基于ALCQ(D)的CBR事例表示及相似性度量

孙晋永[1] 古天龙[2] 常 亮[2] 马林威[2]
(西安电子科技大学计算机学院 西安 710071)[1]
(桂林电子科技大学广西可信软件重点实验室 桂林 541004)[2]

摘 要 针对目前用于CBR事例表示的描述逻辑,如$\mathcal{EL}$、ALC、ALCNR等缺少定性数量约束和有型域约束的问题,将具有定性数量约束和有型域构子的描述逻辑ALCQ(D)应用于CBR中。首先使用ALCQ(D)概念表示有定性数量约束、具体数据类型和数据值约束需求的CBR事例,并对之索引。研究两种主要的具体数据类型:数值类型和符号类型。然后定义ALCQ(D)范式来规范事例的索引表示,最后给出事例相似性度量方法。该度量方法先对事例索引的各个部分进行相似性度量,然后对度量结果进行加权求和得到最终相似性。实验结果表明,ALCQ(D)可以更准确地表示事例,事例相似性度量方法可以更贴切地度量事例的相似性,这对提高事例检索的速度和准确性以及提高CBR系统的效率具有重要意义。

关键词 基于事例推理,描述逻辑,事例表示,事例检索,相似性
中图法分类号 TP18 **文献标识码** A

Research on CBR's Case Representation and Similarity Measure Based on ALCQ(D)

SUN Jin-yong[1] GU Tian-long[2] CHANG Liang[2] MA Lin-wei[2]
(School of Computer Science and Technology, Xidian University, Xi'an 710071, China)[1]
(Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China)[2]

Abstract Focused on the lack of qualified number restrictions and concrete domains restrictions in DLs such as $\mathcal{EL}$, ALC, ALCNR that have been used in CBR's case representation, ALCQ(D) was used with which qualified number restrictions and concrete domains constructor were equipped. First, ALCQ(D) concepts were used to represent and index cases with the requirements of qualified number restrictions, concrete data types and numerical restrictions. Two concrete domain types which are numerical data type and symbolic data type were studied. Second, the normal form of ALCQ(D) was defined to normalize case representations in the form of indexes. Finally the measure method for case similarity was presented, which measures similarities of all parts of the case representations, then weights and summates gained similarities. Experimental results show that ALCQ(D) represents cases more accurately and the measure method for case similarity measures the similarity between cases more adequately. It is very important for increasing the speed of case retrieval, for improving the accuracy of case retrieval, and for improving the efficiency of the CBR system.

Keywords Case-based reasoning(CBR), Description logic(DL), Case representation, Case retrieval, Similarity

1 引言

基于事例推理(Case-Based Reasoning,CBR)是一种重要的基于知识的问题求解和学习方法[1,2],在智能规划、产品设计、故障检测、模式分类、电子商务、软件复用、法律推理、医疗服务等领域得到了广泛应用[3]。

CBR的基本思想是以事例的形式组织过去的经验或经历,通过在历史事例库中检索出与目标事例相似的事例,并根据目标事例与相似事例的差异进行修正,得到目标事例的解。一个完整的CBR系统包括4个循环过程[2]:检索(Retrieve)、重用(Reuse)、修正(Revise)和保留(Retain)。

目前,CBR研究面临的问题有[4]:如何表示事例?如何选择事例索引以有效地组织与存储事例?如何度量事例的相似性?如何检索出相似事例?如何修正历史事例的解使之适用于新的事例?…等等。在知识密集型应用(如语义Web、数据挖掘、软件复用、产品设计等)中大规模事例库的表示、组织、存储及其操作是CBR研究亟待解决的问题。

借鉴人工智能中的知识表示方法,目前已建立了多种CBR事例表示方法[5]。这些方法大致分为两类:特征向量表示法和结构化表示法。特征向量方法用一组特征(属性)/值

到稿日期:2013-05-13 返修日期:2013-10-16 本文受国家自然科学基金(60963010,60903079,61262030,61363030),广西自然科学基金(2012GXNSFBA053169)资助。

孙晋永(1978—),男,博士生,CCF会员,主要研究方向为知识表示与推理、CBR推理,E-mail:sunjy@guet.edu.cn;古天龙(1964—),男,教授,博士生导师,CCF会员,主要研究方向为形式化方法、符号计算等;常 亮(1980—),男,博士,教授,CCF会员,主要研究方向为知识表示与推理、智能规划、形式化方法;马林威(1987—),男,硕士生,主要研究方向为描述逻辑、CBR推理。

对以向量的形式对事例进行表示，但不能描述事例的内部结构。结构化方法可以实现事例的内部结构(如层级结构、网络结构、流结构等)表示。使用框架表示事例是该方法的一种传统形式，由于框架可以看作一阶逻辑的子集，这种表示方法已经用逻辑实现了部分形式化。

事例相似性度量是CBR事例检索过程中最重要的步骤，目的是对当前事例与历史事例的相似程度进行评价。合适的度量方法可以在历史事例库中迅速、准确地查找到所需要的事例。常用的相似性度量方法主要有两种：表层相似性度量和结构相似性度量。表层相似性度量方法是与事例特征向量表示法紧密相关的传统度量方法，直接从事例表示的原始数据中获取事例的相似性度量。结构相似性度量方法依据事例的内部结构进行相似性度量，具有更贴切地度量事例相似性的优点，因此得到了广泛的应用。

描述逻辑(Description Logic，DL)是基于框架知识表示的形式化工具，是一阶谓词逻辑的一个可判定子集。在众多知识表示的形式化方法中，描述逻辑受到人们的特别关注，主要原因在于：具有清晰的模型－理论语义；对概念性知识的处理，特别是对概念分层的处理非常有效；提供了有效的推理服务，实现了知识表达能力和推理可判定性的统一。

从20世纪90年代起，许多学者就尝试将描述逻辑应用于CBR中。正式将描述逻辑引入CBR中的学者是Koehler。Koehler[6,7]给出了描述逻辑ALC和规划逻辑结合的混合知识表示方法，将ALC作为规划事例库的查询/检索语言。Kamp[8]使用能够描述数值、串和符号集合等的描述逻辑C_{TL}，给出了通过C_{TL}的基本推理实现CBR事例检索的3种方法，事例之间的相似性采用数值相似性度量。Salotti和Ventos[9]采用描述逻辑C-CLASSIC概念来表示事例，将事例之间的相似性和相异性形式化地定义为符号概念，事例检索通过C-CLASSIC的自动概念分类来实现。d'Amato等[10-12]给出了描述逻辑ALC、ALN表示下的实例与实例、实例与概念、概念与概念间相似性度量方法。Janowicz[13,14]给出了描述逻辑ALCNR、SHI表示下的事例相似性度量方法。Gomez-Albarran等[15]总结了前期学者的研究成果，提出了基于描述逻辑的CBR系统开发模型，即使用描述逻辑形式化地表示CBR系统中的结构化知识(即事例)，使用描述逻辑的推理机制实现CBR的事例检索、事例修正与学习等。

描述逻辑刻画CBR事例知识的内部结构的优势及其良好的推理能力，尤其是从知识库显式包含的知识中推导出隐含包含的知识的能力，对提高CBR的事例检索的准确性和完备性有很大帮助。不足的是，目前用于CBR事例表示的描述逻辑未能引入一些实际应用领域常见、但重要的约束条件或语义(如数量约束、有型域、时序信息、模糊信息等)，从而不能很准确地表示复杂的CBR事例。这在很大程度上降低了事例检索的准确性，也影响了整个CBR系统的效率。因此，在CBR的实际应用中，为了提高事例描述能力和事例检索的准确性与完备性，有必要将具有数量约束、有型域、时序信息或模糊信息等约束或语义的描述逻辑引入到CBR中。

针对上述问题，本文的主要研究工作是使用具有定性数量约束和有型域构子的描述逻辑ALCQ(D)描述有定性数量约束、具体数据类型和数据值约束需求的CBR事例，并对事例间的相似性进行度量。本文给出了ALCQ(D)范式的定义；在此基础上提出了基于ALCQ(D)的事例相似性度量方法。该度量方法可以更贴切地度量事例间的相似性，进而提高CBR事例检索的速度和准确性，是CBR事例相似性度量的一条可行的、有效的方法。

2 基于ALCQ(D)的事例表示方法

2.1 描述逻辑与CBR事例表示

描述逻辑通过定义应用领域的概念及其结构关系(角色)，刻画领域内的个体信息[16]。在概念和角色描述之上，由构子从简单概念和角色构造出复杂概念和角色。概念对应于逻辑中的一元谓词，角色对应于二元谓词，构子决定着语言的表达能力，类似于逻辑联结词的功能[18,19]。ALC是最基本的一种描述逻辑，构子包括合取、析取、否定、存在性约束和值约束。在ALC的基础上，增加数量约束、函数性约束、定性数量约束和有型域约束，就分别演变为ALCN、ALCF、ALCQ、ALC(D)[20]。

使用描述逻辑表示CBR事例的基本思想是：将事例作为描述逻辑的个体，使用描述逻辑的概念来描述和索引它们。在描述逻辑的分类操作下，可以把事例的索引组织在一个层次结构中。

2.2 ALCQ(D)简介

(1)有型域D[16]

有型域(concrete domain，简称为D)，也称具体域，是对描述逻辑的一个扩展，使之包括数值、字符串和时间等这类有型对象。

定义1(有型域D)　有型域D是一个二元组$(\triangle^D, pred(D))$，其中$\triangle^D$是有型论域，$pred(D)$是谓词集合。任意n元谓词$P \in pred(D)$是有型论域上的n元关系，即：$P^D \sqsubseteq pred(D)^n$。

(2)ALCQ(D)的语法定义与语义解释

设N_C是概念名的集合，如$\{C, D, \cdots\}$；N_R是角色名的集合，如$\{R, S, \cdots\}$；N_f是具体特征名(特征式)的集合，如$\{f, g, \cdots\}$，且N_R、N_f彼此不相交。ALCQ(D)的角色集合是$N_R \sqcup N_f$，特征链是特征式的合成$f_1 \cdots f_n$，n表示整数。

定义2(描述逻辑ALCQ(D))　定义描述逻辑ALCQ(D)的概念集合为满足下列条件的最小集合[16]：

1)若概念名$C \in N_C$，则C是ALCQ(D)概念。

2)若C, E是ALCQ(D)概念，R是ALCQ(D)概念角色，则$(C \sqcap E)$，$(C \sqcup E)$，$(\neg C)$，$(\forall R.C)$，$(\exists R.C)$，$(\geqslant nR.C)$，$(\leqslant nR.C)$都是ALCQ(D)概念。

3)若$u_1, \cdots, u_n$是特征链，$P \in pred(D)$是n元谓词，则$\exists u_1 \cdots u_n.P$是ALCQ(D)概念。

例如，概念"至少有1个孩子的女性(Women)"可以更准确地表示为：

$\text{Women} \equiv \text{Person} \sqcap \text{Female} \sqcap \geqslant 1\text{hasChild}.\text{Person} \sqcap \exists \text{hasAge}.\geqslant_{21}$。

定义3(ALCQ(D)解释)　ALCQ(D)解释[16,21] $I=(\triangle^I, \triangle^D, \cdot^I)$，由非空集合$\triangle^I$(抽象论域)、有型域$\triangle^D$和解释函数$\cdot^I$组成。集合$\triangle^I$与$\triangle^D$不相交。解释函数$\cdot^I$将每个概念映射为$\triangle^I$的一个子集，将每个角色映射为$\triangle^I \times \triangle^I$上的一个二元关系。同时还将每个特征式映射为$\triangle^I \times \triangle^D$的一个子集[20]；即$f(x)=a$，其中$f$是特征式，抽象个体$x \in \triangle^I$，

函数值 $a\in\triangle^D$，记作$\langle x,a\rangle\in f^I$。特征链 $f_1\cdots f_n$ 解释为函数的合成 $u^I=f_1^I\cdots f_n^I$。

(3)ALCQ(D)的知识库

定义 4(ALCQ(D)知识库)　ALCQ(D)知识库 $KB=TBox+ABox$，简写为 KB(TBox,ABox)。

1)TBox,TBox 是概念描述及其关系的公理集，包含概念定义和概念的包含关系。概念定义如：Father≡Man⊓∃ hasChild. Person，即 Father 被定义为"有孩子的男人"。概念包含如：Woman⊑Person，即一个 Woman 也是一个人 Person。

2)ABox,ABox 是描述个体和关系断言的集合。断言陈述了一个个体是某个概念的实例或两个个体间存在的关系。概念断言如：Student (Mary)，关系断言如：hasFriend (Tom, Jack) 。

2.3　基于 ALCQ(D)的事例表示

Sanchez-Ruiz 等[22]使用 Michalski[23]提出的火车数据集作为实例，用描述逻辑 $\mathcal{EL}$ 来描述如图 1 所示的火车实例 train1 和 train2。

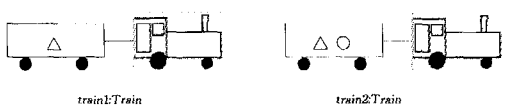

图 1　2 个火车实例

基于 $\mathcal{EL}$，用于表示火车实例 train1、train2 的最具体概念 MSC[22](Most Specific Concept)分别如下。为了表述方便，以 train1、train2 来指代它们。

train1 ≡ Train ⊓ ∃ hasCar. (CloseCar ⊓ ShortCar ⊓ ∃ load. Triangle⊓∃ wheels. Two)；

train2 ≡ Train ⊓ ∃ hasCar. (OpenCar ⊓ ShortCar ⊓ ∃ load. Triangle⊓∃ load. Circle⊓∃ wheels. Two)

现在对图 1 所示的两个火车实例进行扩展，增加其所载货物的复杂程度，得到图 2。

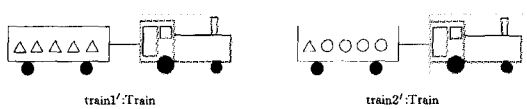

图 2　2 个扩展的火车实例

在图 2 中，火车实例 train1′装载 5 个三角形货物，火车实例 train2′装载 1 个三角形和 4 个圆形货物。每个三角形货物重量为 100Kg，每个圆形货物重量为 150Kg。对于这两个新实例，显然 $\mathcal{EL}$ 的描述能力不够，可以使用 ALCQ(D)来描述它们。于是，表示 train1′的最具体概念 MSC 为：

train1′ ≡ Train ⊓ ∃ hasCar. (CloseCar ⊓ LongCar ⊓ ∃ wheels. Two ⊓ = 5load. Triangle ⊓ ∃ hasTriangleWeight. $=_{100Kg}$)；

表示 train2′的最具体概念 MSC 为：

train2′ ≡ Train ⊓ ∃ hasCar. (OpenCar ⊓ LongCar ⊓ ∃ wheels. Two⊓∃ load. Triangle⊓=4load. Circle⊓∃ hasTriangleWeight. $=_{100Kg}$⊓∃ hasCircleWeight. $=_{150Kg}$)

其中，= 5load. Triangle 是 ≥5load. Triangle⊓≤5load. Triangle 的缩写形式，= 4load. Circle 是 ≥4load. Circle⊓≤4load. Circle 的缩写形式。

与 $\mathcal{EL}$ 相比，显然 ALCQ(D)能更准确地描述新的火车实例，能适应有定性数量约束、具体数据类型和数据值约束知识表示需求的应用领域。这里的每个火车实例对应于一个 CBR 事例。在实际应用中可能还需要描述事例的其它约束或语义，那么有必要进一步扩展 ALCQ(D)的描述能力。

2.4　ALCQ(D)范式

d'Amato[12]和 Janowicz[13]分别提出了描述逻辑 ALN、ALNR 概念的范式。ALCQ(D)是在描述逻辑 ALC 的基础上增加定性数量约束、有型域构子得到的，其概念用于表示 CBR 事例的索引。为了规范 CBR 事例索引的表示形式，本文提出了 ALCQ(D)范式。

定义 5(ALCQ(D)范式)　一个概念描述 D 是 ALCQ(D)范式当且仅当 $D\equiv\perp$ 或者 $D\equiv\top$ 或者 $D\equiv D_1\sqcup D_2\sqcup\cdots\sqcup D_m$。其中，

$$D_i=\prod_{A\in prim(D_i)}A\sqcap\prod_{R\in N_R}\Big[\prod_{E\in ex_R(D_i)}\exists R.E\sqcap\forall R.val(D_i)\sqcap\Big(\prod_{C\in\min_R(D_i)}\geq nR.C\Big)\sqcap\Big(\prod_{C\in\max_R(D_i)}\leq mR.C\Big)\sqcap\Big(\prod_{F\in N_f}\exists F.P\Big)\Big] \quad (1)$$

对于 $i=1,\cdots,m,D_i\neq\perp$。其中，$N_R$ 是角色的集合。具体解释如下：

1)$prim(D_i)$表示概念 D_i 所有的顶层基本概念或其否定组成的集合，如$\{A_1,A_2,A_3,\cdots,A_n\}$。

2)$ex_R(D_i)$表示在概念 D_i 的顶层角色 R 的存在约束下，如$\exists R.C'$，概念 C'的集合，如$\{C_1',C_2',C_3',\cdots,C_n'\}$。

3)$val_R(D_i)$表示在概念 D_i 的顶层角色 R 的值约束下，如$\exists R.C'$，概念 C'的合取，如 $C_1'\sqcap C_2'\sqcap\cdots\sqcap C_n'$。

4)$\min_R(D_i)$、$\max_R(D_i)$分别表示在概念 D_i 的顶层角色 R 的最小、最大定性数量约束下，如$\geq nR.C'$或$\leq nR.C'$，概念 C'的集合，如$\{C_1',C_2',C_3',\cdots,C_n'\}$。

5)$con_F(D_i)$ 表示在概念 D_i 的顶层有型特征(角色)F 下，如$\exists F.P$，谓词 P 的集合，如$\{P_1,P_2,P_3,\cdots,P_n\}$。

6) 解释 2)、3)、4)中的概念 C'以及解释 3)中的概念 $val_R(D_i)$也是 ALCQ(D)范式。

使用重写规则，如$(\forall R.C)\sqcap(\forall R.D)=\forall R.(C\sqcap D)$很容易把每个 ALCQ(D)概念转化成其范式形式[16,17]。本质上，该范式定义了 ALCQ(D)概念上的一个序。依据这个序，任意一个 ALCQ(D)概念都可以转换成范式的形式。将 2.3 节中扩展的火车实例的索引概念 train1′和 train2′与定义 5 的 ALCQ(D)范式进行对比，容易得出 train1′和 train2′是 ALCQ(D)范式的形式。

3　基于 ALCQ(D)的事例相似性度量

在基于描述逻辑的事例表示下，事例的相似性度量方法属于结构相似性度量方法。事例的相似性通过评价由描述逻辑概念表示的事例索引的相似性得到。所有描述逻辑概念的表达式及概念的包含关系的集合是一个拟序集或概念格，可以利用这种良好的知识结构性质来评价事例的相似性。概念与概念之间的相似性具有符号和数值两种表示形式。数值形式的相似性由于具有直观、方便的特点，因此被广泛采用。事例的相似性是其索引概念中的概念、角色及其内部结构相似性的加权和。

3.1　事例相似性度量公式

Janowicz[13]给出了基于 ALCNR 的事例表示和相似性度量方法，并将其用于欧洲某旅游城市的住宿预订门户网站的住宿服务搜索，得到了满意的预订方案。但不足的是，ALC-

NR 不具备定性数量约束和有型域约束，不能准确描述住宿预订应用中具有房间状态、费用等约束的事例，自然不能准确度量它们的相似性，从而不可能检索出最满意的预订方案。这些需求正在 ALCQ(D)的描述能力之内。针对实际应用中的类似需求，本文对 Janowicz 的相似性度量方法进行扩展，提出基于 ALCQ(D)的事例相似性度量方法。

定义 6(基于 ALCQ(D)的相似性度量函数) 假设 C、D 是 ALCQ(D)概念，$C\equiv C_1 \sqcup C_2 \sqcup \cdots \sqcup C_n$，$D\equiv D_1 \sqcup D_2 \sqcup \cdots \sqcup D_m$，其中 C_i、D_j（$i=1,2,\cdots,n$；$j=1,2,\cdots,m$）均为 ALCQ(D)范式。S_C、S_D 分别表示由 C_i、D_j 组成的集合。于是概念 C、D 的相似性度量函数被定义为：

$$\mathrm{sim}_u(C,D)=\sum_{(C_i,D_j)\in SI} w_{ij}\cdot \mathrm{sim}_i(C_i,D_j) \tag{2}$$

式中，$SI\subseteq S_{CP}$，其中 S_{CP} 是 S_C 与 S_D 的笛卡尔积，即 $S_{CP}=S_C\times S_D$。SI 的求解方法如下：

1)令 $SI=\emptyset$；

2)对每一个 $C_i\in S_C$，计算 $\mathrm{sim}_i(C_i,D_j)$，其中 $D_j\in S_D$；选 S_D 中与 C_i 相似度最大的元素 D_j 组成元组(C_i,D_j)，令 $SI=SI\cup(C_i,D_j)$；

3)对每一个 $D_j'\in S_D$，计算 $\mathrm{sim}_i(C_i',D_j')$，其中 $C_i'\in S_C$；选 S_C 中与 D_j'相似度最大的元素 C_i'组成元组(C_i',D_j')，令 $SI=SI\cup(C_i',D_j')$。

4)最后得到的 SI 即为所求的集合。

在式(2)中，w_{ij} 是 C_i 与 D_j 的相似性所占的比重。为了简化，假设 C_i 与 D_j 的每次相似性评价所占比重相同；并考虑 $\mathrm{sim}_u(C,D)$的归一化，设定 $w_{ij}=\frac{1}{|SI|}$，$|SI|$为集合 SI 的基数(即元素个数)。

在式(2)中，

$$\begin{aligned}\mathrm{sim}_i(C_i,D_j)=\frac{1}{\sigma}\Big[&\sum_{(A,B)\in SP}\mathrm{sim}_P(A,B)+\sum_{(R,S)\in SE}\mathrm{sim}_e(ex_R(C_i),ex_S(D_j))+\sum_{(R,S)\in SF}\mathrm{sim}_f(val_R(C_i),val_S(D_j))+\sum_{(R,S)\in SMIN}\mathrm{sim}_m(\min_R(C_i),\min_S(D_j))+\sum_{(R,S)\in SMAX}\mathrm{sim}_m(\max_R(C_i),\max_S(D_j))+\sum_{(F_1,F_2)\in SC}\mathrm{sim}_c(con_{F_1}(C_i),con_{F_2}(D_j))\Big]\end{aligned} \tag{3}$$

记概念 C_i、D_j 的存在约束、值约束、最小数量约束、最大数量约束、有型域的角色集合分别为 N_R^E、N_R^F、N_R^{Min}、N_R^{Max}、N_f 和 $N_R^{E'}$、$N_R^{F'}$、$N_R^{\mathrm{Min}'}$、$N_R^{\mathrm{Max}'}$、N_f'。在式(3)中，$SP\subseteq prim(C_i)\times prim(D_j)$，即 SP 为集合 $prim(C_i)$与 $prim(D_j)$的笛卡尔积的子集。$SE\subseteq N_R^E\times N_R^{E'}$，$SF\subseteq N_R^F\times N_R^{F'}$，$SMIN\subseteq N_R^{\mathrm{Min}}\times N_R^{\mathrm{Min}'}$，$SMAX\subseteq N_R^{\mathrm{Max}}\times N_R^{\mathrm{Max}'}$，$SC\subseteq N_f\times N_f'$。其中，SP、SE、SF、SMIN、SMAX 和 SC 的求解方法与 SI 的类似，此处不再重复。

在式(3)中，$\sigma=|SP|+|SE|+|SF|+|SMIN|+|SMAX|+|SC|$，即 σ 是集合 SP、SE、SF、SMIN、SMAX 和 SC 的基数之和。由此易得，$0\leqslant \mathrm{sim}_i(C_i,D_j)\leqslant 1$。

在式(3)中，

$$\mathrm{sim}_P(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{4}$$

式(4)用于计算基本概念 A、B 的相似性。可选的方法较多，此处采用概念 $A\sqcap B$ 的个体数量与概念 $A\sqcup B$ 的个体数量的比值作为概念 A、B 的相似性。在确定概念 A 和 B 的个体数量时，用到了描述逻辑的 $ABox$ 推理中的实例检查推理。

在式(3)中，

$$\mathrm{sim}_e(ex_R(C_i),ex_S(D_j))=\frac{1}{|SX|}\mathrm{sim}_r(R,S)\cdot\sum_{(E_i,E_j)\in SX}\mathrm{sim}_u(E_i,E_j) \tag{5}$$

式中，$SX\subseteq ex_R(C_i)\times ex_R(D_j)$，即 SX 为集合 $ex_R(C_i)$与 $ex_R(D_j)$的笛卡尔积的子集。SX 的求解方法与 SI 的类似，此处不再重复。式中 $\mathrm{sim}_r(R,S)$为角色 R、S 的相似性，如式(6)所示，计算方法类似式(4)。

$$\mathrm{sim}_r(R,S)=\frac{|R\cap S|}{|R\cup S|} \tag{6}$$

在式(5)中，$\mathrm{sim}_u(E_i,E_j)$为概念 E_i、E_j 的相似性，由式(2)计算。

在式(3)中，

$$\mathrm{sim}_f(val_R(C_i),val_S(D_j))=\mathrm{sim}_r(R,S)\cdot\mathrm{sim}_u(val_R(C_i),val_S(D_j)) \tag{7}$$

式(7)的解释类似式(5)。

3.2 定性数量约束的相似性度量

Janowicz[13]提出的 ALCNR 概念的相似性度量方法只能处理角色的非定性数量约束，如$\geqslant nR$ 和$\leqslant nR$ 的形式。在基于 ALCQ(D)的事例表示前提下，本文对角色的非定性数量约束进行扩展，提出了处理角色的定性数量约束，如$\geqslant nR.C$ 和$\leqslant nR.C$ 的形式的相似性度量方法。这是式(3)的一个组成部分。在式(3)中，

$$\mathrm{sim}_m(m_R(C_i),m_S(D_j))=\mathrm{sim}_r(R,S)\cdot\Big[\frac{1}{|SM|}\sum_{\substack{X_k\in m_R(C_i),\\ Y_l\in m_S(D_j)}}(\mathrm{sim}_u(X_k,Y_l)\cdot(1-\frac{|a-b|}{\max(a,b)}))\Big] \tag{8}$$

式中，$m_R(C_i)$、$m_S(D_j)$分别表示 $\max_R(C_i)$、$\max_S(D_j)$或 $\min_R(C_i)$、$\min_S(D_j)$。$SM\subseteq m_R(C_i)\times m_S(D_j)$，即 SM 为角色 R、S 的最大(最小)数量约束下的概念集合的笛卡尔积的子集。SM 的求解方法与 SI 的类似，此处不再重复。a 是 C_i 中在角色 R 的数量约束下概念 X_k 对应的基数，如$\geqslant aR.X_k$ 或$\leqslant aR.X_k$；b 是 D_j 中在角色 S 的数量约束下概念 Y_l 对应的基数，如$\geqslant bS.Y_l$或$\leqslant bS.Y_l$。$\mathrm{sim}_u(X_k,Y_l)$用于求$\geqslant aR.X_k$ 与$\geqslant bS.Y_l$ 或$\leqslant aR.X_k$ 与$\leqslant bS.Y_l$ 中的定性约束概念 X_k 与 Y_l 的相似性。其中，$\max(a,b)$为 a、b 中的较大者。类似地，记 $\min(a,b)$为求 a、b 中的较小者。

3.3 有型域约束的相似性度量

在 CBR 的实际应用中，具体数据类型和数据值约束很常见，在进行事例相似性度量时具有重要的地位。但目前用于 CBR 的描述逻辑，如 $\mathcal{EL}$、ALC、ALCNR 等，均未引入有型域的语义，这限制了它们对复杂 CBR 事例的表示能力。在基于 ALCQ(D)的事例表示下，本节提出有型域约束下概念的相似性度量方法。这也是式(3)的一个组成部分。在式(3)中，

$$\mathrm{sim}_c(con_{F_1}(C_i),con_{F_2}(D_j))=\begin{cases}\mathrm{sim}(P_1,P_2), & \text{if } F_1=F_2\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

式中，F_1、F_2 分别是概念 C_i、D_j 的顶层有型特征(角色)，$P_1\in con_{F1}(C_i)$，$P_2\in con_{F2}(D_j)$。当 $F_1\neq F_2$ 时，例如 $F_1=$ has-

Weight，F_2=hasHeight，显然 $\mathrm{sim}_c(con_{F1}(C_i), con_{F2}(D_j))=0$。所以，下面仅讨论 $F_1=F_2$ 的情况。

(1)当 P_1、P_2 的有型论域 $\triangle^D$ 为数值型时，P_1、P_2 的论域可视为实数集。

a)P_1、P_2 为等式型谓词

此时，谓词 P_1、P_2 形如概念“hasWeight. $=_{100Kg}$”中的“$=_{100Kg}$”。则 F_1. P_1、F_2. P_2 的形式可分别表示为 $F_1.=(d_1)$、$F_2.=(d_2)$，d_1、d_2 为数值型的确定值。于是

$$\mathrm{sim}(P_1,P_2)=\frac{|d_1-d_2|}{|\max-\min|} \tag{10}$$

式中，max、min 分别为 d_1、d_2 对应属性值取值范围的最大、最小值。

b)若 P_1、P_2 为不等式型谓词

此时，谓词 P_1、P_2 形如概念“hasAge. $\geqslant_{20}$”中的“$\geqslant_{20}$”。P_1、P_2 属于模糊区间属性类型，可以使用模糊集合的贴近度或文献[24]提出的混合概念格相似性度量方法计算出谓词 P_1、P_2 的相似性。

c)若 P_1、P_2 为模糊概念型谓词

例如在概念 hasSize. Large、hasSize. Medium 中，谓词分别为 Large、Medium。当 P_1、P_2 为这种类型的谓词时，它们属于模糊概念属性类型，也可以使用模糊集合的贴近度或文献[24]提出的混合概念格相似性度量方法计算出谓词 P_1、P_2 的相似性。

(2)当 P_1、P_2 的有型论域 $\triangle^D$ 为符号类型时，P_1、P_2 一般为等式型谓词。P_1、P_2 形如概念 hasLable. $=_{'CS005'}$ 中的“$=_{'CS005'}$”。则 F_1. P_1、F_2. P_2 的形式分别为 $F_1.=(d_1)$、$F_2.=(d_2)$，d_1、d_2 为符号型的确定值。这时，

$$\mathrm{sim}(P_1,P_2)=\begin{cases}1, & \text{if } d_1=d_2\\ 0, & \text{otherwise}\end{cases} \tag{11}$$

3.4 相似性度量实例

目前，研究基于 ALCQ(D)的 CBR 事例相似性度量的文献很少，相应数据集也较少。本文对 Michalski[23]的火车数据集进行扩展，得到如图 3 所示的 8 个火车实例。此处的每个火车实例也对应于一个 CBR 事例。分别使用文献[13,22]及本文的相似性度量方法对这些火车事例进行相似性度量，并分析度量结果。

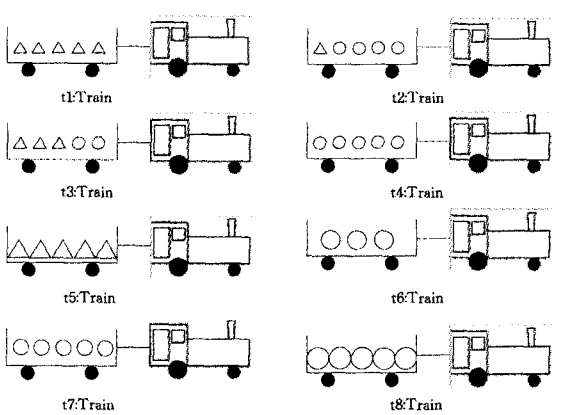

图 3　8 个扩展的火车实例

模糊概念型谓词 Medium 和 Heavy 的隶属度函数的曲线如图 4 所示。其中，Meadium = trz(100, 200, 300, 400, [0, 500])，Heavy=rs(300,400,[0,500])。

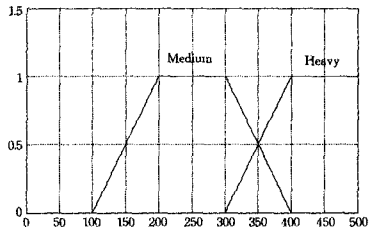

图 4　Medium 和 Heavy 的隶属度函数曲线

基于 ALCQ(D)的火车事例描述如下：

t1≡Train ⊓ ∃ hasCar. (OpenCar ⊓LongCar ⊓ ∃ wheels. Two ⊓ = 5load. Triangle ⊓ ∃ hasTriangleWeight. $=_{100Kg}$)；

t2≡Train ⊓ ∃ hasCar. (OpenCar ⊓LongCar ⊓ ∃ wheels. Two⊓ ∃ load. Triangle⊓=4load. Circle ⊓ ∃ hasTriangleWeight. $=_{100Kg}$⊓ ∃ hasCircleWeight. $=_{150Kg}$)

t3≡Train ⊓ ∃ hasCar. (OpenCar ⊓LongCar ⊓ ∃ wheels. Two⊓=3load. Triangle⊓=2load. Circle⊓ ∃ hasTriangleWeight. $=_{100Kg}$⊓ ∃ hasCircleWeight. $=_{150Kg}$)

t4≡Train ⊓ ∃ hasCar. (OpenCar ⊓LongCar ⊓ ∃ wheels. Two⊓=5load. Circle⊓ ∃ hasCircleWeight. $=_{150Kg}$)

t5≡Train ⊓ ∃ hasCar. (OpenCar ⊓LongCar ⊓ ∃ wheels. Two ⊓ = 5load. Triangle ⊓ ∃ has TriangleWeight. $=_{200Kg}$)

t6≡Train ⊓ ∃ hasCar. (OpenCar ⊓LongCar ⊓ ∃ wheels. Two⊓=5load. Circle⊓ ∃ hasCircleWeight. $=_{280Kg}$)

t7≡Train ⊓ ∃ hasCar. (OpenCar ⊓LongCar ⊓ ∃ wheels. Two⊓=5load. Circle⊓ ∃ hasCircleWeight. Medium)

t8≡Train ⊓ ∃ hasCar. (OpenCar ⊓LongCar ⊓ ∃ wheels. Two⊓=5load. Circle⊓ ∃ hasCircleWeight. Heavy)

此处省略基于描述逻辑 $\mathcal{EL}$ 和 ALCNR 的事例描述。相似性度量结果如表 1 所列。其中，sim(t1,t2)即为求火车实例 t1、t2 的相似性，其它类似。$S_{DL\rho}$ 为文献[22]的度量方法，利用两个事例描述概念的 LCS(Least Common Subsumer 最小公共包含)来计算相似性。S_1 为文献[13]的度量方法，两个事例的相似性为它们的概念描述的各部分的相似性的加权和。S_u 为本文提出的相似性度量方法。

表 1　方法 $S_{DL\rho}$、S_1 和 S_u 的相似性度量结果

序号	相似性度量	$S_{DL\rho}$	S_1	S_u
1	sim(t1,t2)	0.86	0.86	0.65
2	sim(t1,t3)	0.86	0.86	0.7
3	sim(t1,t4)	0.71	0.83	0.75
4	sim(t1,t5)	0.71	0.83	0.97
5	sim(t2,t3)	1	1	0.85
6	sim(t3,t4)	0.86	0.86	0.68
7	sim(t4,t6)	0.71	0.67	0.89
8	sim(t7,t8)	0.71	0.83	0.96

从表 1 可以看出，1)方法 $S_{DL\rho}$ 和 S_1 无法区分火车实例 t1 与 t2 和 t3 以及 t1 与 t4 和 t5 的差别。2)方法 $S_{DL\rho}$ 和 S_1 无法区分 t2 与 t3 的差别；而实际上二者确有差别，本文的方法 S_u 得出 t2 与 t3 的相似性为 0.85，符合客观事实。以上两点均是因为方法 $S_{DL\rho}$ 和 S_1 基于的描述逻辑无法表达定性数量约束，而方法 S_u 可以表达定性数量约束。3)方法 $S_{DL\rho}$ 和 S_1 认为：相比 t6，t4 与 t3 更相似。而实际上无论从货物的形状还是重量上看，t4 与 t6 应该更相似，方法 S_u 的结果证实了这一

点。这是因为方法 $S_{DL\rho}$ 和 S_1 所基于的描述逻辑无法表达有型域约束，而方法 S_u 可以表达有型域约束。4)$t7$ 和 $t8$ 的重量约束为模糊概念，方法 $S_{DL\rho}$ 和 S_1 认为二者的重量不同，而方法 S_u 认为 $t7$ 和 $t8$ 的重量有一定相似性，从而得到了比方法 $S_{DL\rho}$ 和 S_1 高的相似性。以上 4 点说明了，使用描述逻辑 ALCQ(D)描述具有定性数量和有型域约束的复杂 CBR 事例的必要性和准确性。

进一步，本文将文献[13,22]的相似性度量方法的思想应用到描述逻辑 ALCQ(D)上，分别得到了基于 ALCQ(D)的相似性度量方法 $S_{DL\rho}'$、S_1'。使用它们来度量图 3 中的 8 个火车实例的相似性，并与本文的方法 S_u 比较，结果如表 2 所列。

表 2　方法 $S_{DL\rho}'$、S_1'和 S_u 的相似性度量结果

序号	相似性度量	$S_{DL\rho}'$	S_1'	S_u
1	sim(t1,t4)	0.56	0.83	0.75
2	sim(t1,t5)	0.75	0.83	0.97
3	sim(t4,t6)	0.56	0.67	0.89
4	sim(t7,t8)	0.75	0.83	0.96

从表 2 可以看出，1)整体上，由方法 S_1'、S_u 计算出的相似性的数值要高于 $S_{DL\rho}'$。这是因为 $S_{DL\rho}'$方法是一种基于距离的相似性度量方法，而 S_1'、S_u 方法是从事例概念描述的语法和语义上更细粒度地度量相似性。2)方法 S_1'无法区分火车实例 $t1$ 与 $t4$ 和 $t5$ 的差别；而方法 S_u 考虑了它们的定性数量约束和有型域约束，从而将二者区分开。得出 $\mathrm{sim}(t1,t5)>\mathrm{sim}(t1,t4)$是符合客观事实的。3)方法 S_u 得到的相似性 $\mathrm{sim}(t1,t4)$小于方法 S_1'得到的相似性 $\mathrm{sim}(t1,t4)$，这是因为 S_u 考虑到了二者装载的货物外形的差别而方法 S_1'没有考虑。4)对于实例 $t7$ 与 $t8$，由于方法 S_1'认为 $t7$ 与 $t8$ 的重量不同，而 S_u 认为 $t7$ 与 $t8$ 的重量有一定的相似性，从而 S_u 得到了更高的相似性。以上 4 点说明了，相比 $S_{DL\rho}'$和 S_1'，方法 S_u 可以得到更准确的相似性度量结果。从而得出，方法 S_u 优于方法 $S_{DL\rho}'$和 S_1'。

从以上两个实例可以得出，引入描述逻辑 ALCQ(D)描述带定性数量约束和有型域约束的复杂 CBR 事例，提高了事例表示的准确性；使用本文的方法 S_u 来度量它们的相似性，提高了事例相似性度量的准确性。

3.5　讨论

本节将证明定义 6 给出的相似性度量函数是一个相似性函数。首先，给出相似性函数的定义。

定义 7(相似性函数[13])　设 S 是一个元素空间，函数 f 是定义在集合 $S\times S$ 上的一个实值函数。如果 f 满足以下准则，则它是一个相似性函数。

1) $f(a,b)\geqslant 0, \forall a,b\in S$;

2) $f(a,b)=f(b,a)$;

3) $\forall a,b\in S, f(a,b)\leqslant f(a,a)$。

在定义 6 中，由于 $\mathrm{sim}_u(C,D)$被递归地定义为非负数的和，因此 $\mathrm{sim}_u(C,D)\geqslant 0$，满足准则 1)。并且，由于权重 W_{ij} 的存在，$\mathrm{sim}_u(C,D)\leqslant 1$，从而 $0\leqslant\mathrm{sim}_u(C,D)\leqslant 1$。

由于在求 $\mathrm{sim}_u(C,D)$时引入的运算，如求和、集合的交与并、求绝对值、求最大与最小等都是可交换的，因此 $\mathrm{sim}_u(C,D)=\mathrm{sim}_u(D,C)$，满足准则 2)。

对于任意的 ALCQ(D)概念 C、D，可能是基本概念、值约束概念、存在约束概念、数量约束概念、有型域约束概念或者它们的复合。现在以定性数量约束概念为例证明 $\mathrm{sim}_u(C,D)\leqslant\mathrm{sim}_u(C,C)$。设 $C=\leqslant_a R.X, D=\leqslant_b S.Y$，则

$$\mathrm{sim}_u(C,D)=\mathrm{sim}_r(R,S)\cdot\mathrm{sim}_u(X,Y)\cdot\left(1-\frac{|a-b|}{\max(a,b)}\right)$$
$$\leqslant 1=\mathrm{sim}_r(R,R)\cdot\mathrm{sim}_u(X,X)\cdot\left(1-\frac{|a-a|}{\max(a,a)}\right)=\mathrm{sim}_u(C,C)$$

容易证明，其它情况也满足 $\mathrm{sim}_u(C,D)\leqslant\mathrm{sim}_u(C,C)$；从而满足准则 3)。

从以上可得，$\mathrm{sim}_u(C,D)$满足相似性函数的 3 个准则，因此定义 6 给出的相似性度量函数 $\mathrm{sim}_u(C,D)$是一个相似性函数。

结束语　本文先使用描述逻辑 ALCQ(D)来描述 CBR 事例；然后给出了用于索引 CBR 事例的 ALCQ(D)概念的范式定义；最后提出了基于 ALCQ(D)的事例相似性度量公式。ALCQ(D)是在描述逻辑 ALC 的基础上增加完全数量约束、有型域构子得到的，因此可用于描述有定性数量约束、具体数据类型和数据值约束需求的 CBR 事例。这种思路与方法为今后描述更复杂的 CBR 事例指明了方向。

本文提出的相似性度量公式可以全面、贴切地对具有定性数量和有型域约束的 CBR 事例的相似性进行评价，可以提高事例检索的准确性和完备性，为度量复杂 CBR 事例的相似性提供了一条有效的、可行的途径。不够完善的是，基于 ALCQ(D)的事例表示方法尚不能准确地描述具有模糊、时态语义等特性的事例；相应地本文的相似性度量方法也不支持这些特性。

下一步的工作是分析提出的相似性度量公式的时空复杂度，研究其在事例检索中的优化计算方法。另一项工作是研究更复杂 CBR 事例的表示方法，尤其是带模糊语义和时态特性的事例，给出相应的事例相似性度量方法。

参 考 文 献

[1] Koloder J. An introduction to case based reasoning[J]. Artificial Intelligence Review, 1992, 6(1): 3-44

[2] Aamodt A, Plaza E. Case-based reasoning: Foundational issues, methodological variations, and system approaches[J]. AI Communications, 1994, 7(1): 39-59

[3] Bartsch-Sporl B, Lenz M, Hubne A. Case based reasoning: survey and future directions[C]// Lecture Notes in Computer Science 1570. Springer, 1999: 67-89

[4] Lopez De Mantaras R, McSherry D, Bridge D, et al. Retrieval, reuse, revision and retention in case-based reasoning[J]. Knowledge Engineering Review, 2005, 20(3): 215-240

[5] Bergman R, Kolodner J, Plaza E. Representation in Case-Based Reasoning[J]. Knowledge Engineering Review, 2005, 20(3): 209-213

[6] Koehler J. An application of terminological logics to case based reasoning[C]// Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning. San Francisco, 1994: 351-362

[7] Koehler J. Planning from second principles[J]. Artificial Intelligence, 1996, 87(1/2): 145-186

[8] Kamp G. Using description logics for knowledge intensive case-based reasoning[C]// Lecture Notes in Computer Science 1168.

Springer,1996:204-218

[9] Salotti S, Ventos V. Study and formalization of a case based reasoning system using a description logic[C]// Lecture Notes in Computers Science 1488. Springer,1998:286-297

[10] d'Amato C, Fanizzi N, Esposito F. A semantic similarity measure for expressive description logics[C]// Proceedings of Convegno Italiano di Logica Computazionale(CILC05). Rome, Italy, 2005

[11] d'Amato C, Fanizzi N, Esposito F. A dissimilarity measure for ALC description logic[C]// Proceedings of the 21st Annual ACM Symposium of Applied Computing (SAC2006). Dijon, France,2006,2:1695-1699

[12] Fanizzi N, d'Amato C. A similarity measure for the ALN description logic[C]// Proceedings of Convegno Italiano di Logica Computazionale(CILC06). Bari, Italy,2006

[13] Janowicz K. Sim-DL: Towards a semantic similarity theory for the description logic ALCNR in geographic information retrieval [C]// Lecture Notes in Computers Science 4278. Springer, 2006:1681-1692

[14] Janowicz K, Wilkes M. SIM-DLA: A novel semantic similarity measure for description logics reducing inter-concept to inter-instance similarity[C]// Lecture Notes in Computers Science 5554. Springer,2009:353-367

[15] Gomez-Albarran M, Gonzalez-Calero P, Diaz-Agudo B, et al. Modelling the CBR life cycle using description logics[C]// Lecture Notes in Computers Science 1650. 1999:147-161

[16] Baader F, Calvanese D, McGuinness D. The description logic handbook: theory, implementation and applications[M]. Cambridge University Press,2003

[17] Brandt S, Küsters R, Turhan A Y. Approximating ALCN-Concept Descriptions[C]// Proc. of the 2002 Int. Workshop on Description Logics. 2002

[18] 常亮,王娟,古天龙,等. 时态描述逻辑 ALC-LTL 的 Tableau 判定算法[J]. 计算机科学,2011,38(8):150-154

[19] 蒋运承,汤庸,王驹,等. 面向语义 Web 的描述逻辑[J]. 模式识别与人工智能,2007,20(1):48-54

[20] 常亮,史忠植,陈立民,等. 一类扩展的动态描述逻辑[J]. 软件学报,2010,21(1):1-13

[21] Stanchev L. On efficient access to knowledge bases[C]// Proceedings of The 20th Midwest Artificial Intelligence and Cognitive Science Conference. Indiana University-Purdue University Fort Wayne, Fort Wayne,2009

[22] Sanchez-Ruiz A A, Ontanon S, Gonzalez-Calero P A, et al. Measuring similarity in description logics using refinement operators [C]// Lecture Notes in Artificial Intelligence 6880. Springer, 2011:289-303

[23] Larson J, Michalski R S. Inductive inference of VL decision rules [J]. ACM SIGART Bulletin,1977(63):38-44

[24] 鞠可一,周德群,吴君民. 混合概念格在案例相似性度量中的应用[J]. 控制与决策,2010,25(7):987-992

(上接第 218 页)

[5] 解立群,颜清华,陈颖. 从"围观模型"看交流困境——微博社会网络图谱分析[J]. 中国传媒科技,2011,8:92-95

[6] 张佰明. 嵌套性:网络微博发展的根本逻辑[J]. 国际新闻界,2010(6):81-85

[7] Backstrom L, Kumar R, Marlow C, et al. Preferential behavior in online groups[C]// Proceedings of the International Conference on Web Search and Web Data Mining. ACM,2008:117-128

[8] Borgatti S P, Foster P C. The network paradigm in organizational research: A review and typology[J]. Journal of management, 2003,29(6):991-1013

[9] Garton L, Haythornthwaite C, Wellman B. Studying online social networks[J]. Journal of Computer-Mediated Communication,2006,3(1)

[10] Flake G W, Lawrence S, Giles C L. Efficient identification of Web communities[C]// Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2000:150-160

[11] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences of the United States of America,2004,101 (9):2658-2663

[12] Lacoste-Julien S, Sha F, Jordan M I. DiscLDA: Discriminative learning for dimensionality reduction and classification[C]// Advances in Neural Information Processing Systems (NIPS). Vancouver, British Columbia, Canada,2008,21

[13] Hofmann T. Probabilistic latent semantic indexing [C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,1999:50-57

[14] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research,2003,3:993-1022

[15] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[J]. Advances in neural information processing systems,2002,2:849-856

[16] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(8):888-905

[17] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113

[18] Newman M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences, 2006,103(23):8577-8582

[19] Nowicki K, Snijders T A B. Estimation and prediction for stochastic blockstructures[J]. Journal of the American Statistical Association,2001,96(455):1077-1087

[20] Airoldi E M, Blei D M, Fienberg S E, et al. Mixed membership stochastic block models for relational data with application to protein-protein interactions[C]// Proceedings of the international biometrics society annual meeting. 2006

[21] Hofman J M, Wiggins C H. Bayesian approach to network modularity[J]. Physical review letters,2008,100(25):258701