步骤：(1) 多 TBox 间的映射；(2) 多 TBox 间的整合。

定义 2 假设任意 $T \in$ TBox 且 T 采用描述逻辑表示，则 T 由(Con, Rel) 构成，其中 Con 是 T 中所有概念集合，Rel 是 T 中所有概念间二元关系的集合。

由定义 2 可知，对 TBox 的整合操作，主要是对其中的概念以及概念间的关系进行操作。

定义 3 多 TBox 间的映射发现是指寻找任意两个 TBox 的元素(概念或者概念间关系)之间的语义相似程度的操作。文献[4]详细地介绍了形式化表示 TBox 间映射函数：

map: $TB_1 \to TB_2$

$\mathrm{map}(d_{1i}) = d_{2j}$，若 $\mathrm{sim}(d_{1i}, d_{2j}) > t$，其中 t 为阈值，$\mathrm{sim}(d_{1i}, d_{2j})$ 表示 d_{1i} 和 d_{2j} 的相似度，则可认为它们在语义上是完全相等的，将 d_{1i} 映射到 d_{2j}。

定义 4 TBox 的整合是建立在多 TBox 映射基础上的，将 $m(m \geqslant 2)$ 个相关的 TBox 统一整合成一新的全局 TBox 的过程。新的全局 TBox 是 m 个 TBox 的并集，不仅包括原 m 个语义相似部分，同时也包括语义不相似的部分。

2 基于范畴论的多 TBox 整合

由 TBox 的定义可知，它实质上是由概念以及概念间关系所构成的有向图，可把此种有向图看成一个 TBox 范畴。TBox 整合的两个关键部分—TBox 映射和 TBox 整合，它们都可利用范畴论的相关特性实现，即将 TBox 的结构看作操作对象，TBox 的映射可用范畴论中“态射”的功能实现，同样，TBox 的整合利用范畴论中“外推”的功能实现。

2.1 多 TBox 间的映射

定义 5 TBox 范畴：把 TBox 的结构看作对象，TBox 间映射看作态射，则 TBox 的范畴 TBc 可定义为态射函数 $(f, g): T \to T'$，其中，$T = (\mathrm{Con}, \mathrm{Rel})$ 和 $T' = (\mathrm{Con}', \mathrm{Rel}')$ 是 TBox 结构。

上述定义中 $f: C \to C'$ 和 $g: R \to R'$ 的态射保持原有的概念以及概念间关系不变。接下来，利用例子详细介绍利用范畴论的特性实现多 TBox 间的映射。

例 1 图 2 所示：TBox TB_1 和 TB_2，其中 $TB_1 = (\{x_0, x_1, x_2, x_3\}, \{r_1, r_2\})$，$TB_2 = (\{y_0, y_1, y_2\}, \{s_1, s_2\})$。根据态射 (m, n)：$\{x_0, x_1, x_2, x_3\} \times \{r_1, r_2\} \to \{y_0, y_1, y_2\} \times \{s_1, s_2\}$，可实现 TBox 间的映射：$x_0 y_0$，$x_1 y_1$，$x_2 y_1$，$x_3 y_2$，$r_1 s_1$，$r_2 s_2$。新得到的映射保持了原有 TBox 的结构。比如，在 T_1 中，x_1 和 x_2 是 x_0 的子概念，则映射到 T_2 中的 $f(x_1), f(x_2) = y_1$ 是 $f(x_0) = y_0$ 的子概念，同样，T_1 中的 $r_1(x_0, x_3)$ 和 T_2 中的 $g(r_1)(f(x_0), f(x_3)) =$

(s_1, y_0, y_2) 形成映射。

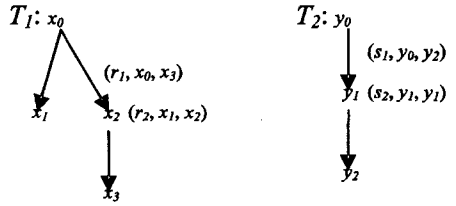

图 2　TBox 映射

2.2　多 TBox 间的整合

外推的在范畴论中的功能是形成融合和(amalgamated sum)[4],文中利用范畴论中外推的性质实现多 TBox 的整合。

定义 6 TBox 的外推:假设有 TB_1、TB_2 和 TB 以及 TBox 态射 $(m_1, n_1): TB \to TB_1$ 和 $(m_2, n_2): TB \to TB_2$,如图3所示,那么称 TB' 和态射 $(m_1, n_1)': TB_1 \to TB'$ 以及 $(m_2, n_2)': TB_2 \to TB'$ 具备 TB_1 和 TB_2 的外推,且满足:(1) $(m_1, n_1)' \circ (m_1, n_1) = (m_2, n_2)' \circ (m_2, n_2)$;(2) 对其它任意 TBox TB'' 的态射 $(m_1, n_1)'': TB_1 \to TB''$ 以及 $(m_2, n_2)'': TB_2 \to TB''$,唯一存在态射 $(m, n): TB' \to TB''$ 使 $(m, n) \circ (m_1, n_1)' = (m_1, n_1)''$, $(m, n) \circ (m_2, n_2)' = (m_2, n_2)''$。

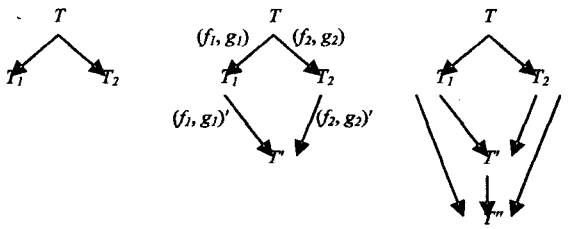

图 3　TBox 外推

基于 TBox 的外推定义以及 TBox 间映射的基础上,下面,文中给出多 TBox 整合算法。设 $TB_i = (C_i, R_i)$(其中 $i = 1, 2$),$TB' = (C'_i, R'_i)$ 分别是 TB_1 和 TB_2 的外推,即需要整合的 TBox,$TB = (C, R)$ 表示的是 TB_1 和 TB_2 的语义交集。初始状态 $TB' = TB_\varphi = (\varphi, \varphi, \perp)$ 表示一个空 TBox,其中 φ 是空集合;$\perp$ 表示函数(此处指 m'_i, n'_i)在所有域内无定义。多 TBox 的外推算法描述如下:

输入:$(f_1, g_1): T \to T_1$, $(f_2, g_2): T \to T_2$

输出:$(f'_1, g'_1): T_1 \to T'$ 和 $(f'_2, g'_2): T_2 \to T'$

初始条件:$T' = T_\varphi$, $f'_i = \perp$, $g'_i = \perp$

(1)/* 针对 C 中每一个 c 在 C' 中相应地添加一新的 c',并定义态射 f'_i,即 $f_i(c)$ 与新的 c' 之间建立映射 */

While(all $c \in C$)

{if($c' \notin C_1 \cup C_2$))

{$C' = C' \cup c'$;

}

$f'_1 = f'_1 \cup (f_1(c)\, c')$;/* $f_1(c)$ 与 c' 之间建立映射 */

$f'_2 = f'_2 \cup (f_2(c) c')$; /* $f_2(c)$ 与 c' 之间建立映射 */

}

(2)/* 将 C_1 中其余的概念也添加到 C' 中 */

While(all $c \in C_1$ 不满足 f_1)

{ $C' = C' \cup c'$;

$f'_1 = f'_1 \cup (f_1(c) c')$;

}

(3)/* 将 C_2 中其余的概念也添加到 C' 中 */

While(all $c \in C_2$ 不满足 f_2)

{ $C' = C' \cup c'$;

$f'_2 = f'_2 \cup (f_1(c)\ c')$;

}

定义并实现 R' 和 r(概念间关系)的外推方法同上,故不再重复阐述。

3　消除数据冗余和不一致性

知识库整合的目的是消除同领域中各知识库间的数据冗余和不一致性,以便用户可共享其他用户所提供的数据[5]。此工作主要通过整合领域内各部门知识库中的 TBox 来完成。在面向大型企业的供应链管理,整合工作是在局部 TBox 的基础上构造全局 TBox,而局部的 TBox 依然独立存在。供应链中的全局 TBox 要求消除各环节 TBox 的数据冗余和不一致性,使企业各部门通过全局 TBox 可准确检索信息。

整合工作主要分三个过程:

(1) 寻找 TBox 间的重叠区域;

(2) 建立概念关联;

(3) 消除数据的冗余性和不一致性。

过程 (1)、(2)在文献[6~8]中有详细阐述且所提方法被绝大部分研究者所认可,因此文中不再重复。过程(3)易被研究者所忽略且目前讨论它的参考文献非常之少。因此,文中重点讨论过程(3)。

供应链中的局部 TBox 经过过程(1)和(2),得到初步整合后的全局 TBox,必存有一些冗余或不一致的数据。这要求对被整合后的内容进行检查,消除冗余和不一致性,对其作进一步整合。

I) 消除冗余数据。如果被整合后的某概念 A 与同一 TBox 中的概念 B 和 C 存在 $A \subseteq B$, $A \subseteq C$ 并且 $C \subseteq B$ 的关系,则需消除概念 A 与 B、C 其中一个包含关系,如图 4 所示。

图 4 描述了包含数控机床的 TBox 的整合,经过过程(1)和(2)后,得到整合后 TBox 中数控机床的结构如图左所示,经系统进一步提出建议,将"操纵杆"和"操作杆"整合。结果如图右所示,然而发现,对于概

念“操纵杆”、“数控机床”、“主控台”三者存在相互包含关系，此时，需删除被整合概念“操纵杆”与其中一个概念间的包含关系，一般而言，建议删除“数控机床”与“操纵杆”间的包含关系。

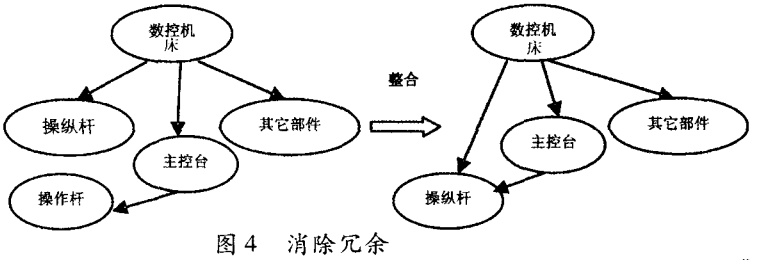

图 4 消除冗余

Ⅱ）消除数据不一致性。多 TBox 整合后，数据不一致性主要存在逻辑表达、原始语义、语言描述等方面的不一致[9,10]。对于逻辑表达而言，目前大部分领域知识库专家采用 DL 来构建知识库，但个别专家也会采用其它逻辑来构建知识库。对此，只需设置一定的规则系统来整合 DL 与其它逻辑，就能有效解决。对原始语义而言，比上述逻辑表达的不一致性要相对复杂些，但现有的 FaCT 系统和 RACER 系统按照 DL 系统提供的一致性检测功能专门设计了解决语义不一致性的优化推理算法[11,12]。对语言描述而言，某一种语言描述的事物可能在另一种语言中无法描述，比如，有些语言不具有否定功能。对此，需修改语法规则系统，扩充语言表达功能。

4 实验验证与结果分析

（1）实验数据。

为了验证上述整合方法的可行性和有效性，文中选取 LUBM[6] 数据集中的部分数据并分成三组：第 1 组包含 400 条术语；第 2 组包含 600 条术语；第 3 组包含 800 条术语。在这 3 组 TBox 基础上构建 6 个查询。

（2）实验条件。

测试环境：软件为 Windows XP SP2+Eclipse 7.0+JDK 1.6.0+PELLET 推理机+Jena 推理机；硬件为 Intel Pentium-Dual 2.0 GHz CUP，2.0GB 的 DDR 内存。

（3）实验结果与分析。

针对上述三组 TBox 和六个查询进行实验仿真，得出系统实际运行时间结果如图 5 所示。

实验结果表明，文中提出的基于范畴论的多 TBox 整合技术可以用来整合知识库，整合后的知识库能够提供快速有效的检索服务，并且检索的结果基本符合要求。

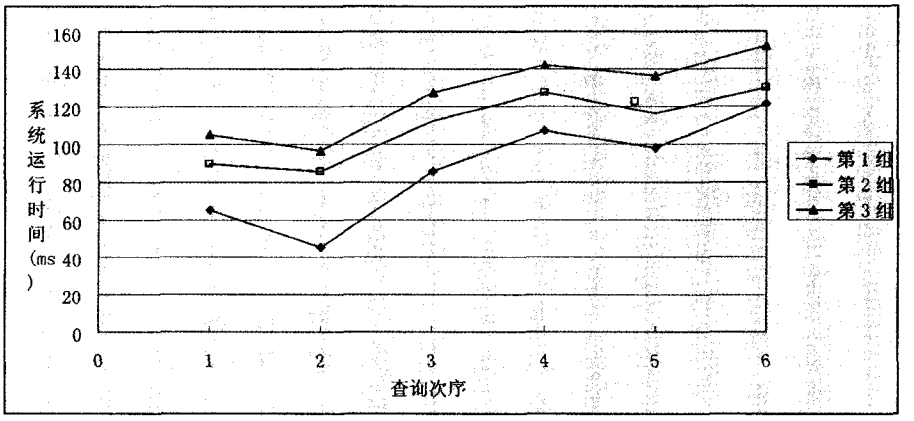

图 5 系统实际运行时间

5 结束语

当前，国内对多知识库的整合研究都是基于集合论，文中从范畴论的角度弥补了集合论的局限性，利用范畴论的“态射”和“外推”技术对多 TBox 进行有效整合。范畴论具有很好的数学理论基础，合理利用它的“态射”理论，可在一定程度上实现多 TBox 的自动映射。

下一步工作主要把范畴论用在多知识库的演化、知识库更新，尤其是实现知识库自主推理方面；进一步完善“态射”理论和“外推”理论，以更好地保证知识库整合的准确性以及有效性。

参考文献：

[1] Baader F, Calvanese D, McGuinness D L, et al. The Description Logic Handbook: Theory, Implementation, and Applications[M]. 2nd Ed. Cambridge: Cambridge University Press, 2007.

[2] 端义锋，胡谷雨，潘志松. 一种基于解释的知识库整合[J]. 电子科技大学学报，2008，4(3)：366-368.

[3] Barr M, Wells C. Category Theory for Computing Science [M]. [s. l.]: Prentice Hall, 1990.

[4] Ehrig M, Staab S. QOM - Quick Ontology Mapping[C]// Proc. of International Semantic Web Conference. Hiroshima, Japan: Springer, 2004.

[5] Alferes J J, Knorr M, Swift T. Queries to Hybrid MKNF Knowledge Bases through Oracular Tabling[C]//Semantic Web Conference. [s. l.]: [s. n.], 2009: 1-6.

[6] Guo Y. Lehigh University Benchmark (LUBM) [EB/OL]. 2006. http:// swat. cse. lehigh. edu/project/lubm.

[7] 唐 杰，梁邦勇，李涓子，等. 语义 Web 中的本体自动映射[J]. 计算机学报，2006(11)：1956-1974.

[8] 许文艳，刘三阳. 知识库系统的逻辑基础[J]. 计算机学报，2009(11)：2123-2126.

[9] 李 未. 世纪之交的知识

（下转第 102 页）

间中的相距很近的点投影到低维空间时所得到的像也是相近的，在此过程中应用到了 Laplacian-Beltrami 算子来保证数据的正确性。但是当 LLE 和 LE 这两种算法在参数选择不当的时候会使全局结构发生变化，不能有效地反应数据之间的映射关系。

另外，Isomap、LLE 和 LE 这三种算法中对于参数的选择也具有一定的困难性和复杂性，因为存在着噪声敏感问题。如果选取的数据集存在较大噪声，那么在算法执行过程中很难突出它们的内在结构。

其次，在进行邻域大小的选取时，若领域值过大则会使原数据的局部信息丢失，若邻域值过小则会导致原来连续的流形分化为若干个子流形，也就是说会出现“空洞”现象[17]，从而导致算法失效。因此在取值的时候要重点考虑噪声、几何结构、数据的取样密度等因素。

最后，从时间和空间上来分析，LLE 算法和 LE 算法用到了稀疏矩阵，这样就节省了计算时间，而 Isomap 算法则是通过图的连接来得到测地距离，增加了计算的复杂性。因此，相比之下，Isomap 算法是一种计算时间代价最高的算法。

4 结束语

流形学习是一个相对复杂的概念，它与拓扑学、微分几何学等数学分支有着紧密的联系。流形学习的研究和应用范围越来越广泛，流形学习算法主要用于维数约简，其中还有一些算法用于图像分析、信息检索、特征认证等多方面的研究[18]。在这些领域如果采用传统的降维方法很有可能得不到想要的效果，而如果采用流形学习算法不但可以进行有效的降维，而且还可以探索出嵌入在高维空间中的低维流形中数据的内在规律。

文中简单介绍了几种流形学习的算法，并对它们各自的特点进行了分析。尽管在过去的几年里流形学习的算法研究取得了很好的成果，但是由于其涉及到的数学分支相对复杂，嵌入在高维数据中的低维流形依然存在着很多值得去深入研究的问题。从另一个角度来说这也意味着流形学习具有更加广泛的应用范围。

参考文献：

[1] 罗四维，赵连伟. 基于谱图理论的流形学习算法[J]. 计算机研究与发展，2006，43(7)：1174-1179.

[2] 徐 蓉，姜 峰，姚鸿勋. 流形学习概述[J]. 智能系统学报，2006，3(1)：45-51.

[3] ZHANG JUNPING, WANG LIJUE. Manifold learning and Applications in Recognition[C] // Intelligent Multimedia Processing with Soft Computing. Heidelberg: Springer-Verlag, 2004.

[4] HE XIAOFEI, YAN SHUICHENG, HU YUXIAO. Face Recognition UsingLaplacianfaces[J]. IEEE, 2005, 27(3): 328-340.

[5] 赵连伟，罗四维，赵艳敞，等. 高维数据流形的低维嵌入及嵌入维数研究[J]. 软件学报，2005，16(8)：1423-1430.

[6] 李小丽，薛清福. 几种流形学习算法的比较研究[J]. 电脑与信息技术，2009，17(3)：14-18.

[7] 詹德川，周志华. 基于集成的流形可视化[J]. 计算机研究与发展，2005，42(9)：1533-1537.

[8] 王自强，钱 旭，孔 敏. 流形学习算法综述[J]. 计算机工程与应用，2008，44(8)：9-12.

[9] 高小方. 流形学习方法中的若干问题分析[J]. 计算机科学，2009，36(4)：25-28.

[10] Zhang Z Y, Zha H Y. Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment [J]. SIAM Journal of Scientific Computing, 2004, 26(1): 313-338.

[11] 何 力，张军平，周志华. 基于放大因子和延伸方向研究流形学习算法[J]. 计算机学报，2005，28(12)：2000-2009.

[12] 解洪胜，王连国. 流形学习中三种非线性降维算法的比较研究[J]. 云南民族大学学报：自然科学版，2009，18(2)：151-156.

[13] 王泽杰. 两类非线性降维流形学习方法的比较分析[J]. 上海工程技术大学学报，2008，22(1)：54-59.

[14] Chang H, Yeung D Y. Robust locally linear embedding[J]. Pattern Recognition, 2006, 39(6): 1053-1065.

[15] 文贵华，江丽君，文 军. 邻域参数动态变化的局部线性嵌入[J]. 软件学报，2008，19(7)：1666-1673.

[16] 周 波. 两种基于谱方法的流形学习算法研究[J]. 云南民族大学学报：自然科学版，2008，17(4)：370-373.

[17] 张振跃，查宏远. 非线性低次逼近与非线性降维[J]. 中国科学 A 辑：数学，2005，35(3)：273-285.

[18] 严峻松，肖 健，周宗潭，等. 非线性流形学习方法的分析与应用[J]. 自然科学发展，2007，17(8)：1015-1025.

(上接第 98 页)

工程与知识科学[M]. 北京：清华大学出版社，2000：5-6.

[10] 张灵峰，夏战锋，彭志平. 基于 TBox 和 ABox 的描述逻辑推理研究[J]. 计算机技术与发展，2010，20(11)：45-48.

[11] 诸葛海. 语义网格的基础理论、模型与方法研究进展[J]. 中国基础科学，2007(6)：27-29.

[12] Giacomo G D. Lenzerini M. Tbox and Abox Reasoning in Expressive Description Logics[C]//Proceedings of the 5th International Conference on Principles of Knowledge Representation and Reasoning. Roma: AAAI Press, 1996: 316-327.