

# 支持虚拟组织的语义基础设施的动态构建方法研究

陈旺虎<sup>1),2),3)</sup> 刘 晨<sup>1),2)</sup> 李厚福<sup>1),2)</sup> 王建武<sup>1),2)</sup>

<sup>1)</sup>(中国科学院计算技术研究所网络与服务计算研究中心 北京 100080)

<sup>2)</sup>(中国科学院研究生院 北京 100039)

<sup>3)</sup>(西北师范大学数学与信息科学学院 兰州 730070)

**摘 要** 提出一种从虚拟组织自治域的资源描述中抽取语义,然后聚合为虚拟组织的语义基础设施的方法.该方法引入了一种领域知识学习算法,用以建立当前语境相关的词法空间,以提高语义抽取和聚合的准确性及自动化程度,并且在语义聚合的过程中隐含了虚拟组织语义到自治域语义的映射,更好的支持了虚拟组织应用的构建和跨自治域资源的透明访问.实验表明,该方法能够适应虚拟组织的动态开放环境、有效支持虚拟组织的语义基础设施构建.

**关键词** 虚拟组织;词法空间;语义抽取;语义聚合  
中图法分类号 TP311

## An Approach to Dynamically Forming Semantic Infrastructure for Virtual Organizations

CHEN Wang-Hu<sup>1),2),3)</sup> LIU Chen<sup>1),2)</sup> LI Hou-Fu<sup>1),2)</sup> WANG Jan-Wu<sup>1),2)</sup>

<sup>1)</sup>(Research Center for Grid and Service Computing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

<sup>2)</sup>(Graduate University of Chinese Academy of Sciences, Beijing 100039)

<sup>3)</sup>(College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070)

**Abstract** The paper proposes an approach to developing the semantic infrastructure for virtual organizations by extracting semantics from resource descriptions in individual autonomous domains and then fusing the extractions. The approach makes the extracted and fused semantics more accurate by using an algorithm to identify the lexical space relative to current scenario and enhances automaticity of the semantic extraction and fusion. The mappings between the semantics of a virtual organization and the corresponding autonomous domains is reflected by the course of semantics fusion, therefore, the approach can well support to integrate applications for virtual organizations and transparently access heterogeneous resources across autonomous domains. Experiments show that the approach suits the dynamic and opening environment of virtual organizations by supporting the development of the semantic infrastructures.

**Keywords** virtual organization; lexical space; semantics extraction; semantics fusion

## 1 引 言

虚拟组织(Virtual Organization, VO)<sup>[1]</sup>的资源

在逻辑上可看作其可访问的各资源提供者的相应资源的集合,我们称该集合为虚拟组织的逻辑资源域.在本文中,我们从资源提供者的角度出发,将虚拟组织的每个资源提供者及其向虚拟组织所提供的资源

称为虚拟组织的一个自治域(Autonomous Domain, AD). 虚拟组织逻辑资源域的语义基础设施是构建虚拟组织的应用、实现虚拟组织各成员之间的资源共享和应用协作的基础. 如图 1 所示, 自治域 1 和自治域 2 的资源构成了虚拟组织 1( $VO_1$ )的逻辑资源域, 该逻辑资源域的语义用以支撑  $VO_1$  的应用的构建.

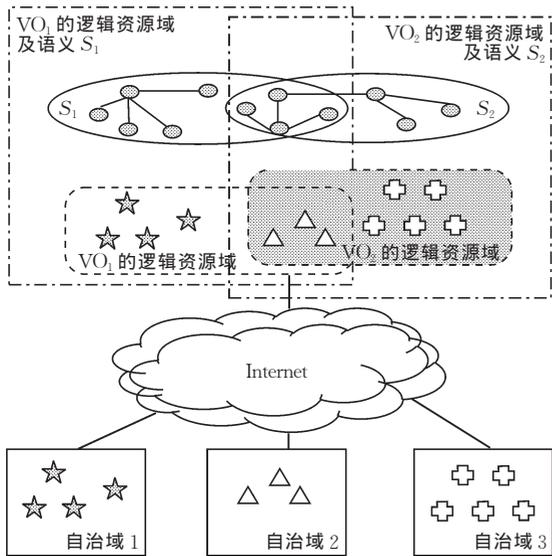


图 1 虚拟组织的逻辑资源域与语义基础设施

虚拟组织本身的特点为其语义基础设施的构建提出了新的挑战. 首先, 虚拟组织面向的是一个动态开放的环境, 其成员自治域的加入和退出是动态的, 因此使得其逻辑资源域可动态变化. 其次, 加入虚拟组织的各个自治域之间事先往往没有形成知识表示和语义描述上的共识. 最后, 因为虚拟组织的发起者往往关心的是一些特定领域的资源, 所以如图 1 所示, 尽管虚拟组织  $VO_1$  可能会因其生命周期的结束而解体, 但新的市场竞争又可能驱动虚拟组织  $VO_2$  的建立, 而  $VO_1$  和  $VO_2$  的逻辑资源域和语义基础设施往往会出现重叠. 也就是说, 一个 VO 在其生命周期结束时, 可为下次创建 VO 提供一些语义基础设施上的遗留成果.

虚拟组织的以上特点决定了传统的由领域专家从零开始构建语义的方式 (build-from-scratch)<sup>[2]</sup> 并不适合于其语义基础设施的构建. 同时, 因为各自治域在知识表示和语义描述上的差异, 考虑到效率和代价问题, 当前的语义聚合方式<sup>[3]</sup> 也因过多的人员参与而很难应用于虚拟组织的语义基础设施构建. 另外, 从图 1 可以看出, 语义的复用对于构建虚拟组织的语义基础设施是非常重要的. 因此, 我们需要探索一种支持虚拟组织的语义基础设施构建

方法.

本文提出一种支持虚拟组织的语义基础设施动态构建方法. 该方法从自治域的资源描述中自动抽取资源语义, 并聚合为虚拟组织逻辑资源域的一致性语义(图 2). 为适应 VO 动态开放的特点和实现 VO 构建过程中语义的复用, 该方法对原有语义聚合技术进行了改进, 有效地支持了自治域直接提供的语义或者新抽取的语义与 VO 当前语义的聚合.

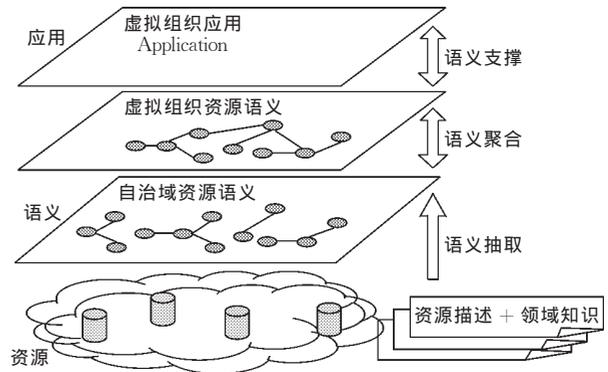


图 2 语义抽取与聚合的基本思想

首先, 该方法通过对领域知识的学习提取出当前语境下的词法空间, 保证了语义抽取过程中概念和概念间关系标注的准确性, 提高了抽取的自动化程度. 其次, 语义的聚合基于当前语境相关的词法空间, 提高了同义词自动标注的准确性, 减少了人员的参与. 最后, 提出向上最大化和向下最小化的思想, 使得虚拟组织逻辑资源域语义到自治资源域语义的映射隐含于语义聚合的过程之中, 使其更适合于跨自治域应用的构建和资源的透明访问. 所谓向上最大化和向下最小化是指: 在语义聚合时尽可能地使概念、关系及其约束具有更丰富的外延, 而在进行语义操作、从逻辑域的语义映射到自治域的语义时, 遵从自治语义的约束, 使得语义的外延最小化.

所要说明的是, 对于语义的抽取方法, 在本文中主要针对关系数据库资源来进行讨论, 而对半结构化和非结构化信息资源的语义抽取过程仅作简单的介绍. 因为, 对于各种资源类型来讲, 当前语境相关的词法空间的建立过程以及本体概念集的建立思想都是相同的(见 3.1 节和 3.2 节).

本文第 2 节分析当前有关语义抽取和聚合的相关研究工作; 第 3 节阐述如何从资源的描述信息中抽取其语义; 第 4 节论述如何通过语义聚合来建立虚拟组织的逻辑资源域的语义; 第 5 节对实验结果进行分析; 第 6 节介绍本文所提方法的应用案例; 最后给出结论——本文提出的方法可以有效支持虚拟

组织的语义基础设施的构建。

## 2 相关工作分析

本体是目前进行语义描述的主要手段,因此在本文中,语义的抽取结果以本体的方式描述,语义的聚合则通过本体的集成来实现。在本体的描述规范中,OWL<sup>①</sup>具有很强的语义描述能力和推理能力。OWL DL 作为 OWL 语言的子语言,因其在描述能力和推理效率上的平衡更是备受青睐。OWL 本体的逻辑基础是描述逻辑(DL)<sup>[4]</sup>。

文献[5]通过分析关系模式的约束属性和非约束属性以及函数依赖(DF),提出了一种将关系数据模式中的元素(包括关系、属性、元组、约束)映射为本体中相应要素(类、属性、实例、约束和公理)的规则。文献[6,7]在分析关系模式的同时,对元组之间的关系也进行了分析,以发现概念之间的继承关系等潜在的语义信息。文献[7]在分析关系模式的基础上,提出了通过对用户查询的设计和分析来修正生成本体的方法。

文献[9]提出一种从文本信息中抽取语义的方法。该方法从语言学的角度出发,分析出文本中所涉及的词及其它们之间的词法关系,然后将其分别转换为概念和概念之间的关系,并进一步转换为文本的语义描述。

Volz 等人提出一种从 XML Schema 中抽取语义的方法。首先,将 XML Schema 转换为一种规则树文法。规则树文法用一个四元组来表示,包括非终结符集、终结符集、开始状态集、产生规则集。然后,将非终结符集和终结符集转换为本体中的概念和角色<sup>②</sup>。

在文献[4]中,其研究方法没有涉及概念之间的继承关系,且仅依赖于关系模式的结构来生成本体,没有考虑关系模式中各要素在词法上的含义和关系;还要求关系模式满足 3NF,并且要获取其函数依赖(FD)集。文献[6~8]同样没有涉及关系模式中各要素在词法上的含义和关系。文献[9]从语法角度分析了文本中的词之间的关系,因此所抽取的概念以及概念之间的关系受制于当前的文本资源的构成。Volz 等人提出的方法同样缺乏概念之间在词法上的关系。

在词法技术的研究方面,最有代表性的是可作为词法数据库的 WordNet<sup>[10]</sup>。WordNet 将英语中的名词、动词、副词和形容词以同义词集合(syn-

onym set,缩写为 synset)的形式组织起来,使得每一个同义词集合可表示一个概念,而同义词集合之间的关系则可反映概念之间的语义关系。文献[11]提出将 WordNet 直接映射为 OWL 本体。文献[12]结合自顶向下和自底向上的方法,试图从 WordNet 中自动抽取概念间的关系,并建立基于 DOLCE<sup>③</sup>的本体。

在文献[11,12]中,没有涉及当前的语境信息对词法含义的影响。例如,指定一个词 match,我们无法确定它所代表的概念是火柴还是比赛,因为其具体含义受到当前所讨论的领域的限制。当然,将 WordNet 整体上作为一个本体的话,对很多的应用来讲会显得过于庞大,影响效率且增加使用难度。在本体集成的相关研究方面,文献[13]将本体的集成分为映射(mapping)、一致化(alignment)和合并(merging)。目前,在本体的集成过程中,同义词的标注大多都依靠手工或词法库的辅助来完成。文献[14]提出了一种借助于 WordNet 来辅助进行同义词标注的方法。文献[15]提出按照词法相似度来辅助完成 OWL 本体的合并。

文献[13]中提出的本体合并方法可以生成一致性的语义基础,但不强调从目标本体到源本体的映射,所以不适合如图 1 所示的跨自治域应用集成的资源访问需求。因为,跨自治域应用的构建需要依赖一个一致性的语义基础,而对自治域资源的最终访问又要依赖于各自自治域的语义。另外,本体映射方法用于维护源本体及其映射关系,不便于构造跨自治域的应用;本体的一致化需要对源本体进行调整,不符合自治域自治性的要求。文献[13]所提出的方法人员干预过多,对操作人员要求过高(要求准确把握领域概念),因此不仅效率受到影响,同时不能很好地适应虚拟组织的动态变化环境;文献[14,15]均不涉及当前的语境信息,因此同义词标注的准确性受到一定的影响。

综上所述,我们认为:在抽取资源的语义时,语义的来源不能只限于资源的描述,需要有词法本体来辅助定位概念和概念间的关系;概念只有处于特定的语境中才会有准确的含义,而且在特定的应用环境中往往只关心其在语义树上的一个分枝;词法空间的

① OWL Web Ontology Language Overview. <http://www.w3.org/TR/owl-features/>.

② Volz R. *et al.* OntoLIFT Prototype. IST Project 2001-33052 WonderWeb Deliverable 11, 2003.

③ DOLCE, 原义指 Descriptive Ontology for Linguistic and Cognitive Engineering.

限定将为概念的抽取明确目标,提高效率.另外,在语义的聚合过程中,同义词的标记要依赖于概念的特定语境;在生成虚拟组织的逻辑资源域的语义时,如果能够隐含其与自治资源域的语义之间的映射,将会更好地支持虚拟组织的应用构建,因为逻辑资源域的语义是构建虚拟组织应用的需要,而语义之间的映射是访问物理资源的需要.

因此,我们将通过对领域知识的学习来提取当前语境相关的词法空间,并结合资源的描述来进行资源语义的抽取;基于语境相关的词法空间完成语义的聚合;将虚拟组织逻辑资源域的语义到各自治资源域的语义的映射隐含于语义的聚合过程之中.

### 3 语义抽取

资源的语义描述通常包含一个概念集和概念关系集.通常情况下,概念的命名方法是取资源描述中的主体词或手工指定名称,显然前者的命名不够准确,而后者效率较低,对人员的要求较高.

本文采用 WordNet 作为统一的词法本体,并通过对领域知识的学习,从中抽取出当前语境相关的同义词集,以辅助概念集和概念间关系的建立.我们将词法本体的上述抽取结果称为当前语境相关的词法空间.

#### 3.1 建立当前语境相关的词法空间

建立当前语境相关的词法空间的基本原理如图 3 所示.其中,领域知识包括自治域当前所涉及的有关命题、术语和文档等.另外,如果当前的自治域是在 VO 一次任务的执行过程中动态纳入的,则领域知识还将包含 VO 的当前语义.要说明的是,当前语境相关的词法空间的建立过程,对各种类型的资源是相同的.

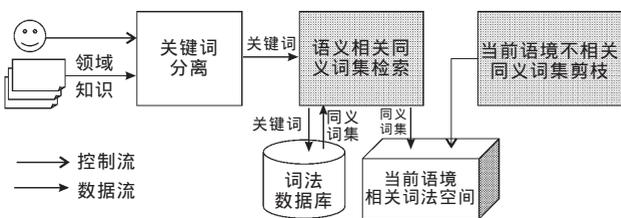


图 3 建立当前语境相关的词法空间示意图

为便于讨论,下面给出 WordNet 中同义词集及同义词集间的基本语义关系的形式化定义.

**定义 1.** 一个词  $term$  的同义词集  $C_i(term)$  表示为  $C_i(term) = \{t | (\exists j)t.sense[j] \equiv term.sense[i]\}$ , 其中,  $t.sense[j] \equiv term.sense[i]$  表示词  $t$  的第  $j$  个

词义和  $term$  的第  $i$  个词义相同.

**定义 2.**  $C(term) = \{C_i(term) | i = 1, 2, \dots, n\}$ , 表示所有与  $term$  相关的同义词集的集合.

**定义 3.**  $Hypernyms(C_1, C_2) ::= C_1$  is a kind of  $C_2$ , 表示同义词集  $C_1$  和  $C_2$  的继承关系(也称为上下位关系).

$Hyponyms(C_1, C_2)$  等价于  $Hypernyms(C_2, C_1)$ .

**定义 4.**  $Holonoms(C_1, C_2) ::= C_1$  is a part of  $C_2$ , 表示同义词集  $C_1$  和  $C_2$  的包含关系(也称为构成关系).

$Meronyms(C_1, C_2)$  等价于  $Holonoms(C_2, C_1)$ .

基于上述定义,我们可以给出如下的定理.

**定理 1.** 若

$(\exists i, j)(C_i(term) \in C(term) \wedge C_j(term_1) \in C(term_1) \wedge R(C_i(term_1), C_j(term)))$ ,

则  $C(term_1)$  与  $C(term)$  语义相关,其中  $R$  表示关系  $Hypernyms$ ,  $Hyponyms$ ,  $Meronyms$  或者  $Holonoms$ .

**证明.** 假设  $R$  为关系  $Hypernyms$ , 给定  $i$  和  $j$ , 如果满足

$C_i(term) \in C(term) \wedge C_j(term_1) \in C(term_1)$  且  $R(C_i(term_1), C_j(term))$ ,

则  $term.sense[j]$  继承  $term_1.sense[i]$ , 所以  $C(term_1)$  与  $C(term)$  语义相关. 当  $R$  表示关系  $Hyponyms$ ,  $Meronyms$  或者  $Holonoms$  时可同样得证. 证毕.

**定理 2.**  $term$  的一个同义词集  $C_i(term)$  被剪枝, 当且仅当满足

$(\forall j)((term \neq term_j) \rightarrow (\forall k)(C_k(term_j) \in C(term_j) \rightarrow \rightarrow R(C_i(term), C_k(term_j))))$  (1)

**证明.** 充分性. 假设  $C_i(term)$  被剪枝, 且  $C_i(term)$  不满足式(1), 则

$(\exists j)((term \neq term_j) \wedge (\exists k)(C_k(term_j) \in C(term_j) \wedge \wedge R(C_i(term), C_k(term_j))))$ .

因此, 给定  $j$  有  $term_j.sense[k]$  与  $term.sense[i]$  在语义上相关, 因此同义词集  $C_i(term)$  与当前语境相关, 不能被剪枝.

必要性. 可进行类似的证明. 证毕.

在此基础上, 我们给出建立当前语境相关的词法空间的算法描述:

1. 初始化并连接 WordNet 词法库;
2. 输入领域知识, 领域知识可表示为  $R(term_1, \dots, term_n)$ , 表示概念  $term_1, \dots, term_n$  之间存在某种关系;
3. 从输入  $R(term_1, \dots, term_n)$  中分离出关键词  $term_1, \dots,$

$term_n$ ;

4. 在 WordNet 中检索每个关键词  $term_i$  的所有同义词集  $C(term_i)$ ;

5. 查找出与每个同义词集  $C(term_i)$  相关的所有同义词集  $C(term_{i1})$ , 即对每个  $C(term_i)$  查找出满足定理 1 的  $C(term_{i1})$ ;

6. 重复上述过程, 直到词法空间不再增长;

7. 利用领域知识所产生的语境信息对同义词集所构成的词法空间进行剪枝, 消除与当前领域不相关的语义分枝, 即对满足定理 2 的同义词集进行剪枝;

8. 算法结束.

当前语境相关词法空间的建立使得抽取语义时概念的捕捉更为准确, 并将概念的语义限定在了特定的语境中; 可帮助建立概念间的关系; 为语义的聚合提供了同义词标注的支撑.

也就是说, 语境相关词法空间为 VO 各自治域提供了一个共同的词法库, 使得语义抽取和聚合更为准确可行. 而语义复用时将 VO 当前语义作为领域知识的一部分, 使得这种方式同样适用于 VO 动态纳入的自治域资源的语义抽取和聚合.

### 3.2 建立本体的概念集

本体的概念集的准确建立是生成本体的基础, 本文中本体概念集的建立依赖于 3.1 节中提到的当前语境相关的词法空间. 本体概念集的建立思想对各种资源类型来讲是相似的 (其基本原理如图 4 所示):

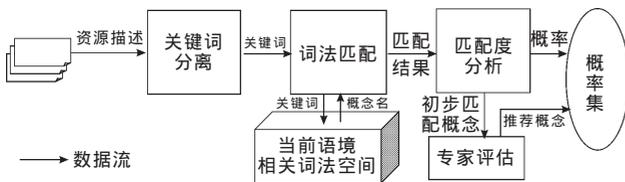


图 4 本体的概念集的建立过程

首先, 对于关系数据库等结构化的数据, 从数据字典中获取关系模式的相关信息; 对于 XML 文档

等半结构化的数据, 获取其 Schema 描述; 对于文本文件, 则依据现有的分词方法, 从当前文本中提取出关键词.

然后, 对于结构化和半结构化的资源, 从资源的描述信息中分离出需要抽象为概念的关键词.

最后, 使用所抽取的关键词, 在当前语境相关的词法空间进行匹配, 并将匹配的关键词加入到本体的概念集中, 以支持本体的生成. 值得注意的是, 关键词可以采用缩写等形式, 所以匹配的结果可能是不完全的. 文献[16]论述了使用 WordNet 且基于词法结构的匹配技术, 本文对此不做研究. 另外, 当领域词法空间中出现一个词的多个同义词集时, 可以将匹配结果提供给领域专家进行评估.

本体概念集中的特定概念与词法空间中的相应同义词集相对应, 因此, 词法空间中同义词集之间的关系可以反映出本体概念集中概念之间的关系. 对于与当前概念有继承、包含等关系的概念, 同样要加入到本体的概念集中, 以丰富所抽取的语义.

### 3.3 生成本体

本文采用 OWL 本体来表示语义的抽取结果. OWL 本体由类、类之间的关系、属性、属性之间的关系、属性约束和实例构成. 属性包括对象属性 (ObjectProperty) 和数据类型属性 (DatatypeProperty), 前者是两个类的实例之间的二元关系; 后者则是类的实例和数据类型之间的二元关系. 下面分析本体的生成过程:

(1) 类的生成. 对于概念集中的每一个概念, 在 OWL 本体中为其生成一个相应的类, 类的名称取概念的名称.

(2) 类之间关系的建立. 根据概念在词法空间中的关系, 在 OWL 本体中建立起相应类之间的关系, 如表 1 所示.

表 1 概念间的关系对照表

概念/类	词法空间中概念间关系	OWL 本体中类间关系	OWL 本体中的表示
$C_1, C_2$	$Hypernyms(C_1, C_2)$	$C_1$ 是 $C_2$ 的子类	SubClassOf
$C_1, C_2$	$Hyponyms(C_1, C_2)$	$C_2$ 是 $C_1$ 的子类	SubClassOf
$C_1, C_2$	$Holonyms(C_1, C_2)$	$C_1$ 和 $C_2$ 的属性联系	ObjectProperty
$C_1, C_2$	$Meronyms(C_1, C_2)$	$C_1$ 和 $C_2$ 的属性联系	ObjectProperty
$C_1, C_2$	属于同一 synset	等价	EquivalentClass
$C_1, C_2, C_3$	$Hypernyms(C_1, C_2)$ 和 $Hypernyms(C_1, C_3)$	复杂类 (类的交)	$C_1 = C_2 \cap C_3$

### (3) 属性的生成

首先, 如果概念  $C_1$  和  $C_2$  之间存在  $Holonyms$  关系, 则意味着每个  $C_1$  的实例与  $C_2$  的一个实例相对应. 因此, 创建对象属性 hasC2, 其定义域 (domain)

和值域 (range) 分别是  $C_1$  和  $C_2$ . 若  $C_1$  和  $C_2$  之间存在关系  $Meronyms$ , 则存在类似的对象属性 hasC1, 两个属性之间存在  $reverseOf$  关系.

其次, 对于关系数据库资源而言, 一个关系模

式往往对应一个概念. 另外, 还需对外键对应的概念构建对象属性  $P_1$ , 其 domain 和 range 分别为参照关系( $r_1$ )和被参照关系( $r_2$ )所对应的类. 属性  $P_1$  的类型为 FunctionalProperty, 其  $minCardinal$  和  $maxCardinal$  均为 1, 表示  $r_1$  的任何一个元组对应的实体必须且只能与一个  $r_2$  的实体发生联系. 同时, 还将存在一个 domain 和 range 分别为  $r_2$  和  $r_1$  对应的类的对象属性  $P_2$ , 其  $minCardinal$  为 0,  $maxCardinal$  为无穷大, 且  $P_1$  和  $P_2$  互为逆属性 (reverseOf).

另外, 由于在关系数据模式中, 每个字段的取值和一个特定的取值域相对应, 因此, 对于在词法空间中不存在、同时在其上又不存在约束的字段, 可在 OWL 本体中为其创建相应的数据类型属性 (DatatypeProperty), 该属性的 domain 和 range 分别为当前的类和取值域对应的数据类型.

如果是半结构化的资源, 如 XML 文档, 则从元素 (element) 之间的构成关系以及元素和属性 (attribute) 之间的关系来辅助定位概念间的关系和属性的属性; 对于文本资源, 则通过句法的分析, 辅助进行概念间的关系和属性的定位. 本文中主要讨论关系数据库资源, 对其余两种资源类型不做详细讨论.

## 4 语义聚合

本文的最终目的在于探索一种支持虚拟组织的语义基础设施的快速构建方法, 而语义聚合是其重要的实现手段. 在本文中, 语义的聚合过程依赖于所构建的当前语境相关的词法空间, 使得同义词的标注更为准确, 并且提高了语义聚合的自动化程度. 前面已经提到, VO 的各自治域在提供资源的同时, 可能会提供其已有的语义描述, 因此, 在建立当前语境相关的词法空间时, 领域知识包括 VO 的当前语义信息、自治域提供的语义信息以及与 VO 当前任务相关的知识. 因此, 这种语义聚合方式支持 VO 语义基础设施的复用.

我们使用 OWL 本体来描述语义, 因此对语义的聚合过程变成了对 OWL 本体的集成过程. 下面给出基于当前领域相关的词法空间, 将两个本体  $O_1$  和  $O_2$  集成本体  $O_3$  的基本算法. 其中,  $C_i$  表示  $O_i$  的概念集;  $P_i$  表示  $O_i$  的属性集;  $I_i$  表示  $O_i$  的实例集.

1. 令  $C_3 = C_1 \cup C_2$ , 符号  $\cup$  表示  $C_3$  从构成范围上是  $C_1$  和  $C_2$  的并集, 但又不是简单地取并集, 该过程中要解决同义

词和一词多义两种语义冲突<sup>[17]</sup>, 具体方法如下:

### 1.1. 标注 $O_1$ 和 $O_2$ 的同义词

1.1.1. 对于任意的  $term \in C_1$  且  $term \in C_2$ , 如果在当前语境相关的词法空间中, 只存在  $term$  的一个同义词集, 则标注本体  $O_1$  的概念  $term$  和本体  $O_2$  的概念  $term$  为同义词;

1.1.2. 对于任意的  $term \in C_1$  且  $term \in C_2$ , 且  $term$  在两个本体中的父类和子类是同义词, 则标注本体  $O_1$  的概念  $term$  和本体  $O_2$  的概念  $term$  为同义词;

1.1.3. 对于任意的  $term_1 \in C_1$  和任意的  $term_2 \in C_2$ , 如果在当前语境相关的词法空间中,  $term_1$  和  $term_2$  属于同一同义词集, 则标注本体  $O_1$  的概念  $term_1$  和  $O_2$  的概念  $term_2$  为同义词;

1.1.4. 对于任意的  $term_1 \in C_1$  和任意的  $term_2 \in C_2$ , 如果两者的属性相同, 则标注本体  $O_1$  的概念  $term_1$  和  $O_2$  的概念  $term_2$  为同义词;

1.1.5. 对于任意的  $term \in C_1$  且  $term \in C_2$ , 如果在当前语境相关的词法空间中, 存在  $term$  的多个同义词集, 且目前无法确定  $term$  是否是同义词, 将结果推荐给领域专家进行标注.

### 1.2. 标注 $O_1$ 和 $O_2$ 中的一词多义词

对于任意的  $term \in C_1$  和  $term \in C_2$  且未确定两者为同义词, 则标注本体  $O_1$  的概念  $term$  和  $O_2$  的概念  $term$  为一词多义.

### 1.3. 在本体中处理同义词和一词多义带来的冲突

1.3.1. 如果  $term$  是  $C_1$  和  $C_2$  的同义词, 则在  $C_3$  中仅仅保留该概念的一个副本; 如果  $term_1 \in C_1$  和  $term_2 \in C_2$  是同义词, 在  $C_3$  中仅标记  $term_1$  和  $term_2$  是等价 (owl: equivalentClass), 在本体的推理过程中, 可获取其带来的知识变化, 而在本体中不需要冗余表示;

1.3.2. 如果  $term \in C_1$  和  $term \in C_2$  是一词多义, 在本体  $O_3$  中生成  $term$  的副本  $term [1]$  和  $term [2]$ , 结果推荐给领域专家进行判定. 若判定结果为同义词, 则标注为同义词, 按照前面的方式将概念  $term$  加入本体  $O_3$ ; 若判定结果为一词多义, 则由领域专家根据两个副本所应属于的同义词集, 分别指定  $term$  在  $O_1$  和  $O_2$  中的概念名称.

1.4. 对于本体  $O_1$  和  $O_2$  的其余概念取并集, 结果加入到  $C_3$  中.

2. 令  $P_3 = P_1 \cup P_2$ , 该过程同样不是简单的取并集, 同时要处理以下的变化:

概念属性的变化. 如果  $term_1$  和  $term_2$  是同义词, 我们在步 1 中已经将其标记为等价的, 但两者的属性刻画可能存在差异, 此时取  $term_1$  和  $term_2$  的属性的并集, 并加入到  $O_3$  中.

由逻辑资源域的语义向各自治资源域的语义进行映射时, 只关心与自治域语义的概念相关的属性, 其余进行舍弃, 显然这样不会影响信息的获取. 这正是我们前面所提到的向上最大化和向下最小化思想的一种体现, 这种思想在语义聚合的其它方面还会得到体现. 显然, 这种思想既保持了各自治域的自治性和支持 VO 应用的构建, 又支持了跨自治域资源的透明访问.

泛化和特殊化关系的变化.概念的增加,使得概念间的关系也发生了改变.通过逻辑运算 $\sqsubset$ 的传递等特性,对 $C_3$ 中的概念之间的泛化和特殊化关系进行调整.

复合概念构造公理的变化.通过 $\cup, \cap, \rightarrow$ 的运算性质,扩展 $C_3$ 中复合概念的构造公理.

基数约束的冲突.在本体 $O_3$ 中,一个属性的基数约束的最大值(最小值)取在本体 $O_1$ 和本体 $O_2$ 中的最大值(最小值).这样做既不会导致集成时的语义丢失,也保证了映射到自治域的语义时,约束值是可恢复的.此处同样应用了向上最大化和向下最小化的思想.

数据类型属性量纲的冲突.我们将概念的取值域的不同和度量标准的不同都归结为量纲的冲突,例如,字符串和数值类型的冲突、温度单位的不同都属于量纲冲突.

如果在本体 $O_1$ 和 $O_2$ 中,概念 $term_1$ 和 $term_2$ 是同义词,且在两个本体中的取值域不同,则在本体 $O_3$ 中选择其中的一个取值域.在本体的映射阶段,当出现值域的冲突时进行自动转换.如果是度量的单位不同,需要给出两种量纲之间的关系<sup>[18]</sup>.

3. 令 $I_3 = I_1 \cup I_2$ ,该过程与步2相类似.本文中主要关心本体中的概念的刻画,而对概念的实例不作详细讨论.

#### 4. 算法结束.

需要说明的是,在本体集成的过程中,可能会产生概念和属性的冗余.例如,如果一个数据类型属性 $DP$ 的 $domain$ 所表示的概念 $C$ 不再作为任何一个对象属性 $OP$ 的 $domain$ ,则认为概念 $C$ 是冗余的,同时,对象属性 $OP$ 也是冗余的.另外,给定两个属性 $P_1(domain_1, range_1)$ 和 $P_2(domain_2, range_2)$ ,如果 $domain_1$ 和 $domain_2$ 是等价的,而且 $range_1$ 和 $range_2$ 是等价的,则认为本体中属性 $P_1$ 或 $P_2$ 可能是冗余的.

另外,对集成结果可以根据描述逻辑的一致性验证算法<sup>[19]</sup>进行证明.限于篇幅,对集成结果的优化和验证,在本文中不做详细讨论.

## 5 实验结果及分析

我们使用Java语言,在Windows操作系统平台上开发了本文的实验原型系统,其运行支撑环境包括WordNet 2.0, wnjn1.6以及jena 2.2工具包和MS SQL Server 2000关系数据库.系统主要包含三个功能模块,即当前语境相关词法空间的创建模块(LSPrunning)、OWL本体的生成模块(OntoExtraction)以及OWL本体的集成模块(OntoIntegration).

实验的场景来源于某体育赛事综合信息服务系统总体规划,并对原场景进行了简化.实验场景的VO中包含三个自治域:分别是比赛组织方( $AD_1$ )、

资讯经纪公司( $AD_2$ )和气象局( $AD_3$ ).我们设定VO最初组建时只包含 $AD_1$ 和 $AD_2$ ,而在任务执行过程中动态纳入 $AD_3$ . $AD_1$ 和 $AD_2$ 分别提供比赛安排数据库Schedule\_DB和运动员历史信息数据库Athlete\_Information\_DB,而 $AD_3$ 在提供资源的同时还提供了其语义描述weather.owl,如图5和代码1所示.

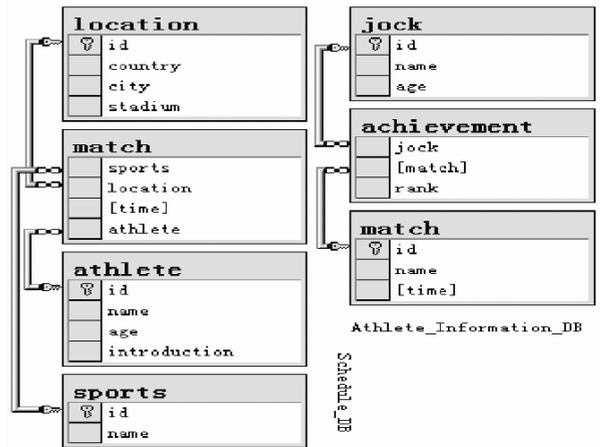


图5  $AD_1$ 和 $AD_2$ 的数据库关系模式

代码1. weather.owl的描述片段.

```
.....
<owl:Class rdf:ID="country"/>
<owl:Class rdf:ID="city"/>
<owl:DatatypeProperty rdf:ID=
    "has_temperature">
    <rdfs:range rdf:resource="xmls# Integer"/>
</owl:DatatypeProperty>
.....
```

实验过程中,对领域知识的表示和预处理过程进行了简化,直接以所选取的领域词汇作为对领域知识进行学习的语料信息,包括match, game, athlete, sports, stadium, location, time, swim, achievement, score, final, semifinal等词汇以及关系模式中的关键词.另外,在本文中对词法匹配<sup>[16]</sup>不做研究,因此实验过程中使用的关键词均为完整的单词.

表2为系统输出的当前语境相关的词法空间与原词法空间的部分对比情况. $TN$ 表示原有词法空间中的语义数; $EN$ 表示当前语境相关的语义数; $\times$ 表示该语义与当前语境无关,相应的同义词集被剪枝; $\checkmark$ 表示该同义词集与当前语境相关.

系统输出的当前语境相关词法空间中match的部分语义链如下:

```
Hypernyms Link: {match} >> {contest, competition} >>
{social_event} >> {event}.
```

Hyponyms Link: { boxing \_\_match &... & field \_\_event  
&... & quarterfinal \_\_final &... &... }=> { match }.

表 2 当前语境相关的词法空间分析

名词	原同义词集	语境相关性	TN	EN
match	match, lucifer	×	9	1
	match >> contest, competition	✓		
	Match >> ighter, light, igniter	×		
	match, mate	×		
	Match >> score	×		
	catch, match	×		
	peer, equal, match, compeer	×		
	couple, mates, match	×		
	Match >> counterpart, ...	×		
athlete	athlete, jock	✓	1	1
rank	Rank >> line	×	5	2
	Rank >> status, position	✓		
	rank and file, rank	×		
	social station, social status, social rank, rank	✓		
	membership, rank	×		
...	...	...	...	...

从表 2 中的结果可以看出,通过对关键词语境相关性的限定,使得关键词的语义比原来更准确,为语义抽取时的概念和概念间的关系标注及语义聚合时的同义词标注提供了很好的基础。

代码 2 为针对  $AD_1$  的资源所生成的本体 Schedule.owl 的描述片段。

#### 代码 2.

```

<owl:Class rdf:ID="contest"/>
<owl:Class rdf:ID="match">
  <rdfs:subClassOf rdf:resource="# contest"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource=
        "# has_sport"/>
      <owl:minCardinality rdf:datatype=
        "&-xsd: nonNegativeInteger">
        </owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
<owl:ObjectProperty rdf:ID="has_sport">
  <rdfs:domain rdf:resource="# match"/>
  <rdfs:range rdf:resource="# sport"/>
</owl:ObjectProperty>
<owl:DatatypeProperty rdf:about="# has_name">
  <rdfs:domain rdf:resource="# athlete"/>
  <rdfs:range rdf:resource="xmls # string"/>
</owl:DatatypeProperty>
...

```

从代码 2 可以看出,因为当前语境相关词法空间的存在,使得语义抽取时概念的标注更为准确。代码 2 中的加黑部分确定了 match 的当前语义。另外生成本体中还包含了一条公理“match is a kind of contest”。

同时,系统为自治域  $AD_2$  生成了本体 Athlete\_Information.owl。

代码 3 为  $AD_1$  和  $AD_2$  对应的本体 Schedule.owl 和 Athlete\_Information.owl 的集成结果片段。

#### 代码 3.

```

<owl:Class rdf:about="http://localhost:8080/
  Schedule.owl# match">
  <rdfs:equivalentClass rdf:resource=
    "http://localhost:8080/
    Athlete_Information.owl # match"/>
</owl:Class>
<owl:Class rdf:about="http://localhost:8080/
  Schedule.owl# Athletet">
  <rdfs:equivalentClass rdf:resource=
    "http://localhost:8080/
    Athlete_Information.owl # jock"/>
</owl:Class>

```

从代码 3 可以看出,系统将两个本体 Schedule.owl 和 Athlete\_Information.owl 中的 match 概念自动标注为同义词。另外, Schedule.owl 中的 match 概念和 Athlete\_Information.owl 中的 jock 概念也被标注为同义词。显然,当前语境相关的词法空间是上述标注过程的保证。

因为同义词所对应概念的属性的合并 in 语义推理的过程中可以体现出来,所以在本体中不需要做冗余的表示。当根据 VO 应用访问物理资源时,自治域本体中不存在的属性将被舍弃、约束条件将被加强,从语义查找的效果来看,该过程隐含了 VO 语义到自治域语义的映射。

当  $AD_3$  被纳入为 VO 的成员自治域时,系统以 VO 的当前语义、 $AD_3$  提供的语义 weather.owl 以及 weather, temperature, humidity, cloudy 等领域词汇作为领域知识,限定当前语境相关的词法空间,并基于该词法空间系统完成了 weather.owl 和 VO 当前本体的集成,集成结果在此不再重复。

新纳入资源的语义抽取以及与当前语义的聚合使得该方法适合于 VO 的动态开放环境,同时也重用了 VO 的当前语义成果。

## 6 应用实例

在某大型体育赛事综合信息服务系统的整体规划中,涉及的资源包括来自赛事、交通、气象、商务、旅游等领域的信息和服务,分别来源于组委会、交通局、气象局、商务局和旅游局等各个资源自治的组织,在面向公众提供赛前、赛时和赛后综合信息服务的任务驱动下,上述组织及其资源被纳入一个虚拟组织中。

综合信息服务系统的原型建立在 VINCA<sup>[20]</sup> 服务网格平台之上。VINCA 服务网格由作者所在的中国科学院计算技术研究所网格中心服务网格研究组自主研制开发,是一整套面向服务架构的软件基础设施,旨在支持业务层面的网络化应用系统的即时构造以及分布式应用的集成和协同。其中, VINCA 服务社区是负责服务、语义、规则等资源管理的基础模块,主要提供服务和语义的注册、服务的查找、服务的虚拟化等功能。可以认为,在本场景中 VINCA 服务社区充当了 VO 的服务和语义资源的公共注册中心。图 6 展示了使用本文的方法在 VINCA 服务网格平台上搭建虚拟组织应用的实例。

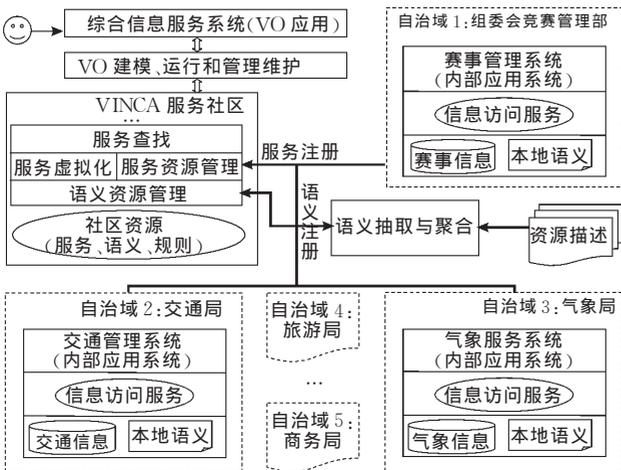


图 6 在 VINCA 服务网格中支撑 VO 应用的实例

如图 6 所示,在构建虚拟组织应用时,从各自自治域的资源描述中自动抽取语义,各自自治域的语义聚合形成虚拟组织应用的语义基础设施。各自自治域提供对物理资源的访问服务,并注册到 VINCA 服务社区。虚拟组织应用依赖聚合所得语义基础设施查找服务,并用以构建业务服务和业务流程,以进一步完成 VO 应用的构建。在应用的执行过程中,被调用服务的语义将被映射为相应自治域的语义,进而完成物理资源的访问。

在该应用实例中,我们以系统整体规划中的用例和资源描述为领域知识,通过本文提出的学习算法,建立了与当前语境相关的词法空间。在赛前运行阶段,VO 根据其资源纳入规则仅纳入旅游、气象、交通、城市文化等自治域的资源,并基于上述语境相关的词法空间通过语义的抽取和聚合构建了 VO 的语义基础设施。在赛事运行阶段,自治域 1 被纳入 VO,以提供赛事信息,此时系统根据 VO 的现有语义基础设施以及自治域 1 的领域知识和资源描述进行其语义的抽取并聚合到 VO 的语义中。在赛后,VO 需要重组,新的语义基础设施将复用原有语义基础设施中的城市文化、商务信息等资源的语义,并将新纳入自治域的资源语义聚合到其中,形成当前 VO 的语义基础设施。

上述的应用实例表明,本文的语义基础设施构建方法减少了人员的参与,并且对 VO 动态开放环境具有很好的适应性,支持对虚拟组织既有语义成果的复用,是构建 VO 应用的有效支撑。

## 7 结 论

实验和应用表明,本文提出的从资源描述中抽取语义并进行聚合的方法,能够适应虚拟组织的动态开放环境,减少语义基础设施构建过程中的人员参与,提高语义基础设施构建的自动化程度,从而有效支持虚拟组织的语义基础设施的动态构建。

在今后的研究工作中,将进一步针对开放环境下的服务资源,研究服务语义的自动抽取,以有效利用各自自治域的服务资源进行跨自治域应用的快速构建。

## 参 考 文 献

- 1 Foster I., Kesselman C., Tuecke S.. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of High Performance Computing Applications*, 2001, 15(3): 200~222
- 2 Uhlig J., Machkova M. *et al.* Creation of architectural ontology: User's experience. In: *Proceedings of the 14th International Workshop on Database and Expert Systems Applications*, DEXA, US, 2003, 65~69
- 3 Bruijn J. D. *et al.* State-of-the-art survey on ontology merging and aligning V1. EKT-project Report D4. 2. 1 (WP4), IST-2003-506826, 2003
- 4 Baader F., Horrocks I., Sattler U.. Description logics as ontology languages for the semantic Web. In: Hutter D., Ste-

- phan W. eds. . Festschrift in Honor of Jorg Siekmann. Lecture Notes in Artificial Intelligence. Berlin Heidelberg, New York: Springer-Verlag, 2003, 228~248
- 5 Stojanovic L. , Stojanovic N. , Volz R. . Migrating data-intensive Web sites into the semantic Web. In: Proceedings of the 17th ACM Symposium on Applied Computing, SAC, 2002, 1100~1107
- 6 Astrova I. . Reverse engineering of relational databases to ontologies. In: Proceedings of the 1st European Semantic Web Symposium, Crete, Greece, 2004, 327~341
- 7 Astrova I. . Extracting ontologies from relational databases. In: Proceedings of the 22 IASTED International Conference on Databases and Applications, Innsbruck, Austria, 2004, 56~61
- 8 Kashyap V. . Design and creation of ontologies for environmental information retrieval. In: Proceedings of the 12th Workshop on Knowledge Acquisition, Modeling and Management, Banff, Canada, 1999, 1~18
- 9 Aussenac-Gilles N. , Biébow B. , Szulman S. . Revisiting ontology design: A methodology based on corpus analysis. In: Proceedings of the 12th International Conference in Knowledge Engineering and Knowledge Management, Juan-les-Pins, France, 2000, 172~188
- 10 Kornilakis H. , Grigoriadou M. *et al.* Using WordNet to support interactive concept map construction. In: Proceedings of the IEEE International Conference on Advanced Learning Technologies, Joensuu, Finland, 2004, 600~604
- 11 He H. , Dayou L. . Learning OWL ontologies from free texts. In: Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, Shanghai, China, 2004, 26~29
- 12 Gangemi A. , Navigli R. , Velardi P. . The OntoWordNet project: Extension and axiomatization of conceptual relations in WordNet. Lecture Notes in Computer Science 2888, 2003, 820~838
- 13 Abels S. , Haak L. , Hahn A. . Identification of common methods used for ontology integration tasks. In: Proceedings of the 1st International Workshop on Interoperability of Heterogeneous Information Systems, Bremen, Germany, 2005, 75~78
- 14 Cañas J. , Valerio A. , Lalinde-Pulido *et al.* Using WordNet for word sense disambiguation to support concept map construction. In: Proceedings of the String Processing and Information Retrieval, Manaus, Brazil, 2003, 350~359
- 15 Richardson B. , Mazlack J. *et al.* Approximate ontology merging for the semantic Web. In: Proceedings of the 23rd International Conference of the North American Fuzzy Information Processing Society Proceedings, Banff, Canada, 2004, 641~646
- 16 Do H. H. , Rahm E. . COMA—A system for flexible combination of schema matching approaches. In: Proceedings of the Very Large Data Bases Conference, Roma, Italy, 2001, 610~621
- 17 Zan C. , O'Brien P. . Domain ontology management environment. In: Proceedings of the 33rd Hawaii International Conference on System Sciences, Maui, Hawaii, 2000, 2964~2972
- 18 Ram S. , Park J. . Semantic conflict resolution ontology (SCROL): An ontology for detecting and resolving data and schema-level semantic conflicts. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(2): 189~202
- 19 Baader F. , Sattler U. . An overview of tableau algorithms for description logics. Studia Logica, 2001, 69: 5~40
- 20 Han Y. , Geng H. *et al.* VINCA—A visual and personalized Business-level composition language for chaining Web-based services. In: Proceedings of the 1st International Conference on Service-Oriented Computing, Lecture Notes in Computer Science 2910, Springer-Verlag, 2003, 165~177



**LIU Chen**, born in 1980, Ph. D. candidate. His re-

**CHEN Wang-Hu**, born in 1973, Ph. D. candidate. His research interests include software integration and service grid, service semantic.

search interests include software integration and service grid, service semantic.

**LI Hou-Fu**, born in 1974, Ph. D. candidate. His research interests include service grid and event-driven architecture.

**WANG Jian-Wu**, born in 1980, Ph. D. candidate. His research interests include software integration and service grid.

## Background

The authors are members of Sino-German Joint Laboratory of Software Integration Technologies group of Institute of Computing Technology, Chinese Academy of Sciences. This paper is primarily supported by the National Natural Science Foundation of China (NSFC) under grant No. 90412005, which involves in the research works about dynamic services

composition including virtual organization in Service Grid. The paper proposes an approach to developing the semantic infrastructure for virtual organizations by extracting semantics from resource descriptions in individual autonomous domains and then fusing the extractions.