ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.2, February 2008, pp.369−378
DOI: 10.3724/SP.J.1001.2008.00369

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于分布RDF(S)模型的信息查询与集成*

李 剑+

(中国科学院 软件研究所 软件工程技术中心,北京 100080)

Information Query and Integration Based on Distributed RDF(S) Model

LI Jian+

(Technology Center of Software Engineering, Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: E-mail: lijian767@hotmail.com

Li J. Information query and integration based on distributed RDF(S) model. *Journal of Software*, 2008,19(2): 369−378. http://www.jos.org.cn/1000-9825/19/369.htm

Abstract: In the cases of Web applications, RDF(S) can be used to semantically describe the distributed information sources in enterprises to make the information retrieval more accurate. In this paper, a distributed RDF(S) model is presented to describe distributed and heterogeneous RDF(S)s. Based on this model, a method to query distributed RDF(S) descriptions is also presented, which can retrieve data of concepts as well as that of instances. By this method, users can retrieve RDF(S) descriptions and the information they describe in the same way and implement distributed RDF(S)s' integration.

Key words: RDF (resource description framework); RDF schema; distributed first-order predicate logic; distributed information integration

摘 要: 在Web应用环境中,可以通过RDF(S)形式描述企业领域内分布信息资源的语义,以提高信息查询的准确性.提出了描述分布异构RDF(S)的分布RDF(S)模型,并基于这一模型给出了实现分布RDF(S)查询的方法,此查询方法既能实现实例层次的查询,也能实现概念层次的查询.基于这一方法,用户能够以统一的形式来查询,获取相关的信息资源,同时还可以实现分布RDF(S)的集成.

关键词: RDF(resource description framework); RDF schema;分布一阶谓词逻辑;分布信息集成

中图法分类号: TP311 **文献标识码**: A

Web应用环境中包含大量的分布信息资源,即使在一个企业应用环境内部,这些信息资源也是相对独立地构建的,它们缺乏统一的语义描述,导致难以准确获取这些信息.通过对这些分布信息赋予确定的可形式理解的语义,可以提高信息查询的准确性.

W3C提出了一种元数据模型RDF(resource description framework)及其模式语言RDF Schema,统称为RDF(S).用它们来描述Web信息,将这些Web信息赋予确定的语义.同时,通过RDF(S)语言还能描述概念实体以及实体之间的关系.总而言之,RDF(S)提供了分布Web环境中对各种信息资源的一种统一的语义描述方式.

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2004AA112010 (国家高技术研究发展计划(863)); the National Basic Research Program of China under Grant No.2002CB312005 (国家重点基础研究发展计划(973))

Received 2006-07-24; Accepted 2006-10-10

在RDF模型中,现实世界被描绘并解释成一系列资源(resource)组成的集合,每个RDF语句可以表示为一个三元组⟨*subject*,*predicate*,*object*⟩,它表示 *subject* 对应资源的 *predicate* 属性的值是 *object* 对应资源或值.在 RDF 基础上,RDF Schema 通过一系列有确定语义的词汇(例如"rdfs:Class")来描述概念层次的语义,在其中可以确切定义类概念和属性概念的语义,RDF Schema 本身也是采用 RDF 语法来进行描述的.我们采用 RDF(S)来描述信息内容,就可以为它们赋予确定的语义,从而实现准确的信息获取.

在分布 Web 环境中,每个本地用户可以相对独立地用 RDF(S)来描述自己感兴趣的领域内的信息,因此,不同用户描述信息的 RDF(S)之间可能存在着语义上的差异性.如何建立它们之间的联系,从而实现分布异构RDF(S)之间的信息关联性,这是分布式信息查询和集成所必须要解决的问题.本文提出的分布RDF(S)模型就描述了各个局域RDF(S)之间的联系,根据这些联系可以实现分布RDF(S)查询以及分布RDF(S)集成.

1 局域RDF(S)模型

为了表示多个局域RDF(S)描述之间的关系,我们首先给出将局域RDF(S)描述及其解释转换到一阶谓词公式及其解释的方法.

1.1 局域RDF(S)的描述及解释

RDF(S)中可以同时包含实例层次和概念层次的描述:实例层次的描述表示实例资源的语义;概念层次的描述定义类概念、类概念之间的关系以及属性概念的相关信息;同时,其中还可以定义实例层次与概念层次之间的实例-类从属关系.RDF(S)可以为所描述的信息赋予语义,同时,RDF(S)描述自身也包含着可供查询的信息.

图1表示了3个局域RDF(S)描述,它们描述了大学领域的信息模型.它们是相对独立地构建的,存在着语义上的异质性,并且它们之间存在着内在联系.图 1(a)的 RDF(S)1 将一个大学主页和一个教授的个人主页分别看作是学校和教授实例,并描述了学校的名字以及教授名字等信息,这些都是实例层次的描述.图 1(c)的 RDF(S)3 描述了大学部分领域中的类概念以及类概念之间的关系(例如子类关系 *subclass*),其中包括学校、人员等类定义,同时描述了属性概念的定义域(*dom*)和值域(*range*),这些都是概念层次的描述.图 1(b)的 RDF(S)2 描述了教授、研究生这些类概念以及类概念的实例,同时定义了实例所从属的类(通过 *type* 实例-类从属关系定义),它同时包括了概念层次和实例层次的描述.

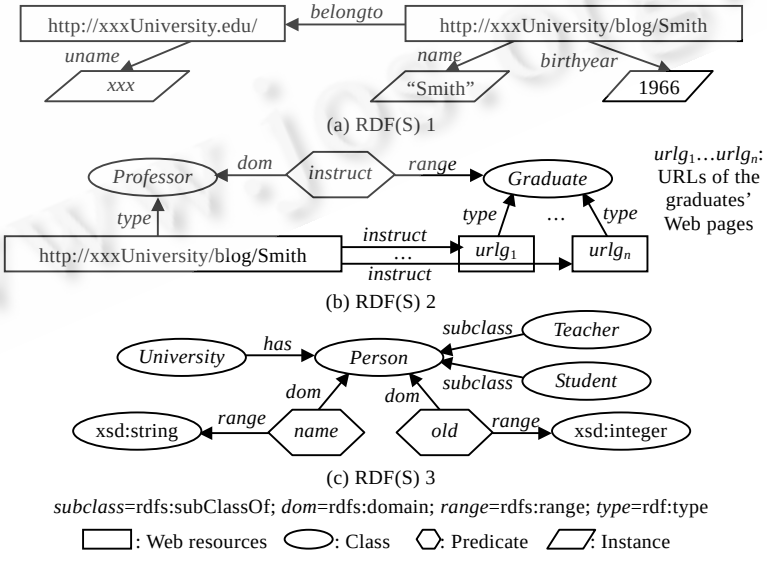

Fig.1 Three distributed RDF(S) descriptions

图1 3个分布的RDF(S)描述

定义 1. 一个局域 RDF(S)描述 *rdf* 可以表示为一个二元组$\langle A,T\rangle$,A 为 *rdf* 中的所有符号集合,T 为 *rdf* 所有语句集合;每个语句$\tau(\tau\in T)$为一个三元组,$\tau=\langle subject,predicate,object\rangle$,其中 $subject,predicate,object\in A$.

定义 2. 对局域 RDF(S)描述 *rdf*(*rdf*=$\langle A,T\rangle$)的解释 $\bar{I}$ 为一个六元组:$\langle DR,DP,IR,IP,IEXT,ICEXT\rangle$,其中:

DR 为 *rdf* 所描述的资源集合,资源集合包含多种类型的值,其中有实例值(例如字符串或者数字值),有 URL 对应的 Web 资源,同时还包括类概念和属性概念.

DP 为属性概念的集合,$DP\subseteq DR$.

$IR:A\to DR$,将 *rdf* 中的符号 $a(a\in A)$解释为一个对应资源.例如,图 1(a)中的表示网页 Web 资源的 URL: "http://xxxUniversity/blog/Smith"和"http://xxxUniversity.edu/",被分别解释为网页表示的个人 P_{Smith} 和学校 U_{xxx},对应网页包含描述它们的信息.

$IP:A\to DP$,将 *rdf* 中的描述属性的符号 $a(a\in A)$解释为一个对应的属性.例如,图 1(a)中的符号"*name*"被解释成名为 *name* 的属性概念.

$IEXT:DP\to DR\times DR$,将属性解释为二元向量集合,向量的两个资源分量之间满足对应的属性关系,$IEXT(p)=\{\langle d_i,d_j\rangle|d_i,d_j\in DR,d_i,d_j$ 满足 p 属性关系(从 RDF(S)描述可以推导出$\langle d_i,p,d_j\rangle$成立)$\}$.例如,图 1(a)中$\langle P_{Smith}$,"Smith"$\rangle\in IEXT(name)$,它表示 P_{Smith} 的 *name* 属性的值为"Smith".

$ICEXT:DR\to\rho(DR)$(ρ表示幂集操作),将类概念资源解释成此类所包含的实例资源的集合,$ICEXT(c)=\{d|d\in DR,d$ 是 c 类的实例$\}$.例如,图 1(b)中的 $ICEXT(Graduate)=\{Gurlg_1,\ldots,Gurlg_n\}$,$Gurlg_1,\ldots,Gurlg_n$ 分别为 $urlg_1,\ldots,urlg_n$ 网页所描述的研究生.

下文中用 $\bar{I}.xxx$ 表示 $\bar{I}$ 解释六元组中的各个部分:$\bar{I}.DR=DR,\ldots,\bar{I}.ICEXT=ICEXT$.

在 RDF(S)的解释中,资源之间存在着一些推导性的关系.例如,如果 $d\in ICEXT(c_1)$(d 是 c_1 类的实例),并且$\langle c_1,c_2\rangle\in IEXT(IR($"rdfs:subClassOf"$))$($c_1$ 是 c_2 的子类),则可以推导出 $d\in ICEXT(c_2)$(d 是 c_2 类的实例).上述推导关系在文献[1,2]中有详细的描述,这里不再赘述.

我们假设任何对局域 RDF(S)描述 *rdf* 的一致解释都将 *rdf* 中相同的符号解释为同一个概念,对于表示 Web 资源的 URL 以及表示数值的符号,其解释都应该相同;对于表示类和属性的符号,假设任何一致解释都将其解释为同一个类概念和属性概念;但是对于不同的一致解释,类概念的实例值集合可能不同.

如果 RDF(S)描述 *rdf* 的一个解释 $\bar{I}$ 满足上述的限制关系,则称 $\bar{I}$ 对 *rdf* 是一致的解释,下文中涉及的对 RDF(S)描述的解释 $\bar{I}$ 都默认是一致的解释.

1.2 局域RDF(S)的逻辑表示

根据 RDF(S)描述 *rdf* 及其解释 $\bar{I}$ 的定义,我们可以将其转换成对应的一阶谓词公式集合及对一阶谓词公式的解释.RDF(S)对应的一阶谓词中存在着下述形式的谓词表示:

(1) $tri(s,p,o).tri(s,p,o)$成立表示 s,p,o 资源之间满足 RDF(S)中的三元组关系$\langle s,p,o\rangle$;

(2) $class(c,i).class(c,i)$成立表示资源 i 是资源 c 类的一个实例;

(3) $relation(p,s,o).relation(p,s,o)$成立表示资源 s 和资源 o 之间满足 p 属性关系;

(4) $x=d$,表示资源等价关系.

定义 3. 一个局域 RDF(S)描述 *rdf*(*rdf*=$\langle A,T\rangle$)对应的一阶谓词形式的局域语言 L 是包含如下一阶谓词语句的集合:

$tri(s,p,o)$,如果对于每个语句$\tau\in T,\tau=\langle subject,predicate,object\rangle$,$s,p,o$ 分别为 $subject,predicate,object$ 在 *rdf* 的任何解释 $\bar{I}$ 下对应的资源,或者对于 *rdf* 的任何解释 $\bar{I}$,都有$\langle s,o\rangle\in\bar{I}.IEXT(p)$;

$class(c,i)$,如果在 *rdf* 的任何解释 $\bar{I}$ 中,都有 $i\in\bar{I}.ICEXT(c)$;

$relation(p,s,o)$,如果在 *rdf* 的任何解释 $\bar{I}$ 中,都有$\langle s,o\rangle\in\bar{I}.IEXT(p)$.

其中,*tri* 语句表示 *rdf* 描述中包含的三元组关系以及可以从 *rdf* 推导出的三元组关系,*class* 和 *relation* 语句分别表示从 *rdf* 描述中能够推导出来的实例-类从属关系以及属性关系.

定义 4. 对于 RDF(S)描述 *rdf* 的解释 $\bar{I},\bar{I}$ 对应的一阶谓词的解释 I 表示为$\langle dom,C^I,P^I\rangle$结构形式,其中:

dom 表示解释 I 的域,$dom=\bar{I}.DR$;C^I 函数表示对常量的解释,$\forall u\in \bar{I}.DR, C^I(u)=u$;

P^I 函数表示对谓词的解释:$P^I(tri)=\{\langle s,p,o\rangle|\forall s,p,o\in \bar{I}.DR,\langle s,o\rangle\in \bar{I}.IEXT(p)\}$;

$$P^I(class)=\{\langle c,i\rangle|\forall c,i\in \bar{I}.DR, i\in \bar{I}.ICEXT(c)\};$$

$$P^I(relation)=\{\langle p,s,o\rangle|\forall s,p,o\in \bar{I}.DR,\langle s,o\rangle\in \bar{I}.IEXT(p)\};$$

$$P^I(=)=\{\langle u,u\rangle|\forall u\in \bar{I}.DR\}.$$

下文中属于解释 I 的 3 个组成部分 dom,C^I,P^I 分别表示为 $I.dom,I.C^I,I.P^I$.

定义 5. 对于由 RDF(S)对应谓词组成的一阶谓词语句 ls,如果 ls 在 I 解释下成立,那么称 I 满足 ls,表示为 $I|=ls$.如果对于局域 RDF(S)描述对应的局域语言 L 中所有的语句 ls,都有 $I|=ls$,那么称 I 满足 L,表示为 $I|=L$.

由上面的定义可知,如果局域 RDF(S)描述 rdf 对应的局域语言为 L,对于任何 rdf 的解释 $\bar{I}$,设 $\bar{I}$ 对应的一阶谓词解释为 I,那么 I 为满足 L 的一个解释($I|=L$).

定义 6. 对于对应局域 RDF(S)描述的局域语言 L 和一阶谓词语句 ls,如果对于任何满足 L 的解释 $I(I|=L)$都有 $I|=ls$,则 ls 是语言 L 的逻辑推论,表示为 $L|=ls$.

定义 7. 对于对应局域 RDF(S)描述的局域语言 L,$q(x_1,\ldots,x_n)\leftarrow le$ 表示对应 L 的一个局域 RDF(S)查询,le 为由 RDF(S)对应的谓词构成的一阶谓词公式,$x_1,\ldots,x_n$ 为 le 中的变元,则此局域查询的结果为

$$LQR(L,le)=\{\langle d_1,\ldots,d_n\rangle|\forall a, L|=le[a], a \text{ 为对 } le \text{ 中变元的赋值,其中,} a(x_1)=d_1,\ldots,a(x_n)=d_n\}.$$

对于局域 RDF(S)描述的查询可以看作是对 RDF(S)图以及其推导图的查询,推导图中包含了根据 RDF(S)描述已有的关系推导出来的关系.Gutierrez 在文献[3]中对此进行了详细描述.

局域查询只能查询到局域 RDF(S)描述中符号的解释值,这些解释值在任何满足局域语言 L 的解释 I 下,通过赋值 a 使 $le[a]$成立,但是它不能查询分布环境下所有满足条件的值.例如,设图 1 中 RDF(S)2 描述对应局域 RDF(S)描述 rdf_2 和局域语言 L_2,rdf_2 中“http://xxxUniversity/blog/Smith”对应解释为 P_{Smith}.对于任何满足 L_2 的解释 I_2,$P_{Smith}\in I_2.dom$,I_2 满足 $class(Professor,P_{Smith})$(表示 P_{Smith} 属于 $Professor$ 类概念),所以,$\langle P_{Smith}\rangle$是局域查询“$q(x)\leftarrow class(Professor,x)$”的一个查询结果.而对于图 1 中 RDF(S)3 对应的 rdf_3 及局域语言 L_3,rdf_3 只描述了概念层次的实体,所以,查询“$q(x)\leftarrow class(Teacher,x)$”对应的局域查询结果为空.在分布环境中,实际上 P_{Smith} 也是 $Teacher$ 类的一个实例,存在一个满足 L_3 的解释 I_3,其中,$P_{Smith}\in I_3.dom,\langle Teacher,P_{Smith}\rangle\in I_3.P^I(class)$.

2 分布RDF(S)模型

为了实现 Web 环境下分布异构 RDF(S)描述之间以及它们所描述的 Web 资源之间的信息关联性,需要有一种统一的模型来描述它们.借鉴 Ghidini 提出的局域模型理论[4]和分布逻辑模型[5]中的模型间关联方式,我们提出用于表示分布 RDF(S)描述之间语义关联的分布 RDF(S)模型.

定义 8. 设两个局域 RDF(S)描述 rdf_i,rdf_j 对应的局域语言为 L_i,L_j,I_i,I_j 分别为满足 L_i,L_j 的一个解释,r_{ij} 为 I_i,I_j 解释域 $I_i.dom$ 和 $I_j.dom$ 中值之间的对应关系,$r_{ij}\subseteq I_i.dom\times I_j.dom$,若$\langle d_i,d_j\rangle\in r_{ij}$,则 $d_i\in I_i.dom, d_j\in I_j.dom$.

定义 9. 对于两个局域 RDF(S)描述 rdf_i,rdf_j 对应的局域语言 L_i,L_j,I_i,I_j 分别为满足 L_i,L_j 的一个解释,当满足下述条件时,对应关系 r_{ij} 是一个一致的对应关系:

(1) 若$\langle ic_i,ic_j\rangle\in r_{ij},\langle c_i,c_j\rangle\in r_{ij}$,则$\langle c_i,ic_i\rangle\in I_i.P^I(class)\Leftrightarrow\langle c_j,ic_j\rangle\in I_j.P^I(class)$;

(2) 若$\langle p_i,p_j\rangle\in r_{ij},\langle s_i,s_j\rangle\in r_{ij},\langle o_i,o_j\rangle\in r_{ij}$,则$\langle s_i,p_i,o_i\rangle\in I_i.P^I(tri)\Leftrightarrow\langle s_j,p_j,o_j\rangle\in I_j.P^I(tri)$,

$$\langle p_i,s_i,o_i\rangle\in I_i.P^I(relation)\Leftrightarrow\langle p_j,s_j,o_j\rangle\in I_j.P^I(relation).$$

其中,条件(1)表示如果一个局域内类资源和实例资源有从属关系,那么它们对应另一个局域内的类和实例也有从属关系;条件(2)表示如果一个局域中的 3 个资源满足 RDF(S)三元组关系,那么它们分别对应的另一个局域内的 3 个资源也满足三元组关系.下文的分布 RDF(S)模型中所涉及的 r_{ij} 对应关系默认都是一致的对应关系.

r_{ij} 通过构建两个局域 RDF(S)描述中实体之间的对应关系来表示局域 RDF(S)之间的联系.r_{ij} 对应关系的语义可以是多方面的:它可以是不同局域 RDF(S)中异名同义实体之间的对应(这样表示便于解决不同局域之间的异名同义冲突),也可以是其他关联关系(例如子类关系)所导致的对应.

值得注意的是,r_{ij} 并不一定是等价关系.例如,设满足图 1 中 RDF(S)2,RDF(S)3 对应局域语言 L_2,L_3 的一个解释分别为 I_2,I_3,并且 r_{23} 为 I_2,I_3 之间的一个一致对应关系,I_2 中的 *Professor* 类可以对应 I_3 中的 *Teacher* 类,$\langle Professor,Teacher\rangle\in r_{23}$,但是这两个类不是等价的.同时,$\langle d_i,d_j\rangle\in r_{ij}\rightarrow\langle d_j,d_i\rangle\in r_{ji}$ 也并不一定成立,例如$\langle Teacher,Professor\rangle\notin r_{32}$,因为有些属于 *Teacher* 类的教师并不是教授.

假设全局中存在的 n 个分布局域 RDF(S)描述 rdf_i,每个 rdf_i 对应的局域语言为 L_i,令 In 为包含 n 个元素的索引序号集合,$i\in In$,这些索引序号分别对应这 n 个局域 RDF(S)描述,那么,这 n 个局域语言构成的对应分布 RDF(S)描述的分布语言集合为$\{L_i\}_{i\in In}$.

定义 10. 对于对应分布 RDF(S)描述的分布语言集合$\{L_i\}_{i\in In}$,$I_i(i\in In)$分别为满足 L_i 的一个解释$(I_i|=L_i)$,设 $r_{ij}(i,j\in In,i\neq j)$为 I_i 和 I_j 之间存在的一致对应关系,则分布解释 $I_{FRDF}=\langle\{I_i\}_{i\in In},\{r_{ij}\}_{i,j\in In,i\neq j}\rangle$,并且分布解释 I_{FRDF} 满足$\{L_i\}_{i\in In}$,表示为 $I_{FRDF}|=_d\{L_i\}_{i\in In}$("$|=_d$"表示分布逻辑下的满足关系,下同).

定义 11. 对于任何 $r_{lm},r_{mn},r_{ln}\in\{r_{ij}\}_{i,j\in In,i\neq j}$,如果对于所有$\langle d_l,d_m\rangle\in r_{lm}$,$\langle d_m,d_n\rangle\in r_{mn}$,都有$\langle d_l,d_n\rangle\in r_{ln}$,那么称 $\{r_{ij}\}_{i,j\in In,i\neq j}$ 是传递性的.

下文中用于表示分布 RDF(S)描述资源之间对应关系的$\{r_{ij}\}_{i,j\in In,i\neq j}$ 都是传递性的.

定义 12. 对应分布语言集合$\{L_i\}_{i\in In}$ 的分布一阶谓词公式 e(设仅用与联结词"∧"连接)可以表示为

$$e::=i:\psi(d_1,\ldots,d_n)|i:\psi(x_1,\ldots,x_n)|e\wedge e,$$

其中,$i\in In$,$i:\psi$为属于局域 i 的局域一阶谓词公式,$d_1,\ldots,d_n$ 和 $x_1,\ldots,x_n$ 分别为常量和变元.

定义 13. 设 I_{FRDF} 为满足$\{L_i\}_{i\in In}$ 的一个分布解释$(I_{FRDF}=\langle\{I_i\}_{i\in In},\{r_{ij}\}_{i,j\in In,i\neq j}\rangle,I_{FRDF}|=_d\{L_i\}_{i\in In})$,对于分布一阶谓词公式 e,设 fa 是一个赋值函数的集合:$fa=\{a_i\}_{i\in J}$,$J\subseteq In$,J 为所有 e 中涉及到的局域索引,a_i 是对应 e 中 i 局域谓词公式 ψ的赋值函数:$a_i(x_1)=d_1\wedge\ldots\wedge a_i(x_n)=d_n,d_1,\ldots,d_n\in I_i.dom$,并且对于所有 $j\in J,j\neq i$,假设 e 中 i 局域谓词公式 ψ和 j 局域谓词公式 ϕ中同时出现的变元为 $x_1,\ldots,x_m$,则在对 j 局域谓词公式 ϕ中的变元进行赋值的函数 a_j 中,对于每个 $k(1\leq k\leq m)$,都有$\langle a_j(x_k),a_i(x_k)\rangle\in r_{ji}$,其他不同时出现的变元则可以单独赋值.

令 $e[fa]$表示将 e 中的变元用 fa 赋值函数集合进行赋值的结果.当特定的条件成立时,分布解释 I_{FRDF} 满足对应形式的分布一阶谓词公式 e'(表示为 $I_{FRDF}|=_d e'$,"$|=_d$"表示分布逻辑下的满足),其具体描述如下:

(1) $I_{FRDF}|=_d i:\psi(d_1,\ldots,d_n)$,当 $I_i|=\psi(d_1,\ldots,d_n)$时,$I_i$ 是 I_{FRDF} 中 i 局域的解释,下同;

(2) $I_{FRDF}|=_d i:\psi[fa]$,当 $I_i|=\psi[a_i]$时,a_i 赋值函数属于 fa 赋值函数集合,$\psi[a_i]$表示对ψ中变元根据 a_i 赋值函数赋值;

(3) $I_{FRDF}|=_d(e_1\wedge e_2)[fa]$,当 $I_{FRDF}|=_d e_1[fa]$并且 $I_{FRDF}|=_d e_2[fa]$时.

上述条件(1)~条件(3)给出了分布一阶谓词公式的满足性条件和判断方法.其中,赋值函数集合 fa 表示对于不同局域的 $i:\psi$,$j:\phi$局域一阶谓词公式之间出现的公有变元 x,按照 r_{ji} 中的$\langle d_j,d_i\rangle(\langle d_j,d_i\rangle\in r_{ji})$对应关系,分别赋予它们对应的值.例如,根据图 1 中的分布 RDF(S)描述,对于分布谓词公式 $3:tri(University,has,x)\wedge 2:tri(x,instruct,Graduate)$,如果包含满足 L_2,L_3 的解释 I_2,I_3 的 I_{FRDF} 中存在对应关系$\langle Professor,Teacher\rangle\in r_{23}$,则可以将 2 和 3 中的 x 变元分别赋值为 *Professor* 和 *Teacher*,且 $I_{FRDF}|=_d 3:tri(University,has,Teacher)\wedge 2:tri(Professor,instruct,Graduate)$.

用户更倾向于用关系形式而不是对应关系 r_{ij} 来表示分布 RDF(S)描述之间的联系,根据 RDF(S)描述对应的一阶谓词形式,我们可以定义分布 RDF(S)描述之间的关联关系.

定义 14. 对于分布 RDF(S)描述对应的分布语言集合$\{L_i\}_{i\in In}$,$b_1:\phi_1,\ldots,b_m:\phi_m$ 以及 $h:\psi(x_1,\ldots,x_n)$分别为 $b_1,\ldots,b_m,h$ 索引局域$(b_1,\ldots,b_m,h\in In)$的局域谓词公式,并且$\phi_1,\ldots,\phi_m$ 中所有变元的集合包含$\{x_1,\ldots,x_n\}$.那么,分布语言集合$\{L_i\}_{i\in In}$ 中局域语言之间的一个关联关系 γ可以表示为两种形式:

①型 γ: $b:x=d\xrightarrow{\wedge}h:x=d'$,其中,$b,h\in In,b\neq h,d$ 和 d'为表示资源的常量.

②型 γ: $b_1:\phi_1\wedge\ldots\wedge b_m:\phi_m\rightarrow h:\psi(x_1,\ldots,x_n)$,其中,$b_1,\ldots,b_m,h\in In$,并且 $H(\gamma)=h:\psi(x_1,\ldots,x_n)$,$H(\gamma)$称为 γ的头部;$B(\gamma)=b_1:\phi_1\wedge\ldots\wedge b_m:\phi_m$,$B(\gamma)$称为 γ的体部.

①型关联关系表示 i,j 局域 RDF(S)描述的资源 d,d'之间的对应关系,$\langle d,d'\rangle\in r_{ij}$.②型关联关系表示如果有资源 $d_1,\ldots,d_k$(对应 $B(\gamma)$中出现的变元),它们使 $B(\gamma)$成立(通过对应赋值,它们满足 $B(\gamma)$表示的条件),那么它们中的

$d_1,\ldots,d_n$(对应 $x_1,\ldots,x_n$)使在 h 局域中 $\psi(d_1,\ldots,d_n)$也成立.

定义 15. 如果满足分布语言集合$\{L_i\}_{i\in In}$的一个分布解释为$I_{FRDF}(I_{FRDF}=\langle\{I_i\}_{i\in In},\{r_{ij}\}_{i,j\in In,i\neq j}\rangle,I_{FRDF}|=_d\{L_i\}_{i\in In})$,那么,对于其局域语言之间的一个关联关系$\gamma$:

(1) 当γ是①型关联关系时,设γ为 $b{:}x=d \hat{\rightarrow} h{:}x=d'$,如果 I_{FRDF} 的$\{r_{ij}\}_{i,j\in In,i\neq j}$ 中满足条件:$d\in I_b.dom,d'\in I_h.dom$,并且$\langle d,d'\rangle\in r_{bh}$,则 I_{FRDF} 满足γ,表示为 $I_{FRDF}|=_d\gamma$;

(2) 当γ是②型关联关系时,设γ为 $b_1{:}\phi_1\wedge\ldots\wedge b_m{:}\phi_m\rightarrow h{:}\psi(x_1,\ldots,x_n)$,对于任何使 $I_{FRDF}|=_d(b_1{:}\phi_1\wedge\ldots\wedge b_m{:}\phi_m)[fa]$成立的赋值函数集合 fa,以及所有 fa 中对 $x_1,\ldots,x_n$ 的赋值 $a_k(x_1)=d_1,\ldots,a_p(x_n)=d_n$,其中,$k,\ldots,p\in In$,都有$\langle d_1,d_1\rangle\in r_{kh},\ldots,\langle d_n,d_n\rangle\in r_{ph}$,并且 $d_1,\ldots,d_n\in I_h.dom,I_h|=\psi(d_1,\ldots,d_n)$,则 I_{FRDF} 满足γ,表示为 $I_{FRDF}|=_d\gamma$.

设Γ为γ组成的集合,如果对每个$\gamma,\gamma\in\Gamma$,都有 $I_{FRDF}|=_d\gamma$,则 I_{FRDF} 满足Γ,表示为 $I_{FRDF}|=_d\Gamma$.

我们可以通过关联关系表示局域RDF(S)之间的联系,满足关联关系γ的分布解释 I_{FRDF} 的$\{r_{ij}\}_{i,j\in In,i\neq j}$中存在着相应的资源对应关系.例如,图 1 中描述的各个分布 RDF(S)之间存在着下述几个关联关系的例子:

(1) 通过关联关系 $2{:}x=Professor \hat{\rightarrow} 3{:}x=Teacher$ 和 $2{:}x=Graduate \hat{\rightarrow} 3{:}x=Student$,我们可以将 RDF(S)2, RDF(S)3 描述的 *Professor*,*Graduate* 类分别与 *Teacher*,*Student* 类对应起来,满足此关联关系的 I_{FRDF} 中$\langle Professor,Teacher\rangle\in r_{23},\langle Graduate,Student\rangle\in r_{23}$.通过关联关系 $1{:}x=P_{Smith} \hat{\rightarrow} 2{:}x=P_{Smith}$,我们可以将 RDF(S)1, RDF(S)2 描述中 Web 资源对应的实例概念对应起来,它们是用来表示 Smith 教授的同一网页.

(2) 通过关联关系 $1{:}x=U_{xxx}\rightarrow 3{:}class(University,x)$($U_{xxx}$=http://xxxUniversity.edu/,它表示 *University* 类的实例 *xxx* 大学),满足此关联关系的 I_{FRDF} 中,$U_{xxx}\in I_3.dom,\langle U_{xxx},U_{xxx}\rangle\in r_{13}$,并且$\langle University,U_{xxx}\rangle\in I_3.P^I(class)$.在 RDF(S)3 局域描述对应的 L_3 中不包含这一实例的描述,而满足此关联关系的 I_{FRDF} 中,满足 L_3 的解释 I_3 包含这一类从属信息,因此,根据此关联关系对 $3{:}class(University,x)$进行分布查询,应该可以查到此信息.

(3) 通过关联关系 $1{:}relation(birthyear,x,NOWYEAR–y)\rightarrow 3{:}relation(old,x,y)$,我们可以将个人($x$)的年龄($y$)与其出生年份联系起来(*NOWYEAR*–$y$,*NOWYEAR* 表示当前年份),在满足此关联关系的 I_{FRDF} 中,$\langle P_{Smith},P_{Smith}\rangle\in r_{13}$,$\langle 40,40\rangle\in r_{13}$,同时,$\langle old,P_{Smith},40\rangle\in I_3.P^I(relation)$,从而提供了 P_{Smith} 的年龄信息描述.

(4) 关联关系 $1{:}relation(belongto,x,y)\wedge 2{:}relation(instruct,x,z)\rightarrow 3{:}relation(has,y,z)$表示如果教授 x 属于学校 y,那么他所指导的研究生 z 也属于学校 y,在满足此关联关系的 I_{FRDF} 中,$\langle P_{Smith},P_{Smith}\rangle\in r_{13},\langle U_{xxx},U_{xxx}\rangle\in r_{12}$,$\langle G_{urlgi},G_{urlgi}\rangle\in r_{23}$($G_{urlgi}$为对应研究生网页描述的研究生).关联关系 $1{:}relation(belongto,x,y)\rightarrow 3{:}relation(has,y,x)$表示 *belongto* 属性关系实际上是 *has* 属性关系的逆向属性关系.

根据已有的关联关系可以推导出一些关联关系,例如,对于类概念之间的对应关系 $2{:}x=Professor \hat{\rightarrow} 3{:}x=Teacher$,我们可以推导出关联关系:$2{:}class(Professor,x)\rightarrow 3{:}class(Teacher,x)$,它表示如果一个教师实例属于 *Professor* 类,那么他也属于 *Teacher* 类;根据 RDF(S)定义,②型的关联关系中的谓词(除了"="以外)都可以转换为 *tri* 谓词,例如将 $class(x,y)$转换为 $tri(y,type,x)$、$relation(x,y,z)$转换为 $tri(y,x,z)$,在此不详细加以描述.关联关系集合Γ同时包含用户定义的关联关系以及由此推导出的关联关系.

根据关联关系集合Γ中的所有关联关系γ,通过构建γ体部 $B(\gamma)$中的局域谓词表达式(将其作为图结点)到γ头部 $H(\gamma)$的局域谓词表达式的有向连接边,我们可以构建Γ的谓词依赖图.如果此依赖图中不存在闭环路径,则称Γ是非循环依赖的.本文中的Γ都默认是非循环依赖的.

定义 16. 对于分布语言集合$\{L_i\}_{i\in In}$ 和包含 L_i 分布语言间关系的关联关系集合Γ,如果对所有满足$\{L_i\}_{i\in In}$并满足Γ的分布解释 $I_{FRDF}(I_{FRDF}|=_d\{L_i\}_{i\in In}$并且 $I_{FRDF}|=_d\Gamma)$,都有 $I_{FRDF}|=_dgs$,gs 为根据定义 12 定义的分布一阶谓词语句(不包含变元的分布一阶谓词公式),那么,gs 是$\{L_i\}_{i\in In}$和Γ的逻辑推论,表示为

$$\langle\{L_i\}_{i\in In},\Gamma\rangle|=_dgs.$$

在上述定义中,对于多个分布的局域 RDF(S)描述,我们可以构建它们之间的关联关系集合Γ,从而将它们联系起来,那么,这些局域 RDF(S)描述中原有的信息知识$\{L_i\}_{i\in In}$ 和关联关系集合Γ一起构成了整个分布 RDF(S)环境中所包含的信息知识,而它们的逻辑推论 $gs(\langle\{L_i\}_{i\in In},\Gamma\rangle|=_dgs)$则可以看作是此分布 RDF(S)环境包含的或者可以推导出的信息知识,以分布一阶逻辑语句的形式来表示.

3 分布RDF(S)查询

根据上一节中定义的分布 RDF(S)模型,我们可以定义在分布模式下的分布查询.通过分布 RDF(S)查询,我们可以获取分布环境下涉及多个局域的 RDF(S)描述信息以及它们所描述的 Web 信息内容.

定义 17. 对于分布语言集合$\{L_i\}_{i\in In}$和包含 L_i 分布语言间关系的关联关系集合Γ,对分布 RDF(S)的一个查询可以表示为 $q(x_1,\ldots,x_n)\leftarrow e$,设 e 是定义 12 中定义的分布一阶谓词公式,$e=b_1:\phi_1\wedge\ldots\wedge b_m:\phi_m,(b_1,\ldots,b_m\in In,\phi_1,\ldots,\phi_m$中所有变元的集合包含$\{x_1,\ldots,x_n\})$.

设 fa'为类似 fa 的赋值函数集合,对于不同局域表达式中共同出现的变元 x,除了等值赋值以外,它根据Γ中①型关联关系γ的资源对应关系赋予相关联的值(例如,根据 $2{:}x=Professor\hat{\rightarrow}3{:}x=Teacher$ 关联关系,它可以将 $3{:}tri(University,has,x)\wedge 2{:}tri(x,instruct,Graduate)$中 3 和 2 局域谓词公式共同出现的 x 分别赋予值:*Teacher* 和 *Professor*,这种赋值称为关联赋值),对于只在某局域内独立出现的变元则单独赋值.那么,对于分布语言集合$\{L_i\}_{i\in In}$和关联关系集合Γ,此 $q(x_1,\ldots,x_n)\leftarrow e$ 查询的分布查询结果为

$$GQR(\{L_i\}_{i\in In},\Gamma,e)=\{\langle d_1,\ldots,d_j/\ldots/d'_j,\ldots,d_n\rangle|\forall fa',\langle\{L_i\}_{i\in In},\Gamma\rangle|=_d e[fa']\},$$

$d_1,\ldots,d_j/\ldots/d'_j,\ldots,d_n$ 分别为 fa'赋值函数集合对变元 $x_1,\ldots,x_n$ 的赋值,其中 $d_j/\ldots/d'_j$ 表示因对在不同局域中共同出现的变元 x_j 进行关联赋值,导致 x_j 对应的结果值可能为多个.

局域查询只能查询到局域 RDF(S)描述中满足条件的值,而分布查询则需要查询通过关联关系 Γ附加上的来源于多个局域满足条件的值,因此,需要将分布查询转换到各个有关联的局域上进行局域查询,并将局域查询结果依照一定关系集成起来,构成分布查询结果.

基于分布 RDF(S)模型,我们设计了一种根据关联关系集合Γ将分布查询转换到局域查询的方法.设对于分布语言集合$\{L_i\}_{i\in In}$ 和关联关系集合Γ的分布查询为 $q(x_1,\ldots,x_n)\leftarrow e$,其中,$e=b_1:\phi_1\wedge\ldots\wedge b_m:\phi_m$,分布查询结果为 $GQR(\{L_i\}_{i\in In},\Gamma,e)$,则其分布查询方法如下:

(1) 如果Γ为空,或者Γ中不存在这样的②型$\gamma{:}e$ 中存在属于 h 局域($h\in In$)的某谓词公式ϕ_i,它等于 $H(\gamma)(h:\psi(x_1,\ldots,x_k))$中谓词公式$\psi$对变元进行一个 $S(S=\{x_1/x'_1,\ldots,x_p/x'_p\}$,其中 $x'_1,\ldots,x'_p$既可能是变元也可能是常量)置换的结果(既$\phi_i=\psi(x_1,\ldots,x_k)S$,$\psi S$ 表示将ψ中的 $x_1,\ldots,x_p$ 分别置换为 $x'_1,\ldots,x'_p$),那么,

$$GQR(\{L_i\}_{i\in In},\Gamma,e)=LQR(L_1,\phi_1)\infty\ldots\infty LQR(L_m,\phi_m).$$

其中,∞根据相邻表达式中共同出现的自由变元对应的查询分量,以及Γ中①型关联关系γ定义的局域 RDF(S)之间的资源对应关系进行连接操作.例如,$i:\phi_i,j:\phi_j$ 之间的共同变元为 $x_1,\ldots,x_k$,设它们的查询结果中分别有 $rq_i=\langle\ldots,d_1,\ldots,d_k,\ldots\rangle$和 $rq_j=\langle\ldots,d'_1,\ldots,d'_k,\ldots\rangle$,$d_1,\ldots,d_k$ 和 $d'_1,\ldots,d'_k$ 为 rq_i,rq_j 中对应 $x_1,\ldots,x_k$ 的查询结果值,如果对所有的 $p(1\le p\le k)$,在Γ中都存在 $i{:}x=d_p\hat{\rightarrow}j{:}x=d'_p$ 这样的关联关系,那么将它们连接起来:$rq_i\infty rq_j=\langle\ldots,d_1/d'_1,\ldots,d_k/d'_k,\ldots\rangle$,并将其作为查询结果;否则不进行连接,不加入结果集.

(2) 否则,如果 e 中属于 h 局域的谓词公式ϕ_i是②型γ的 $H(\gamma)(h:\psi(x_1,\ldots,x_k))$中谓词公式$\psi$对变元进行一个 S 置换的结果,设 $B(\gamma)S$ 表示将γ的体表达式中对应变元进行 S 置换,$e\{h:\phi_i/B(\gamma)S\}$表示将 e 中的 $h:\phi_i$ 谓词公式替换成 $B(\gamma)S$(如果ϕ_i是 e 中 $h:\phi'_i$ 中的一部分,则先将其分裂为 $h:\phi_i\wedge h:\phi''_i$ 两部分),替换后对 $e\{h:\phi_i/B(\gamma)S\}$进行重整,属于同一局域的局域谓词公式进行合并,那么,

$$GQR(\{L_i\}_{i\in In},\Gamma,e)=GQR(\{L_i\}_{i\in In},\Gamma-\{\gamma\},e)\cup GQR(\{L_i\}_{i\in In},\Gamma,e\{b:\phi_i/B(\gamma)S\}).$$

这一算法的主要思想是:对于分布查询表达式 $q(x_1,\ldots,x_n)\leftarrow e$,$e$ 中谓词公式 $h:\phi_i(x_1,\ldots,x_m)(h\in In)$的分布查询结果集合(设其为型如$\langle d_1,\ldots,d_m\rangle$的元组组成的集合)应该包含两部分内容:

一部分是 h 局域的局域 RDF(S)查询结果$\langle d_1,\ldots,d_m\rangle$(存在于 $GQR(\{L_i\}_{i\in In},\Gamma-\{\gamma\},e)$中),根据局域 RDF(S)查询定义,对于任何满足 L_h 的解释 I_h,都有 $I_h|=\phi_i(d_1,\ldots,d_m)$.因此,对于任何满足$\{L_i\}_{i\in In}$ 并且满足Γ的 I_{FRDF},都有 $I_{FRDF}|=_d h:\phi_i(d_1,\ldots,d_m)$,则$\langle\{L_i\}_{i\in In},\Gamma\rangle|=_d h:\phi_i(d_1,\ldots,d_m)$,所以,$\langle d_1,\ldots,d_m\rangle$是全局查询结果的一部分.

另外一部分是根据②型关联关系 $B(\gamma)\rightarrow h:\psi(x_1,\ldots,x_k)(\phi_i(x_1,\ldots,x_m)=\psi(x_1,\ldots,x_k)S)$来源于 $B(\gamma)$的查询结果,设其为$\langle d_1,\ldots,d_m\rangle$(存在于 $GQR(\{L_i\}_{i\in In},\Gamma,e\{b:\phi_i/B(\gamma)S\})$中),假设它是对 $B(\gamma)S$ 分布查询的正确结果,则对于任何满

足$\{L_i\}_{i\in In}$并且满足Γ的I_{FRDF},都有$I_{FRDF}|=_d B(\gamma)S\{x_1/d_1,\ldots,x_m/d_m\}(\{x_1/d_1,\ldots,x_m/d_m\}$表示分别对变元$x_1,\ldots,x_m$赋予$d_1,\ldots,d_m$值);同时根据定义,满足$\gamma$的$I_{FRDF}$中存在着对应关系$\langle d_1,d_1\rangle,\ldots,\langle d_m,d_m\rangle$,它们分别属于$B(\gamma)S$中各局域$b(b\in In)$到$h$的对应关系$r_{bh}$(如果$d_1,\ldots,d_m$中的某些值是在递归调用时根据其他②型$\gamma$关系来源于其他局域查询时,可以通过$r_{ij}$关系的传递性证明),并使得当$I_{FRDF}|=_d B(\gamma)S\{x_1/d_1,\ldots,x_m/d_m\}$时,$I_{FRDF}|=_d(h:\psi)S\{x_1/d_1,\ldots,x_m/d_m\}$,即$I_{FRDF}|=_d h:\phi_i(d_1,\ldots,d_m)$.因此,对任何满足$\{L_i\}_{i\in In}$和$\Gamma$的$I_{FRDF}$,都有$I_{FRDF}|=_d h:\phi_i(d_1,\ldots,d_m)$,所以,$\langle d_1,\ldots,d_m\rangle$也是正确的分布查询结果.

同时,在对局域查询结果进行连接时,算法根据Γ中①型关联关系进行连接,不满足局域 RDF(S)描述资源之间对应关系的结果值不被加入结果集.

综上所述,采用上述算法所获得的结果是满足条件的查询结果集.

对于图 1 中的 RDF(S)1~RDF(S)3 对应的分布语言集合$\{L_i\}_{i\in In}$,$In=\{1,2,3\}$,以及上一节中描述它们之间关系的关联关系集合Γ,存在着下述查询例子:

Q_1:$q(x',y')\leftarrow 1{:}relation(belongto,x',U_{xxx})\wedge 3{:}class(Teacher,x')\wedge relation(old,x',y')$,

Q_2:$q(x',y')\leftarrow 3{:}class(University,x')\wedge relation(has,x',y')$,

Q_3:$q(x',y',z')\leftarrow 3{:}tri(University,has,x')\wedge 2{:}tri(y',dom,x')\wedge tri(y',range,z')$.

其中,Q_1表示查询属于U_{xxx}大学的教师以及他们的年龄,Q_2表示查询大学以及大学中的所有人,Q_3表示查询和*University*类满足*has*属性关系的类概念以及此概念所有属性及属性的值域.Q_1,Q_2是实例层次的查询,Q_3是概念层次的查询.

在对Q_1进行分布查询时,$class(Teacher,x')\wedge relation(old,x',y')$在局域 3 的局域 RDF(S)查询结果为空,根据$2{:}class(Professor,x)\rightarrow 3{:}class(Teacher,x)$和$1{:}relation(birthyear,x,NOWYEAR-y)\rightarrow 3{:}relation(old,x,y)$关联关系进行查询谓词公式替换,$Q_1$对应的分布查询结果为

$$LQR(L_1,relation(belongto,x',U_{xxx})\wedge relation(birthyear,x',NOWYEAR-y'))\infty LQR(L_2,class(Professor,x')),$$

其中,对局域 1、局域 2 的局域 RDF(S)查询结果分别为$\langle P_{Smith},40\rangle$和$\langle P_{Smith}\rangle$,它们根据①型$\gamma$关系$1{:}x=P_{Smith}\hat{\rightarrow}2{:}x=P_{Smith}$对$x'$变元的查询结果连接,生成分布查询结果$\langle P_{Smith},40\rangle$.$Q_2$分布查询根据关联关系$1{:}relation(belongto,x,y)\rightarrow 3{:}relation(has,y,x)$和$1{:}relation(belongto,x,y)\wedge 2{:}relation(instruct,x,z)\rightarrow 3{:}relation(has,y,z)$以及$1{:}x=U_{xxx}\rightarrow 3{:}class(University,x)$被转换成两部分:$GQR(\{L_i\}_{i\in In},\Gamma,1{:}x'=U_{xxx}\wedge relation(belongto,y',x'))$和$GQR(\{L_i\}_{i\in In},\Gamma,1{:}x'=U_{xxx}\wedge relation(belongto,x,x')\wedge 2{:}relation(instruct,x,y'))$,分别查询学校所有的教师以及所有的研究生,这两部分的结果用$\cup$并集关系合并.对Q_3进行查询时,对应$3{:}tri(University,has,x')$的局域 RDF(S)查询结果$\langle Teacher\rangle$和对应$2{:}tri(y',dom,x')\wedge tri(y',range,z')$的局域查询结果$\langle Professor,instruct,Graduate\rangle$根据$2{:}x=Professor\hat{\rightarrow}3{:}x=Teacher$关联关系进行连接,构成查询结果$\langle Professor/Teacher,instruct,Graduate\rangle$.

通过以上分布查询方法,根据分布 RDF(S)之间的关联关系,用户可以获取涉及多个局域、满足查询条件且相关联的 RDF(S)信息以及它们语义描述的信息内容.

4 分布RDF(S)集成

通过集成分布的局域 RDF(S)描述,用户能够统一地访问全局范围内多个分布的 RDF(S)描述以及它们描述的 Web 信息内容.基于上文中的分布 RDF(S)模型以及分布查询方法,我们可以设计一种分布 RDF(S)的集成方法.

在集成分布的 RDF(S)描述时,对于已有的分布语言集合$\{L_i\}_{i\in In}$和L_i分布语言间关联关系集合Γ,我们可以构建一个表示全局概念的 RDF(S)描述,全局 RDF(S)描述包含所有局域 RDF(S)中概念层次的实体,并按照局域 RDF(S)中的关系以及分布 RDF(S)描述之间的关联关系构建全局 RDF(S)描述中实体之间的联系.

对于图 1 中 3 个局域 RDF(S)描述对应的分布语言集合$\{L_i\}_{i\in In}$($In=\{1,2,3\}$)和描述它们之间关系的关联关系集合Γ,我们可以构建图 2 中的 RDF(S)描述,它包括了 RDF(S)1~RDF(S)3 中所有概念层次的实体,设其对应的语言为L_G.其中,通过Γ中的$2{:}class(Professor,x)\rightarrow 3{:}class(Teacher,x)$关联关系,我们可以在*Professor*类和*Teacher*

类之间构建“rdfs:subClassOf”符号描述的子类属性关系.

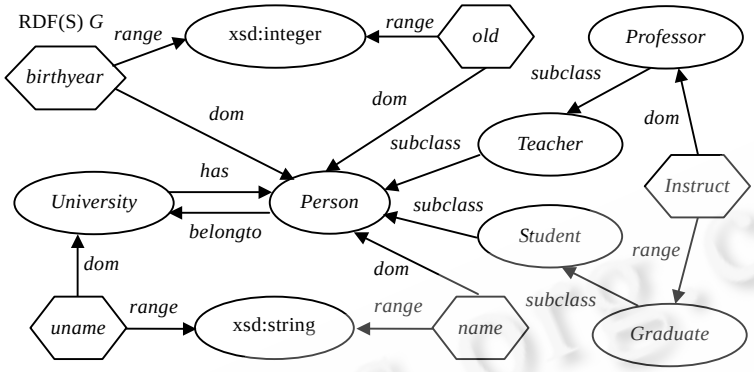

Fig.2 The global RDF(S) description to express global concepts

图 2 表示全局概念的全局 RDF(S)描述

同时,我们可以构建 L_G 与分布语言$\{L_i\}_{i\in In}$之间的关联关系集合Γ',其中包括:(1) 对于类概念,构建其与来源 RDF(S)类之间的类从属对应关系和类对应关联关系,例如,2:*class*(*Professor*,*x*)→*G*:*class*(*Professor*,*x*),2:*x*=*Professor* $\hat{\rightarrow}$ *G*:*x*=*Professor*;(2) 对于属性概念,构建其与来源 RDF(S)之间的满足属性条件的对应关系以及属性对应关联关系,例如,1:*relation*(*uname*,*x*,*y*)→*G*:*relation*(*uname*,*x*,*y*),1:*x*=*uname* $\hat{\rightarrow}$ *G*:*x*=*uname*;(3) 同时,还可以构建分布 RDF(S)描述中的信息实例与 L_G 的概念之间的从属关系,例如,2:$x=U_{xxx}$→*G*:*class*(*University*,*x*).

通过加入上述 L_G 和Γ',可以构建新的分布语言集合$\{L_i\}_{i\in In}\cup\{L_G\}$和分布语言间关联关系的集合$\Gamma\cup\Gamma'$.

对全局各个分布 RDF(S)描述的查询可以统一表示为对 L_G 描述的概念和实例查询,例如,查询所有学校以及这些学校中所有的人可以表示为 *q*(*x*,*y*)←*G*:*class*(*University*,*x*)∧*relation*(*has*,*x*,*y*);查询所有 *Person* 类的子类可以表示为 *q*(*x*)←*G*:*relation*(*subclass*,*x*,*Person*).若全局查询 *Q* 形式为 $q(x_1,\ldots,x_n)\leftarrow e$(*e* 为仅包含 *G* 局域的分布一阶逻辑公式),则 *Q* 的查询结果为 $GQR(\{L_i\}_{i\in In}\cup\{L_G\},\Gamma\cup\Gamma',e)=\{\langle d_1,\ldots,d_j/\ldots/d'_j,\ldots,d_n\rangle|\forall fa',\langle\{L_i\}_{i\in In}\cup\{L_G\},\Gamma\cup\Gamma'\rangle|=_d e[fa'],fa'$为定义 17 中定义的根据$\Gamma\cup\Gamma'$中关联关系的赋值函数集合}.可以按照上一节描述的分布查询方法将此全局查询转换到局域查询上,局域查询结果集成后就是满足条件的全局查询结果.通过上述方法就可以实现分布 RDF(S)的集成与查询.

5 相关工作

Ghidini 在文献[4,5]中提出了用于实现分布语言互操作性的局域模型理论和分布逻辑模型.Serafini 在文献[6,7]中介绍了基于分布逻辑模型理论的分布关系数据模型,这一模型及其方法解决了分布关系数据库的数据实例层次的集成问题.本文提出的分布 RDF(S)模型是基于分布逻辑的,与上述分布模型及其应用的不同之处在于:为了适应分布 RDF(S)的查询和集成的要求,本文提出的分布 RDF(S)模型可以同时集成概念层次和实例层次的数据,而不同于关系数据库集成中仅仅集成实例层次的数据.本文提出了基于分布语言之间的关联关系集合Γ的查询方法,而不是基于分布数据实例之间的对应关系 r_{ij},使用规则比使用数据对应可以更方便地表示分布 RDF(S)之间的关系;同时,此关联关系可以表示涉及多个局域的联系,而不仅仅是两个局域之间的对应关系[5].通过以上关联关系的定义以及本文给出的分布查询方法,用户可以更灵活地描述局域 RDF(S)之间的联系,并实现对分布 RDF(S)描述的查询.

Franconi 建立了一种基于限制性规则的查询模型[8],并对各种基于规则的查询变换方法进行了分析.Bruijn 提出了一种将 RDF(S)描述转换成一阶谓词逻辑的方法[9],这样,可以利用逻辑推导来实现对 RDF(S)的查询,在其表示中将属性作为谓词,这样会导致无法表示对属性的查询.而在本文提出的 *tri*(*s*,*p*,*o*)谓词中,属性 *p* 可以被设置为变量,从而可以实现概念层次对属性的查询.文献[8,9]中都只考虑单一信息源情况下的查询处理,而本文

提出的基于关联关系的查询变换方法则是针对涉及多个局域的分布 RDF(S)描述的.

Staab 使用 RDF(S)来描述表示领域模型的本体(ontology)[10].Xiao 则提出了一种采用 RDF(S)本体来集成各种信息资源(包括局域 RDF 描述以及 XML 文档)的方法[11].实际上,本文中只包含概念层次描述的全局 RDF(S)描述就可以看作是一种全局本体,它可以用来表示所集成的分布 RDF(S)描述的全局统一语义.

6 结束语

在 Web 应用环境中可能存在多个具有独立语义的 RDF(S)描述.通过构建分布 RDF(S)模型,可以将这些分布异构的 RDF(S)描述联系起来,实现其信息关联性.同时,可以将分布 RDF(S)查询转换为局域 RDF(S)查询,局域查询结果进行特定的集成后就是分布 RDF(S)查询所需要的结果.

本文提出了描述分布异构 RDF(S)的分布 RDF(S)模型,并基于这一模型给出了实现分布 RDF(S)查询的方法.此查询方法既能实现实例层次的查询,也能实现概念层次的查询.同时,在上述基础上,本文还给出了集成全局 RDF(S)描述的方法.

今后的工作主要集中在 RDF(S)集成方法的研究上,主要考虑如何自动构建表示全局概念的 RDF(S)描述以及如何解决分布 RDF(S)描述之间的异质冲突等问题.

References:

[1] Hayes P. RDF semantics. W3C Recommendation, 2004. http://www.w3.org/TR/rdf-mt

[2] Patel-Schneider P, Simeon J. The yin/yang Web: A unified model for XML syntax and RDF semantics. IEEE Trans. on Knowledge and Data Engineering, 2003,15(4):797−812.

[3] Gutierrez C, Hurtado C, Mendelzon AO. Foundations of semantic Web databases. In: Deutsch A, ed. Proc. of the ACM Symp. on Principles of Database Systems (PODS). New York: ACM Press, 2004. 95−106.

[4] Ghidini C, Giunchiglia F. Local models semantics, or contextual reasoning=locality+compatibility. Artificial Intelligence, 2001, 127(2):221−259.

[5] Ghidini C, Serafini L. Distributed first order logics. In: Baader F, Schulz KU, eds. Frontiers of Combining Systems 2. Berlin: Research Studies Press, 1998.

[6] Serafini L, Giunchiglia F, Mylopoulos J, Bernstein PA. The local relational model: Model and proof theory. Technical Report, 0112-23, Istituto Trentino di Cultura, IRST, 2001.

[7] Serafini L, Ghidini C. Local models semantics for information integration. Technical Report, 9702-04, Istituto Trentino di Cultura, IRST, 1997.

[8] Franconi E, Tessaris S. Rules and queries with ontologies: A unified logical framework. In: Ohlbach HJ, Schaffert S, eds. Proc. of the 2nd Int'l Workshop on Principles and Practice of Semantic Web Reasoning. LNCS 3208, New York: ACM Press, 2004. 50−60.

[9] Bruijn J, Franconi E, Tessaris S. Logical reconstruction of RDF and ontology languages. In: Fages F, Soliman S, eds. Proc. of the 3rd Workshop on Principles and Practice of Semantic Web Reasoning (PPSWR 2005). Berlin: Springer-Verlag, 2005.

[10] Staab S, Erdmann M, Maedche A, Decker S. An extensible approach for modeling ontologies in RDF(S). In: Grütter R, ed. Proc. of the 1st Workshop on the Semantic Web at the 4th European Conf. on Digital Libraries. Hershey: IGI Publishing, 2000. 18−20.

[11] Xiao H, Cruz IF, Hsu F. Semantic mappings for the integration of XML and RDF sources. In: Proc. of the Workshop on Information Integration on the Web (IIWeb 2004). 2004.

李剑(1976－),男,湖北武汉人,博士,主要研究领域为语义 Web,网络信息处理.