

基于描述逻辑的组合知识库推理^{*}

卢道设^{1,2}, 杨世瀚^{1,2†}, 吴尽昭^{1,2}, 熊玲芳^{1,2}

(1. 广西民族大学 信息科学与工程学院, 南宁 530006; 2. 广西混杂计算与集成电路设计分析重点实验室, 南宁 530006)

摘要: 由于知识库的定义不同, 相近领域的知识库不能相互利用已有知识进行推理。为了联合不同的知识库进行推理, 通过对描述逻辑的表现形式进行扩展提出一种组合描述逻辑, 并基于概念的相似性将不同的领域概念进行关联, 给出组合描述逻辑的语法及语义以及相应的 Tableau 算法。通过实例表明, 组合描述逻辑可以利用已有知识进行推理; 组合描述逻辑可以将不同的知识库进行结合, 为借用不同知识库的知识进行推理提供一条新的途径。

关键词: 描述逻辑; 推理; 相似; 跨领域; 知识库

中图分类号: TP18

文献标志码: A

文章编号: 1001-3695(2012)12-4503-04

doi:10.3969/j.issn.1001-3695.2012.12.026

Interdisciplinary reasoning on description logic

LU Dao-she^{1,2}, YANG Shi-han^{1,2†}, WU Jin-zhao^{1,2}, XIONG Ling-fang^{1,2}

(1. School of Information Science & Engineering, Guangxi University for Nationalities, Nanning 530006, China; 2. Guangxi Key Laboratory of Hybrid Computational & IC Design Analysis, Nanning 530006, China)

Abstract: In order to reasoning on interdisciplinarily with existed knowledge bases (KBs), this paper proposed a combined description logics based knowledge representative, relating to different fields by the similarity of concepts. First it gave the syntax and semantics of the combine description logics, then presented a Tableau algorithm for the logics' grammar. The experiment shows that the combine description logic can reason on inter-disciplinary in KBs. With the combined formalism, it can integrate the different KBs so that providing a new support for interdisciplinary reasoning.

Key words: description logic; reasoning; similar; interdisciplinary; knowledge base (KB)

0 引言

随着语义网^[1]的发展, 描述逻辑受到越来越多的关注。除了其本身的知识表达能力很强之外, 关键是它提供了可判定的推理服务, 并且目前已有可靠完备的推理算法。

当前许多领域都在开发基于描述逻辑的知识库。但即使是相邻甚至是同一领域, 由于使用的本体^[2]的不同, 也会构造出不同的知识库, 这样就造成了很多内在的知识在表示上不可推理, 且重用难度增大。例如城市公共交通领域的公交站 A 与地铁轨道交通领域的地铁站 B 距离很相近, 人们一般将它们当做是同一个站点, 然而在各自的领域当中所定义的概念的不同造成“A 与 B 是同一站点”这一类知识无法被各自知识库所使用。为了解决重用问题以及借用相近领域的知识进行合理的推理, 目前主要的处理方法是进行结构类似映射、重新抽象领域词汇以及本体映射^[1,3]。这类方法在使用的时候都要对词汇和结构进行详细的分析, 需要耗费相当多的时间与精力。本文的研究是基于不同的知识库, 通过提出一种扩展形式将不同领域的知识库进行结合, 并借用上述例子中的距离概

念, 通过个体的空间距离合理定义, 把不同知识库间的知识联系起来, 从而把潜在的知识挖掘出来。

1 描述逻辑

描述逻辑 (description logic, DL) 是一种知识表示的形式化语言^[4,5], 也称为术语逻辑。它建立在概念 (concept) 与关系 (role) 之上, 是一阶逻辑的可判定子集, 在具有适当定义的语义的同时具有很强的表达能力。描述逻辑一般由四部分组成: 概念构造符集合 Σ 、概念及关系集合、实例断言集合及推理机制。不同的描述逻辑语言的区分取决于概念构造符的选择, 概念构造符的不同决定了表达能力的强弱。一个基于描述逻辑的知识库一般包含 TBox (terminology box) 与 ABox (assertion box) 两方面。TBox 是概念及关系集合, ABox 是实例集合。描述逻辑的推理主要分为基于概念及关系的包含检测推理和基于实例断言的一致性检测推理。

1.1 基本的描述逻辑语法及语义

ALC 是描述逻辑语言中最基本的描述语言, 其他描述语

收稿日期: 2012-04-07; **修回日期:** 2012-05-21 **基金项目:** 国家自然科学基金资助项目 (60973147); 广西自然科学基金资助项目 (2011GXNSFA018154); 广西教育厅科研资助项目 (201012MS274); 中国科学院 2010 年度“西部之光”人才培养计划资助项目; 广西自然科学基金创新研究团队资助项目 (2012JJGAG0002); 广西混杂计算与 IC 设计分析重点实验室开放基金资助项目 (HCIC201110)

作者简介: 卢道设 (1987-), 男, 广东阳江人, 硕士研究生, 主要研究方向为描述逻辑、可满足性模理论; 杨世瀚 (1972-), 男 (通信作者), 讲师, 博士, 主要研究方向为形式化方法、软件工程 (dr.yangsh@gmail.com); 吴尽昭 (1965-), 男, 教授, 博导, 主要研究方向为形式化方法、自动推理、验证技术; 熊玲芳, 女, 硕士研究生, 主要研究方向为描述逻辑、可满足性模理论。

言都是在 ALC 基础上扩展而来的,并以概念构造符的不同加以区分。ALC 的概念构造符集合 $\Sigma = \{\neg, \sqcap, \sqcup, \exists, \forall\}$ 。以下给出 ALC 语法的定义。

定义 1 设 N_C 是原子概念的集合, N_R 是原子关系的集合, 则 ALC 概念由如下规则确定:

- 如果 $A \in N_C$, 则 A 是概念;
- 如果 A 是概念, 则 $\neg A$ 也是概念;
- 如果 A, B 是概念, 则 $A \sqcap B, A \sqcup B$ 也是概念;
- 设 $R \in N_R, C$ 是概念, 则 $\exists R. C, \forall R. C$ 是概念。

其中, N_C 包括全域概念(顶概念) $\top$ 与底概念 $\perp$ 。

由描述逻辑的概念及概念间的结构关系组成的集合称为 TBox, 而描述关于具体个体事实的公理集称为 ABox。假设个体名为 a, b, c, C 是概念, R 是关系, 则 ABox 的断言形如 $C(a), R(b, c)$ 。

定义 2 解释 $I = (\Delta^I, \cdot^I)$ 由非空集合 Δ^I 和解释函数 $\cdot^I$ 组成。称 Δ^I 为 I 的论域(包括所讨论领域中的所有个体), 解释函数 $\cdot^I$ 将概念映射到 Δ^I 中的子集, 将关系映射到 $\Delta^I \times \Delta^I$ 的子集。

定义 3 解释 I 满足 Tbox T , 当且仅当 I 满足 T 中所有的概念定义及公理, 则称 I 是 T 的一个模型。

定义 4 解释 I 满足 ABox A 中的每一个断言, 那么 I 满足 A , 则称 I 是 A 的一个模型。

由前面的定义可以得出 ALC 语法及语义, 如表 1 所示。

表 1 ALC 的语法及语义表

概念构造算子	语法	语义
原子概念	A	$A^I \subseteq \Delta^I$
原子关系	R	$R^I \subseteq \Delta^I \times \Delta^I$
概念的交	$C \sqcap D$	$C^I \cap D^I$
概念的并	$C \sqcup D$	$C^I \cup D^I$
否定	$\neg C$	$\Delta^I \setminus C^I$
存在量词	$\exists R. C$	$\{x \mid \exists y. (x, y) \in R^I \wedge y \in C^I\}$
值限定	$\forall R. C$	$\{x \mid \forall y. (x, y) \in R^I \rightarrow y \in C^I\}$
顶概念	$\top$	Δ^I
底概念	$\perp$	$\emptyset$

1.2 描述逻辑推理

推理机制是描述逻辑构成的主要成分之一。一般基于描述逻辑的知识库就 TBox 而言有概念包含检测的可满足性推理, 就 ABox 而言有一致性检测推理^[5]。

可满足性指的是概念 C 关于 TBox T , 若存在 T 的一个模型 I , 满足 $C^I \neq \emptyset$, 称 C 是关于 T 可满足的。可满足性问题是描述逻辑推理的核心问题, 其他问题诸如基于概念的包含检测、基于实例的一致性检测问题都可以转换为可满足性问题。当前大多数描述逻辑语言都使用 Tableaux 算法来检测概念的可满足性^[6]。

为检测概念 C 的可满足性, Tableau 算法约束系统^[7,8]从 $S(x) = \{C \mid (x \text{ 是概念 } C \text{ 的实例}, S(x) \text{ 是关于 } x \text{ 的概念集合})$ 开始, 并通过如表 2 所示的规则进行概念扩展演算, 构造出概念 C 的模型 I 。若模型 I 可以被构造出来, 则称 C 是可满足的, 否则称 C 是不可满足的。

定义 5 若 x 是 ABox 中的具体的个体, $L(x)$ 是 x 关于 TBox T 的概念集合, 若存在概念 C , 有 $\{C, \neg C\} \subseteq L(x)$, 则称 $L(x)$ 包含一个冲突。

定义 6 如果在构造模型 I 的过程中, 出现了冲突或是没有规则(如表 2 所示)可以再应用, 则称此模型 I 是完全的。如

果能构造出概念 C 的完全无冲突的模型 I , 则称概念 C 是可满足的, 否则是不可满足的。

表 2 Tableau 算法规则表

规则名称	规则定义
$\sqcap$ 规则	如果 $(1) (C_1 \sqcap C_2) \in S(x); (2) \{C_1, C_2\} \not\subseteq S(x)$; 那么 $S(x) = S(x) \cup \{C_1, C_2\}$
$\sqcup$ 规则	如果 $(1) C_1 \sqcup C_2 \in S(x); (2) \{C_1, C_2\} \cap S(x) = \emptyset$; 那么 $S(x) = S(x) \cup \{C_1\}$ 或 $S(x) = S(x) \cup \{C_2\}$
$\exists$ 规则	如果 $(1) (\exists R. C_1) \in S(x); (2)$ 没有实例 z , 使得 $R(x, z)$ 成立并且 $C_1 \in L(z)$; 那么新增实例 z , 使得 $R(x, z)$ 成立并且 $L(z) = \{C_1\}$
$\forall$ 规则	如果 $(1) (\forall R. C_1) \in S(x); (2)$ 存在实例 y 使得 $R(x, y)$ 成立且 $C_1 \in L(y)$; 那么 $L(y) = L(y) \cup \{C_1\}$

2 组合描述逻辑

描述逻辑作为知识表示的形式化基础, 具有很强的知识表示能力和推理能力。但不同的知识库因为定义了不同的 TBox, 在现实社会领域中对于相同或相近对象事物在表示上不能很好地相互认知, 且不能推理出潜在的知识, 这就约束了描述逻辑对现实世界的推理能力。为了达到借用知识库间的知识进行领域间的知识库推理, 本文对描述逻辑的知识表示形式进行了扩展, 通过对两个不同知识库作语义上的积来构建一种新的知识表示的形式, 称为组合描述逻辑。

知识库包含 TBox 和 ABox 两个部分。因为 TBox 是概念集和关系集的集合, 因此记知识库 DL 为三元组^[9] $DL = (N_C, N_R, A)$, 其中 N_C 表示概念集, N_R 表示关系集, A 表示个体公理集。

假设存在两个不同的基于 ALC 的知识库 $DL_1 = (N_{C_1}, N_{R_1}, A_1)$ 和 $DL_2 = (N_{C_2}, N_{R_2}, A_2)$, 其中 N_{C_1}, N_{R_1}, A_1 和 N_{C_2}, N_{R_2}, A_2 分别为 DL_1 和 DL_2 的原子概念集、原子关系集、断言集。

定义 7 称 $DL^2 = DL_1 \times DL_2$ 为 DL_1 和 DL_2 的组合知识库, $DL^2 = (N_C, N_R, A^2)$, 那么 DL^2 的 N_C, N_R, A^2 满足以下条件:

- $N_C = \{(A, C) \mid A \in N_{C_1} \wedge C \in N_{C_2}\}$ 是概念的集合, 称为 DL^2 的原子概念集合, 也称为组合概念;
- 若 $(A, C) \in N_C, A(a) \in A_1$ 并且 $C(c) \in A_2$, 那么 $(A, C)(a, c) \in A^2$;
- 若 $R_1 \in N_{R_1}, R_1(a, b) \in A_1$ 且个体 $((a, x), (b, y)) \in A^2$ 或 $R_2 \in N_{R_2}, R_2(c, d) \in A_2$ 且个体 $((x, c), (y, d)) \in A^2$, 则存在组合关系 $R' \in N_R$, 使得 $R'((a, x), (b, y)) \in A^2$ 或 $R' \in N_R$ 使得 $R'((x, c), (y, d)) \in A^2$ 。为了方便, 对于 $R \in N_R$, 可记为 (R_1, R_2) , 其中 $R_1 \in N_{R_1}$ 或 $R_2 \in N_{R_2}$ 。

定义 8 设 N_C 是 DL^2 原子概念的集合, N_R 是原子关系的集合, 则 DL^2 组合概念由如下规则确定:

- 若 $(A, C) \in N_C$, 则 (A, C) 是组合概念;
- 若 (A, C) 是组合概念, 则 $\neg(A, C)$ 是组合概念;
- 若 $(A, C), (B, D)$ 是组合概念, 则 $(A, C) \sqcap (B, D), (A, C) \sqcup (B, D)$ 是组合概念;
- 设 $R \in N_R, (B, D)$ 是组合概念, 则 $\exists R. (B, D), \forall R. (B, D)$ 是组合概念。

定义 9 DL^2 的解释 $I = (\Delta^I, \cdot^I)$ 由非空集合 $\Delta^I = \Delta_1^I \cup \Delta_2^I$ 和解释函数 $\cdot^I$ 组成。其中 I_1, I_2 分别是 DL_1, DL_2 的一个解释。

描述逻辑存在语法与语义解释的对应, 本文也给出 DL^2 组合概念的语法及语义解释, 如表 3 所示。

表 3 DL² 组合概念的语法及语义表

概念构造算子	语法	语义
原子概念	(A, C)	$(A, C)^I \subseteq \Delta^I$
原子关系	R	$R^I \subseteq \Delta^I \times \Delta^I$
概念的交	$(A, C) \sqcap (B, D)$	$((A, C) \sqcap (B, D))^I = (A, C)^I \cap (B, D)^I$
概念的并	$(A, C) \sqcup (B, D)$	$((A, C) \sqcup (B, D))^I = (A, C)^I \cup (B, D)^I$
否定	$\neg(A, C)$	$\Delta^I \setminus (A, C)^I$
存在量词	$\exists R. (B, D)$	$\{(x_1, y_1) \mid \exists (x_2, y_2). ((x_1, y_1), (x_2, y_2)) \in R^I \wedge (x_2, y_2) \in (B, D)^I\}$
值限定	$\forall R. (B, D)$	$\{(x_1, y_1) \mid \forall (x_2, y_2). ((x_1, y_1), (x_2, y_2)) \in R^I \rightarrow (x_2, y_2) \in (B, D)^I\}$
顶概念	$\top^2$	Δ^I
底概念	$\perp^2$	$\emptyset$

3 组合描述逻辑的推理

要将一个领域范围的知识关联另外领域的知识,就要建立不同领域概念及关系间的相关性,通过个体把这种相关性联系起来是切合实际的方式。

定义 10 对于 $C \in N_{C_1}, D \in N_{C_2}$ 且有 $c \in C^{I_1}, d \in D^{I_2}$, 定义函数 $f: N_{C_1} \times N_{C_2} \rightarrow R$ 有 $f(c, d) = \text{distance}(c, d) \leq n$, 其中实数 $n \in R$ 为空间距离。若给定距离 n , 个体 c 与 d 满足 f , 则称 c 与 d 是基于距离相近的, 简称相近; 同时称概念 C 与 D 是基于距离相似的, 简称相似。

定义 11 对于 $R_1 \in N_{R_1}, R_2 \in N_{R_2}$, 有 $R_1(c, d)$ 与 $R_2(c', d')$, 其中 $c \in C_1^{I_1}$ 与 $c' \in C_2^{I_2}, d \in D_1^{I_1}$ 与 $d' \in D_2^{I_2}$, 若概念 D_1 与 D_2 是相似的, 并且 C_1 与 C_2 是相似的, 称 R_1 与 R_2 是相似的。

组合知识库把两个不同的知识库关联起来, 通过概念、关系的有序构成组成新的概念对和新的关系。这种组合在表现形式上既包含了原来各自领域的知识表示, 也包含了领域间隐含的知识, 经过适当的语法及语义定义, 就形式表示而言, 这种组合形式具备一定的领域间知识推理服务能力。

3.1 一般概念向组合概念的扩展

一般的描述逻辑推理使用 Tableau 算法检测概念的可满足性。通过前面的介绍可知, Tableau 是通过概念构造出满足概念的模型从而判断概念是否满足。那么, 对于一般 DL 进行检测的概念形如 $(\exists R. A) \sqcap (\exists R. B) \sqcap \neg(\exists R. (A \sqcap B))$ 。为了在组合知识库上进行推理获取更多的知识, 对这些概念进行推理时概念扩展。要进行扩展的概念及其领域的概念称为原概念, 与原概念组合成组合概念的另一领域的概念称为相似概念。那么有如下扩展规则:

a) 扩展组合表示的概念时, DL 概念必须符合否定标准形式(NNF), 扩展后也是否定标准形式。

b) 对一般概念向组合概念进行扩展时, 概念构造符不变。

c) 对于关系 R 的扩展, 若存在相似关系 R' , 则选取相似关系, 扩展成组合关系, 如 $((a, x), (b, y))$, 其中有 $R(a, b), R'(x, y)$; 否则随机选取。

d) 对于概念 A 的扩展, 若存在相似概念 A' , 则选取相似概念, 扩展成组合形式概念, 如 (A, A') , 否则随机选取。

例如 DL 下概念 $(\exists R. A) \sqcap (\exists R. B) \sqcap \neg(\exists R. (A \sqcap B))$, 扩展成组合形式后的概念形如 $(\exists (R, R'). (A, A')) \sqcap (\exists (R, R'). (B, B')) \sqcap (\forall (R, R'). \neg(A, A') \sqcup (B, B'))$ 。

3.2 推理

组合描述逻辑的推理目标是利用不同知识库里面的知识

进行推理。假若在某领域存在 $R(a, b)$, 而在另一领域存在 $R'(c, d)$, 而 c 与 b 满足函数 f , 那么是否存在一个关系从 a 到 c 呢? 显然, 在各自的知识库下得出的结论是不存在的, 只有在组合知识库下才可以得出存在的结论。在组合关系中使用“or”进行组合, 在保持各自的知识库的现有关系的知识不丢失的同时也造成了很多冗余知识的出现, 如 $R((a, b), (c, d))$ 中 a 与 c 的概念不满足相似性, 对它们进行推理一般是没有任何意义的。为了减少这样的冗余知识的影响保证推理有意义, 在推理的时候对于组合关系 R 限定为是互为相似的关系 R_1 与 R_1' 。

组合推理是在原 Tableau 算法上进行改进, 以适应所引入的相似概念和相似关系。那么一般的推理算法就是在知识库的个体集合中找到满足相应概念的个体的存在来判断是否满足。算法的步骤与一般的 ALC Tableau 一样, 主要有 $\sqcap$ 规则、 $\sqcup$ 规则、 $\exists$ 规则、 $\forall$ 规则。对于 $\sqcap$ 规则、 $\sqcup$ 规则和 $\forall$ 规则, 组合推理的 Tableau 的意义与 ALC Tableau 一致; 但对于 $\exists$ 规则, 主要有以下两种情况需要考虑:

a) 怎样处理不是由互为相似关系所组合的组合关系, 对于本文的推理有无意义, 算法是否应该停止;

b) 在互为相似关系的组合关系当中, 怎样处理 R_1 或 R_1' 在当前个体中并不实际存在。

对于问题 a), 若组合关系不是由互为相似的关系所组合, 那么不同知识库间的知识无法传递下去, 相似概念必定是不满足的, 算法应该停止。对于问题 b), 在 (R_1, R_1') 组合关系当中, 若 R_1 在当前的推理的上下文中出现而 $R_1(a, b)$ 在原来知识库中并不存在, 那么在原来知识库中的推理也是不满足的, 所以在组合知识库的推理当中, 它应当停止; 对于 R_1' 也是同样的道理。

3.2.1 组合推理 Tableau

一般的 ALC 通过 Tableau 构造出概念的模型树, 模型的个体对应树的节点, 节点上的标记是一组 ALC 概念集合, 树的边的标记是某个 ALC 关系名。本文的组合推理也是构造模型树, 不同的是组合形式的个体对应树的节点, 组合形式的关系名对应树的边。

与一般的 Tableau 不同, 本文在每个节点的标注上使用两个标注如 $L_1((x, y)), L_2((x, y))$, 其中 $L_1((x, y))$ 代表组合形式的概念集以及原概念集合, 称为原概念集标注; $L_2((x, y))$ 标注代表组合形式的概念集以及相似概念集合, 称为相似概念集标注。

原概念集标注 $L_1((x, y))$ 用来验证在无须利用其他知识库知识进行推理的时候能否满足原概念, 而相似概念集标注 $L_2((x, y))$ 表示的是原知识库需要借用其他知识库知识进行推理是否满足概念。当组合关系 R 是由原知识库关系 R_1 及其相似关系 R_1' 组合的时候, 若 R_1 在原知识库不使个体的关系成立, 则在原知识库中是不满足的, 此时原概念集应该停止, 但相似概念集可能会满足, 所以算法过程不停止, 直至相似集也不满足为止。

定义 12 对于概念集 $L((x, y))$, 如果存在概念 $\{(A, C), \neg(A, C)\} \subseteq L((x, y))$ 或 $\{A, \neg A\} \subseteq L((x, y))$, 称为存在冲突; 如果原概念集和相似概念集都存在冲突, 则称为完全冲突; 只有其一存在冲突, 称为部分冲突。

对于是原概念集冲突的非完全冲突的模型树, 称概念是相

似可满足的;称相似概念集冲突的非完全冲突的模型树为原概念可满足的;如果是完全冲突的模型树,称概念是不可满足的。

在构造模型树中,本文通过应用算法规则(如表4所示)不断地扩展组合概念,如果 $L_1((x,y))$ 或 $L_2((x,y))$ 的标注集中出现了冲突,那么对应的概念集构造过程停止,否则,直到应用完所有的规则为止。因此,下面给出进行推理的算法描述:

a) 根据函数 f 计算出相似概念集和相似关系集。

b) 初始化并构造组合知识库,构造出原子概念集 N_C 、组合关系 N_R 和断言集 A^2 。

c) 将待证的概念 D 扩展成组合概念。

d) 初始化模型树,并不断应用表4规则对组合概念进行扩展演算,直到无规则可应用或是出现完全冲突。

伪代码如下:

```
(simConceptSet, simRelationsSet) = input(f, c, d);
( $N_C, N_R, A^2$ ) = initialize( $DL^2$ );
combineConcept = expand(D);
while(apply the Tableau rules)
    if(originSat != false && simisat != false)
        (originSat, simisat) = sat(combineConcept);
    end if
end while
```

表4 组合 Tableau 规则表

规则名称	规则定义(其中 $i=1,2$)
$\sqcap$ 规则	如果 $(1)((A_1, C_1) \sqcap (A_2, C_2)) \in S_i((x, y))$; (2) $\{(A_1, C_1), (A_2, C_2)\} \not\subseteq S_i((x, y))$; 那么 $S_i((x, y)) = S_i((x, y)) \cup \{(A_1, C_1), (A_2, C_2)\}$
$\sqcup$ 规则	如果 $(1)(A_1, C_1) \sqcup (A_2, C_2) \in S_i((x, y))$; (2) $\{(A_1, C_1), (A_2, C_2)\} \cap S_i((x, y)) = \emptyset$; 那么 $S_i((x, y)) = S_i((x, y)) \cup \{(A_1, C_1)\}$ 或 $S_i((x, y)) = S_i((x, y)) \cup (A_2, C_2)$
$\exists$ 规则	如果 $(1)(\exists R.((A_2, C_2))) \in S_i((x, y))$; (2) 没有节点 (x', y') , 使得 $R((x, y), (x', y'))$ 成立并且 $(A_2, C_2) \in L_i((x', y'))$; 那么新增节点 (x', y') , 使得 $R((x, y), (x', y'))$ 成立并且 $L_1((x', y')) = \{(A_2, C_2)\}$ 。若 R 是由一对相似关系 R_1, R_1' 构成, 则有 $L_2((x', y')) = \{(A_2, C_2)\}$, 否则, $L_2((x', y')) = \{(A_2, C_2), \neg(A_2, C_2)\}$; 若 R_1 (或 R_1')不存在实例 $R_1(x, x')$ (或 $R_1'(y, y')$), 则有 $L_1((x', y')) = \{A_2, \neg A_2\}$ (或 $L_2((x', y')) = \{C_2, \neg C_2\}$)
$\forall$ 规则	如果 $(1)(\forall R.((A_2, C_2))) \in S_i((x, y))$; (2) 存在节点 (x', y') 使得 $R((x, y), (x', y'))$ 成立且 $(A_2, C_2) \notin L_i((x', y'))$, 那么 $L_i((x', y')) = L_i((x', y')) \cup \{(A_2, C_2)\}$
展开规则	对于 $(\forall A, C) \in L_i((x, y))$ 或 $S_i((x, y))$ 有 $L_1((x, y)) \cup \{A\}, L_2((x, y)) \cup \{C\}$ 或 $S_1((x, y)) \cup \{A\}, S_2((x, y)) \cup \{C\}$

3.2.2 算法性质

定理1 对任一组合概念 D , Tableau 算法都停机。事实上, 算法是否停机取决于算法规则的特征, 不会无止境地使用算法规则。已知 Tableau 算法对任一 ALC 概念都停机^[7], 即使组合概念的形式有别于一般描述逻辑的概念, 但算法本身还是基于基本的 Tableau 算法, 因此算法必停机。

定理2 组合知识库推理复杂度不会高于基于 ALC 的 Tableau 算法本身的复杂度。ALC 的 Tableau 算法的复杂度是 PSPACE-完全的^[7], 而组合知识库是结合两不同知识库的知识, 它的规模大小取决于构成它的两个不同知识库的规模。但由于知识库的概念集及断言集是有限的, 假设不同的两个概念集个体数分别为 C_1, C_2 , 则组合概念的个数最大为 C^2 , 其中 $C = \max\{C_1, C_2\}$, 假设每个概念下都有最大的 n 个体, 则组合概念的个体数为 $n^2 C^2$; 概念关系集元素个数分别是 m_1, m_2 , 组合概念个体数为 $n^2 C^2(m_1 + m_2)$, 那么生成组合知识库的复杂度不会超过 $O(n^5)$ 。而推理主要依据 Tableau 算法, 其本身的复杂度是 PSPACE-完全的, 因此组合知识库的推理的复杂度取

决于 Tableau 算法的复杂度。

3.2.3 实例

假设有如下两个知识库:

$KB_1: Bstat(a), Bstat(b), Bstat(c), Bstat(d), trans(c), trans(d), R(a, b), R(b, c), R(c, d)$;

$KB_2: mstat(a'), mstat(b'), mstat(c'), mstat(d'), R'(a', b'), R'(b', c'), R'(c', d')$ 。

个体 c 与 c' 满足 f , 那么在 KB_1 中检测是否存在一个概念 $Bstat$ 的个体 x 与 d' 有关系 R , 并且 x 是概念 $trans$ 的一个实例。先将问题描述转换成一般的概念描述: $trans \sqcap \exists R. Bstat(d')$ 。算法步骤如下:

a) 计算相似概念集与相似关系集。

$simConceptSet = \{(Bstat, mstat), (trans, mstat)\}$

$simRelationsSet = \{(R, R')\}$

b) 初始化组合知识库, 构造出原子概念集 N_C 、组合关系 N_R 和断言集 A^2 。

$N_C = \{(Bstat, mstat), (trans, mstat)\}$

$N_R = \{(R, R'), (R, R_1), (R_2, R')\}$

$A^2 = \{(Bstat, mstat)(a, a'), (Bstat, mstat)(a, b'), \dots, (trans, mstat)(c, a'), \dots, (trans, mstat)(d, d'), (R, R')((a, a')(b, b')), (R, R')((b, b')(c, c')), \dots, (R, R_1)((a, a')(b, a'), \dots, (R_2, R')((d, c')(d, d'))\}$

其中: (R, R_1) 表示存在原概念的关系, 相似概念的关系不存在; (R_2, R') 表示相似概念存在关系, 但原概念不存在关系。

c) 扩展概念为组合概念。

$(trans, mstat) \sqcap \exists (R, R'). (Bstat, mstat)(d, d')$

d) 初始化模型树应用 Tableau 规则进行演算。

(a) 初始化模型树, 有节点 (x, y) , 其标注为 $S_1((x, y)) = S_2((x, y)) = \{(trans, mstat) \sqcap \exists (R, R'). (Bstat, mstat)\}$;

(b) 应用 $\sqcap$ 规则, 得到 $S_1((x, y)) = S_2((x, y)) = \{(trans, mstat) \sqcap \exists (R, R'). (Bstat, mstat), (trans, mstat), \exists (R, R'). (Bstat, mstat)\}$;

(c) 应用 $\exists$ 规则, 产生新的节点 (x', y') , 其标注为 $N_1((x', y')) = N_2((x', y')) = \{(Bstat, mstat)\}$, 因为 R 中并没有出现 d' , 所以原来关系中没有这个实例, 则 $N_1((x', y')) = \{(Bstat, mstat), Bstat, \neg Bstat\}$, $N_2((x', y')) = (Bstat, mstat)\}$;

(d) 应用展开规则, 则有 $N_2((x', y')) = \{(Bstat, mstat), mstat\}$, $S_2((x, y)) = \{(trans, mstat) \sqcap \exists (R, R'). (Bstat, mstat), (trans, mstat), \exists (R, R'). (Bstat, mstat), mstat\}$;

(e) 无规则可再应用, 算法停止。结果为相似可满足的, 即存在个体 (c, c') 满足组合概念。

4 结束语

描述逻辑是知识表示的形式化, 不仅知识表示的能力强大, 而且具备很强的推理能力, 目前描述逻辑应用于如语义网、配置等诸多领域。在不同的知识库上, 它们各自描述某一特定的领域, 而现实生活当中, 这些领域可能是交叉、关联的。本文是在基于语义积的形式上通过对两个不同的知识库组合关联起来, 借用知识库间的概念相似来进行关联。事实上, 要联合不同的知识库并将它们真正结合起来, 还有很多工作要做, 未来的工作主要是对组合描述逻辑进行 (下转第4511页)

摘录效率(以 F 为标准)与该学科语料的句子平均信息量(表7)基本成反比,即语料的句子平均信息量越大,摘录效率越低。如果注意到计算机科学的语料中存在影响语料规模统计的因素(如公式等),这个规律就更加可靠了。究其原因,如果每个句子包含的词语越多且文本包含的句子数越少,即句子平均信息量越大,则优选过程中删除一个句子所损失的关键信息就越多,摘录效率自然也就越低。

除上述用于测定方法自身效率的实验外,本课题组还基于相同语料将本文方法与文献[8~11]中的方法进行了对比性实验(具体步骤请参看相关文献)。对比性实验中所有的方法选用了相同的输入矩阵,并进行了相同的LSA,因此实验结果所反映的事实上就是不同关键句子选择算法对方法整体效率带来的影响。该实验用准确率、召回率和 F 度量值的平均值作为比较参数,实验结果如表9所示。

表9 本文方法与其他方法的对比分析 /%

对比项	本文方法	文献[8]	文献[9]	文献[10]	文献[11]
P 的平均值	75.9	65.4	69.1	71.7	66.3
R 的平均值	71.8	70.2	68.5	67.9	71.2
F 的平均值	73.8	67.7	68.8	69.7	68.7

表9的结果显示,与文献[8~11]的同类方法相比,本文方法在各个比较参数的对比中都具有一定优势。可见改进后的方法不仅实现了全程无监督,而且在整体效率上有较大提升,更具应用潜质。

4 结束语

通过引入LSA,本文提出了一种无监督的中文摘录方法,该方法借助SVD将主题信息抽象为潜在概念,并利用句子与潜在概念的语义相关度来判断句子与主题的语义相关度,由此完成关键句子的选取。实例及实验证明,本文方法可行且有效。

作为本文方法的核心和基础,LSA在发挥其功效的同时也不可避免地遇到了如下两个问题:

a) LSA所处理的输入矩阵其构建基础是向量空间模型,受限于独立性假设(单向量模型的分量或多向量模型的分向量表征的特征相互独立),此模型在支撑文本到输入矩阵转换时丢失了所有关联信息^[14],而这些关联信息(如词序、句法语义约束等)很可能成为提升摘录效率的辅助条件。

b) 当主题信息集中于个别句子时,输入矩阵可能出现数据不均的状况,此时,LSA的分析效率会呈现一定程度的下降,

这对摘录结果无疑会产生负面影响。

为解决上述问题,本研究下一步将重点思考如何用富含关联信息的模型(如图模型)来构建输入矩阵。此外,如何将LSA的应用范围从单文本摘录升级到多文本摘要也将是本研究努力的方向。

参考文献:

- [1] 刘兴林,郑启伦,马千里.一种基于主题词集的自动文摘方法[J].计算机应用研究,2011,28(4):1322-1324.
- [2] 陈燕,龙建勋.基于明确语义分析的自动文摘算法[J].计算机工程,2011,37(3):183-185.
- [3] 余正涛,樊孝忠,郭剑毅,等.基于潜在语义分析的汉语问答系统答案提取[J].计算机学报,2006,29(10):1889-1893.
- [4] 石晶,戴国忠.基于PLSA模型的文本分割[J].计算机研究与发展,2007,44(2):242-248.
- [5] 盖杰,王怡,武港山.潜在语义分析理论及其应用[J].计算机应用研究,2004,21(3):9-13.
- [6] 徐永东,徐志明,王晓龙.基于信息融合的多文档自动文摘技术[J].计算机学报,2007,30(11):2048-2054.
- [7] GONG Yi-hong, LIU Xin. Generic text summarization using relevance measure and latent semantic analysis[C]//Proc of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2001: 19-25.
- [8] STEINBERGER J, JEZEK K. Using latent semantic analysis in text summarization and summary evaluation[C]//Proc of ISIM. 2004: 93-100.
- [9] MURRAY G, RENALS S, CARLETTA J. Extractive summarization of meeting recordings[C]//Proc of the 9th European Conference on Speech Communication and Technology. 2005: 112-115.
- [10] OZSOY M G, ALPASLAN F N. Text summarization using latent semantic analysis[J]. Journal of Information Science, 2011, 37(4): 405-417.
- [11] 马亮,何婷婷,李芳,等.以关键词抽取为核心的文摘句选择策略[J].中文信息学报,2008,22(6):50-54.
- [12] 张瑾,王小磊,许洪波.自动文摘评价方法综述[J].中文信息学报,2008,22(3):81-88.
- [13] KONTOSTATHIS A, POTTENGER W M. A mathematical view of latent semantic indexing: tracing term cooccurrences, LU-CSE-02-006[R]. Bethlehem: Lehigh University, 2002.
- [14] 张晓艳.新闻话题和关联追踪技术研究[D].长沙:国防科学技术大学,2010.

(上接第4506页)进一步的理论探讨,改进完善Tableau算法及其规则,探讨算法的复杂性以及实际应用的可行性等。

参考文献:

- [1] 陆建江,张亚非,苗壮,等.语义网原理与技术[M].北京:科学出版社,2006.
- [2] 王金环,李宝敏.基于本体DL的语义推理研究[J].计算机技术与发展,2009,19(11):94-100.
- [3] WACHE H, VOEGELE T, VISSER U, et al. Ontology-based integration of information: a survey of existing approaches[C]//Proc of Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence. 2001: 108-117.
- [4] BAADER F, MCGUINNESS L D, NARDI D, et al. The description logic handbook: theory, implementation and application[M]. Cam-

bridge: Cambridge University Press, 2003.

- [5] 石莲,孙吉贵.描述逻辑综述[J].计算机科学,2006,33(1):194-197.
- [6] HORROCKS I. Description logics in ontology applications[C]//Proc of the 9th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods. Berlin: Springer, 2005: 2-13.
- [7] 梅靖,林作铨.从ALC到SHOQ(D):描述逻辑及其Tableau算法[J].计算机科学,2005,32(3):1-11.
- [8] HORROCKS I, SATTler U. A tableau decision procedure for SHOIQ[J]. Automated Reasoning, 2007, 39(3): 249-276.
- [9] GLIMM B, HORROCKS I, LUTZ C, et al. Conjunctive query answering for the description logic SHIQ[C]//Proc of the 20th International Joint Conference on Artificial Intelligence. 2007: 399-404.