ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.8, August 2007, pp.2049−2062
DOI: 10.1360/jos182049

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

医学图像网格基于语义的信息集成方法*

金 海[1,2+], 孙傲冰[1,2], 郑 然[1,2], 何儒汉[1,2], 章 勤[1,2], 吴 松[1,2]

[1](华中科技大学 计算机科学与技术学院 服务计算技术与系统教育部重点实验室,湖北 武汉 430074)

[2](华中科技大学 计算机科学与技术学院 集群与网格计算湖北省重点实验室,湖北 武汉 430074)

Semantic-Based Medical Information Integration Scheme for Medical Image Grid

JIN Hai[1,2+], SUN Ao-Bing[1,2], ZHENG Ran[1,2], HE Ru-Han[1,2], ZHANG Qin[1,2], WU Song[1,2]

[1](Services Computing Technology and System Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

[2](Cluster and Grid Computing Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

+ Corresponding author: Phn: +86-27-87543529, Fax: +86-27-87557354, E-mail: hjin@hust.edu.cn, http://grid.hust.edu.cn/hjin/

Jin H, Sun AB, Zheng R, He RH, Zhang Q, Wu S. Semantic-Based medical information integration scheme for medical image grid. *Journal of Software*, 2007,18(8):2049−2062. http://www.jos.org.cn/1000-9825/18/2049.htm

Abstract: A semantic-based information integration scheme for MedImGrid (medical image grid) is presented, which creates parent-ontology (HL7-RIM ontology) based on HL7-RIM (health level 7 referenced information model), and adopts hybrid means to construct the hierarchical structure of MedImGrid global and local ontologies. The HL7 (health level 7) gird middleware is developed based on Agent and middleware technology, which gives the semantic parsing capability to HL7 intelligent Agent to support grid service encapsulation and uniform access of heterogeneous data sources. The interrelations of data modes at ontology layer are denoted with ontology tag and used to support the semantic parsing and mapping between different medical data sources referring to MedImGrid ontologies. MedImGrid prototype is based on CGSP2 (China grid support platform v2.0) and adopts global and local semantic mapping loosely coupled means, and its special layered structure makes resource sharing and matching across systems and hospitals more efficient.

Key words: medical image grid; ontology tag; medical information integration; semantic mapping; HL7 (health level 7) intelligent Agent

摘 要: 提出了一种医学图像网格 MedImGrid(medical image grid)基于语义的信息集成方法.基于 HL7 RIM(health level 7 referenced information model)生成父本体(HL7-RIM ontology),采用混合方式(hybrid means)建立 MedImGrid 全局和局部本体的分级结构.结合代理和中间件技术开发了 HL7(health level 7) Grid 中间件,实现了具有

* Supported by the National Natural Science Foundation of China under Grant Nos.60673174, 90412010 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant Nos.2006AA02Z347, 2006AA01A115 (国家高技术研究发展计划(863))

Received 2007-03-01; Accepted 2007-04-26

医疗语义解析功能的 HL7 智能代理,以支持对异构数据源的 Grid Service 封装与统一访问.基于本体标记表达异构数据模式的语义模型在本体层的相关关联,参照 MedImGrid 各级本体实现数据源间的语义解析和映射.MedImGrid 原型系统基于 CGSP2(China grid support platform v2.0),采用了局部与全局语义映射松耦合的构架,其特有的层次结构使得网格环境下跨系统/医院的信息集成更加有效.

关键词: 医学图像网格;本体标记;医疗信息集成;语义映射;HL7(health level 7)智能代理

中图法分类号: TP391 **文献标识码:** A

Internet 中的数据呈爆炸性地增长,在一些专业领域,数据量的增长速度更加惊人.当前,一家大型医院一个工作日新增的医疗数据就接近 10GB,且随着医院数字化程度的提高,这个数字有着不断增加的趋势.如何对这些数据进行有效的组织和管理,实现医疗信息的广泛共享,以提高医疗资源的利用率,降低社会整体医疗成本,同时发现海量医疗数据的内在联系和深层规律,挖掘医疗信息的最大潜力,对国家医疗卫生水平的提高有着重要的意义.

网格技术以实现互联网上所有资源(计算资源、存储资源、数据资源等)的全面共享与协作,将互联网整合成一台巨大的超级计算机,为用户提供“即连即用”式的服务为目标[1].信息集成是网格研究的一个重要分支,众多研究人员在相关领域进行了深入的研究,提出数据网格、信息网格、知识网格等与信息互操作和信息集成相关的网格原型系统.MedImGrid(medical image grid)是华中科技大学集群与网格计算湖北省重点实验室开发的一个网格原型,旨在利用网格技术整合分布式资源的优势,屏蔽网格内部海量信息的异构性,实现异构、多源医疗信息的集成;并在此基础上借助于网格计算技术,支持医学图像处理、病理建模、流行病预警等计算密集型医疗应用研究[2].

1 信息集成技术概述

信息集成的主要目的是屏蔽底层信息源的异构性,实现基于不同软、硬件平台的信息系统在不同的数据模式、通信协议、查询语言、并发性控制与数据一致性维护规则条件下,信息源访问控制、互操作、事务响应等一系列问题的集成解决方案,为用户提供逻辑上统一的数据视图和信息访问接口.

当前,分布式异构信息集成的研究主要集中在两个方向上,即结构级方法和语义级方法[2,3].结构级的集成方法着重解决由不同的数据表述语言、数据表达方式、数据模型所引起的异构,如Stanford大学基于CORBA体系结构实现的INFOBUS系统[4]和基于Mediator/Wrapper架构实现的TSIMMIS[5]系统.语义级的集成方法着重解决由术语定义、概念结构和相互关系的差异所造成的异构,如Ontobroker[6]和Infosleuth[7]项目通过对基于Ontology的数据模型的管理和互操作提供语义的支持.结构级的集成系统大多采用数据视图集成的方式.其特点是实现比较简单,信息源相对比较固定,但是可扩展性较差.语义级方法具有可扩展性好、动态适应性强、支持语义操作等特点,但是实现比较复杂,涉及到本体的创建.

网格环境下海量的、动态的和自治的信息源,对信息集成方法的自动化程度和动态适应性提出了进一步的要求.比如,DataGrid项目基于SRB(storage resource broker)中间件提供了一种与应用无关的元数据服务,使各类异构数据系统集成能够基于统一的元数据驱动和访问机制实现[8];e-Science的eDiaMoND[9]项目基于OGSA-DAI[10]中间件提供的网格数据服务(grid data service,简称GDS)实现了对医疗关系数据库的集成;HealthGrid的MammoGrid项目基于ALice的AliEN环境以对等的架构向用户提供虚拟的数据资源视图[11,12].网格环境下,为海量的数据源创建统一的语义模型比较困难.因此,相关本体模型的构建只针对一些专业应用领域,如针对生物医学应用的myGrid项目[13].

医疗信息集成应用是信息集成领域的难点,涉及到多种医疗信息系统(如医院信息系统HIS、医疗影像存储和通信系统PACS、放射线信息系统RIS等)和应用领域(如流行病学和基因医学),采用了多种医疗应用协议(如HL7(health level 7),OpenEHR,EHRcom,ENV13606 等)、工业标准(如ICD,SNOMED,LOINC,UMLS等)和医疗应用操作流程.由于数据源的数据模式众多,信息交换模型构造困难,使得医院之间以及医院信息系统之间缺乏有

效的数据共享机制.MedImGrid将结构级和语义级的信息集成方法进行有效的结合,通过综合多代理技术和中间件技术将各种医疗信息系统封装为统一的Grid Service资源,并以Agent Server作为网格与医院的医疗信息系统进行数据交换的接口;基于混合本体架构构建MedImGrid本体,在保证各医疗领域本体和应用本体独立性的同时,基于共享词汇表在本体间建立了数据模式的语义关联;通过元数据模型与语义模型的统一,加速了从语义模型到底层数据源的映射.MedImGrid的Agent Server节点部署了语义操作组件,支持基于本体的语义映射操作,能够智能地解析用户的请求,适应了网格环境下医疗信息集成的需要.MedImGrid的目标是将网格信息集成从简单地对数据源的整合发展为集信息互操作、数据处理(预/后期处理)、语义交互、工作流等技术于一体的自动化流程[14].

2 网格环境下的信息集成

2.1 网格环境下信息集成的语义模型

信息集成的目的是向用户提供符合需求的数据处理结果,因此,数据的来源和处理过程对于用户应该是透明的.网格环境下,由于涉及众多的信息源和信息处理服务,用户不可能对这些系统的数据模式都有全面的了解,因此对信息集成的自动化程度提出了更高的要求.

定义 1(网格信息集成). 本文将 MedImGrid 网格信息集成定义为一个三元组 $K=\langle Q,S,W\rangle$,其中,Q 为网格信息集成请求,S 为完成请求需要访问的信息源集合,W 为完成请求需要访问的信息处理系统(或 grid service)的集合.Q,S 和 W 都工作在一定的数据模式下,即与一定的语义环境绑定.

$Q=\langle Q_{Req},Q_{Mode}\rangle$,其中,$Q_{Req}$为请求的描述字符集,如SQL或XML查询语句;$Q_{Mode}$为$Q_{Req}$遵循的数据模式或语义模型.

$S=\langle S_{ID},S_{Mode},Q_S,R_S\rangle$,其中,$S_{ID}$为信息源的标识集合;$S_{Mode}$,$Q_S$和$R_S$为与$S_{ID}$对应的集合,其中,$S_{Mode}(i)$为$S_{ID}(i)$对应的数据模式,$Q_S(i)$为$S_{ID}(i)$能够识别的对应请求,$R_S(i)$为$S_{ID}(i)$对$Q_S(i)$的处理结果.

$W=\langle W_G,M_W,S_{IN},R_{PRO}\rangle$,设$Q$被分解为$m$个处理步骤,则有$m$个信息处理系统的集合$W_G$与其对应.$M_W$,$S_{IN}$和$R_{PRO}$为与$W_G$对应的集合,其中,$M_W(i)$为$W_G(i)$对应的数据模式,$S_{IN}(i)$为$W_G(i)$的输入(它由$R_S$的子集组成),$R_{PRO}(i)$为$W_G(i)$的输出.

MedImGrid信息集成的过程可以被看作信息请求和中间结果在数据通道中流动的过程.如图1所示,请求Q在执行过程中分别从 n 个数据源中获取信息,并经 m 个串行或可并行的中间步骤处理,形成最终结果 Res.从中我们可以看出,网格信息集成的核心问题是建立 Q,S 和 W 三者之间数据模式或语义模型的转换和映射关系.

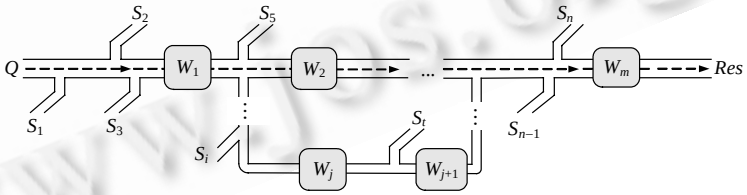


Fig.1 Semantic model of grid information integration

图 1 网格环境下信息集成的语义模型

2.2 网格信息集成方法的评价

网格环境下信息源众多、相关关系复杂,如何以最小的成本在最大范围内实现安全、高效的信息互操作是网格信息集成追求的目标.本文定义了从以下 6 个方面来对比和评价网格环境下信息集成实现方案的性能:

(1) 集成效率(efficiency),即集成方案对于异构信息系统集成所表现出的整体效率,主要表现为集成系统整体响应速度、信息查全率、查准率等.集成效率是评价集成方案的关键参数,它决定了集成方案是否可以为用户所接受.

(2) 集成代价(cost),即信息集成的实现在保证原有个体信息系统正常工作的前提下所需要的软件和硬件

的总开销,包括增设的服务器设备、开发相关中间件和配套软件所要耗费的成本等.它关系到集成方案能否实现或在怎样的一个范围内实现.

(3) 集成安全性(security),即集成实现方法对原有个体信息系统的数据安全性、运行稳定性可能造成的影响.部分集成方法绕过信息系统的安全控制模块,以直接操作数据库的方式实现信息集成,尽管实现方法简单,但对信息系统的安全性可能带来负面的影响.

(4) 自适应能力(auto-adaptation),即信息集成方案对于新增的异构数据源是否能够支持数据源模式的自学习,以及对于新增系统的通信协议、数据模式、查询语言等的自动适应.它直接决定了信息集成方法的可扩展性和适用范围.网格环境下,只有鲁棒的、支持自适应扩展的方案才是切实可行的.

(5) 人工干预度(manual degree),即集成方案对于新加入的异构数据源需要人工参与的工作量.网格环境中,面对数量庞大、动态和异构的数据源,开发、管理和维护的开销将是非常庞大的.因此,较小的人工干预度才能保证系统的稳定性和易维护性.

3 MedImGrid 语义环境构建

3.1 网格环境下本体的构建方法

本体是作为“概念模型的明确的规范说明”来引入的.因此,可以使用本体来描述医疗信息系统(HIS,PACS和RIS等)的数据模式,并基于本体建立语义关联以及实现数据模式的转换,这为MedImGrid实现跨医院和跨地区的信息集成提供了新的途径.网格环境下,通过将异构数据源模式的语义模型在语义层构建相关关联,实现基于本体的数据模式的描述与共享有多种途径,主要包括以下 3 种策略[15,16]:

(1) 单本体策略(single ontology)使用单一的全局本体来对网格范围内的数据源的数据模式进行语义描述.这种方法适合于待整合的数据源拥有接近或相同的数据模式,但不能针对不同描述粒度的数据模式进行调整,无法适应数据源的动态变化.

(2) 多本体策略(multiple ontologies)的每一数据源分别由各自的本体描述,这些本体可以是一些领域本体的并集.它不需要网格中存在一个全局共享本体或是本体最小承诺.数据源可以独立地按照自己的方法去创建本体,能够较好地支持个体本体的更新.但在集成过程中,需要对所有可能的本体对象进行比对.

(3) 混合本体策略(hybrid ontologies)集成以上两种方法的优势.它由全局本体(global ontology)和局部本体(local ontologies)组成,即每一个数据源仍采用其自身的本体进行描述,但需要通过一个全局本体建立各局部本体的相关关联.混合本体共享同一个基本词汇表,这些基本词汇可以组合描述更加复杂的语义.混合本体策略为各数据源本体(局部本体)的新增概念增加一个标签,使用共享词汇表的语汇进行说明,使得新增概念可以被其他本体理解.全局本体同样也是基于共享词汇表构建,它可以从不同的局部本体中抽取语义信息完成自己知识库的更新.全局本体和各数据源本体基于共享的词汇表,使本体间的语义映射和转换成为可能.此外,混合本体使得新数据源可以在无改动的情况下较容易地加入,且支持各本体独立的知识获取和语义更新[17].

3.2 MedImGrid本体构建

MedImGrid采用混合本体的方法构建相关本体,其层次架构如图 2 所示.HL7 标准是ANSI(美国国家标准化组织)支持的医疗信息系统间通信标准,在世界范围内得到了广泛的采用[18].我们采用扩展的HL7 RIM(referenced information model)词汇表来构建MedImGrid的混合本体,这与我们基于HL7 智能代理实现信息集成的方法相对应.MedImGrid各级本体的父本体(HL7-RIM ontology)原型通过将HL7 模型转换成OWL(ontology Web language)格式并导入Protégé 2000[19]来实现.HL7-RIM Ontology可以有效地用于各种应用本体(医疗信息源本体)的创建、转换和映射.网格平台本体仓库中也可以同时存储其他粒度的医疗领域本体(如基因学本体、分子医学本体等).当MedImGrid本体无法解析用户请求时,平台可以用这些本体来进行语义匹配.MedImGrid本体的创建和更新也可以依赖于各级元数据,从数据源的元数据中获取有用的信息,并通过监控元数据的变化来实现MedImGrid各级本体的更新.

定义 2(MedImGrid 本体). 本文将 MedImGrid 本体定义为一个 4 元组 $O=\langle A,T,X,I\rangle$,其中,A 为本体属性集,T 为概念集,I 为基本术语集,X 为实例集.

$A=\langle N,D,M\rangle$,其中,N 为本体名称,D 为本体域,M 为本体概念关联度矩阵.

$T=\langle C,P,R,\delta\rangle$,其中,$C$ 为原子类概念,P 为原子属性集,R 为原始关系集,δ为概念间关系矩阵.

定义 3(本体标记). MedImGrid 为网格中每一个本体定义了唯一的本体标记 *Tag*,并将所有信息集成操作与唯一的本体标记绑定,使得平台在解析请求时能够通过本体标记掌握操作请求中的概念或实例所对应的本体(语义环境),以高效地支持请求的语义自描述和互操作.

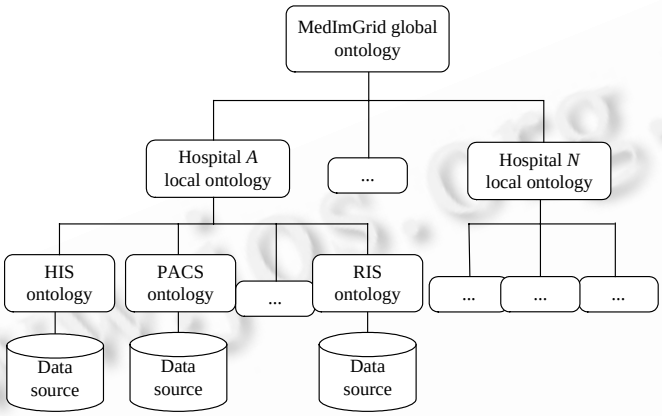


Fig.2 Hierarchical structure of hybrid ontology of MedImGrid

图 2 MedImGrid 混合本体的分级结构

MedImGrid 通过建立本体关联树来记录平台混合本体的层次结构和相关关系.MedImGrid 平台的所有信息集成操作都与某一本体唯一对应(如图 3 所示),使得信息集成操作和处理结果能够通过本体标记表述自己应用的数据模式所对应的本体.当目标本体与当前的本体不相匹配时,平台获取两者的本体标记,通过执行相关语义映射和转换操作,完成数据在不同数据模式间的转换.

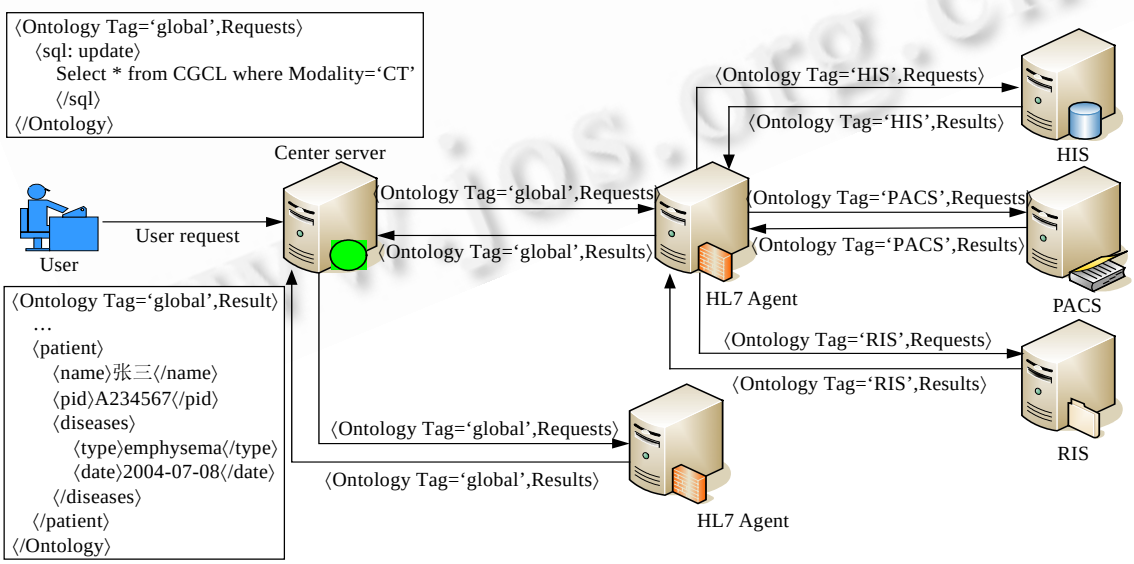


Fig.3 Semantic requests and operations owning ontology tag within MedImGrid

图 3 MedImGrid 中持有本体标记的语义请求和操作

4 语义映射和转换的实现

本体间映射的方法有多种,如KRAFT项目中采用的定义映射(defined mapping)的方法、OBSERVER系统采用的词汇关联(lexical relation)的方法以及DWQ采用的顶级本体(top-grounding ontology)关联等[17].本文采用一种基于概念/实例的映射方法.它基于概念/实例间的映射关系,为不同数据模式间的翻译和转换产生映射策略Schema.

当本体间具有以下关系时,它们是有可能进行语义映射的:(1) 直接关联的本体之间;(2) 具有共同的父本体或祖本体的本体之间.当两个子本体O_1和O_2没有直接关系时,容易证明两者可以通过共同的父本体建立映射关系,即先将O_1中的概念映射到父本体O,再通过O映射到O_2中.MedImGrid中的各本体基于同一的父本体HL7-RIM且使用同一基本词汇表进行描述,以简化本体间的语义映射和转换的实现.语义映射的过程实际上是将一个本体O_1中的概念c_1、属性p_1和关系r_1映射为另一本体O_2中的概念c_2、属性p_2和关系r_2的过程,如图4所示[20,21].

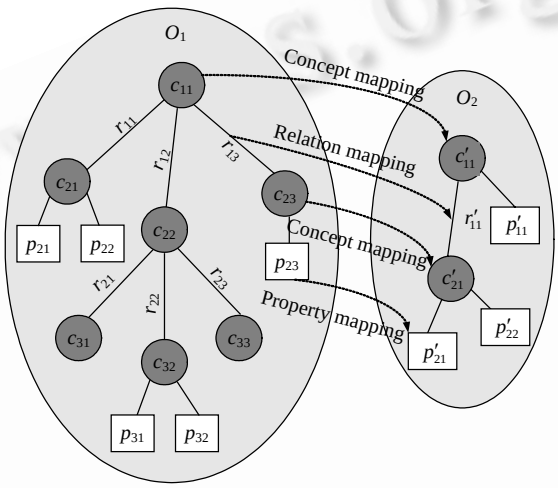


Fig.4 Mapping functions across ontologies

图4 本体间映射函数

定义 4(本体映射函数). 对于任意两个本体O_1和O_2,本体映射函数M实现了将O_1中的c_1,p_1和r_1映射到O_2中且满足$(M(c_1)\subseteq C_2)\wedge(M(r_1)\subseteq R_2)\wedge(M(p_1)\subseteq P_2)$.当$M$为$O_1$和$O_2$间的入射关系时,称$M$为本体单向映射函数;当$M$为$O_1$和$O_2$间的双射关系时,称$M$为本体全向映射函数;当$M$仅为$O_1$和$O_2$子集间的映射关系时,称$M$为本体局部映射函数.

M是MedImGrid产生不同数据模式间映射策略Schema的基础.M的创建过程实际上是不同本体的概念根据自己的定义表述字段(使用共享词汇表描述)或连接关系进行匹配的过程.对于没有歧义关系存在的完备本体,容易证明映射函数M具有以下性质:

(1) $T_1\subseteq T_2\Rightarrow M(T_1)\subseteq T_2$.

(2) $M(O_1)=M(T_1)\cup M(X_1)$;实例集X由概念、属性和关系的集合组合而成,即

$$X=\left\{\bigcup_{i=1}^{m}\bigcup_{j=1}^{n}\bigcup_{k=1}^{t}c_i p_j r_k\right\},M(X)=M(C)\cup M(R)\cup M(P) \tag{1}$$

(3) O为父本体,O_i,O_j为O的派生本体,存在O_i到O的映射M_i,O到O_j的映射M_j,那么,必然存在从O_i到O_j的映射$O_i \xrightarrow{t} O_j$,且$t\subseteq M_i\times M_j$.

定义 5(本体关联度). MedImGrid本体以树状关系来组织,本体间联结边的权值反映了本体间的紧密度.对

于MedImGrid本体结构内任意两个本体O_i和O_j,定义$\kappa(\overbrace{O_i,O_j})\in[0,1]$为本体关联度,它表示本体关系树中,本体间的连结强度.κ具有以下性质:

(1) 当O_i和O_j为非相关本体时,$\kappa(\overbrace{O_i,O_j})=0$;

(2) 当O_i,O_j,O_k互为相关本体时,

$$\kappa(\overbrace{O_i,O_k})=\kappa(\overbrace{O_i,O_j})\times\kappa(\overbrace{O_j,O_k}) \tag{2}$$

当O_i和O_j具有共同父本体时,通过计算各概念的映射完整度integrity来计算$\kappa(\overbrace{O_i,O_j})$,如图 5 所示.其中,每一概念的映射integrity为该概念的属性、子概念、关系映射成功率的加权平均值.当O_i中所有的概念、属性、关系可以映射到O_j时,有$\kappa(\overbrace{O_1,O_2})=1$.

```
float ontoCorrelation (ontology O1,O2)
 {float correlation=0; int count=0;
   for (all ci in O1)
     {float integrity=0;
     if (ci can map to O2)
      {count+=1;
       float p=ci property mapping success rate;
       float s=ci subconcept mapping success rate;
       float r=ci relation mapping success rate;
       integrity=w1*p+w2*s+w3*r;
       correlation+=integrity;}
      }
   return correlation/count;}
```

Fig.5 Ontology correlation-degree counting program

图 5 本体关联度计算程序

定义 6(概念关联度). 概念关联度μ表征概念间的语义相似程度.将 MedImGrid 本体中的概念使用加权边进行连接,加权边的权值为概念关联度的值,$\mu\in[0,1]$.通常将不同的概念相关关系分为如下几类:

(1) 两个概念完全相同(sameOf),即$c_i=c_j$,则$\mu=1$;

(2) 两个概念为同义词(synonymOf),表示相同类的对象;

(3) 概念间存在包含关系(subOf),如c_i是c_j的子概念;

(4) 概念存在某种联系(intersectionOf),如c_i和c_j不在同一关系树上,但存在其他联系;

(5) 概念间不存在任何关联(unrelatedOf),则$\mu=0$.

对于情况(2)~情况(4),μ通常通过对实例集(或元数据库)的词频解析获取.

定理 1(关系的可传递性). 对于概念c_i和c_j,存在关系r,使得$c_i\xrightarrow{r}c_j$成为有实际含义的语义表达,称r为语义连接关系.并且,如果在 3 个概念c_i,c_j,c_k中存在$(c_i\xrightarrow{r_i}c_j)\wedge(c_j\xrightarrow{r_j}c_k)$,则一定有$c_i\xrightarrow{r_k}c_k$存在,此时,称关系$r_i$和$r_j$是可传递的.

由此可推出无直接连接的概念之间的概念相关度计算公式,如对于 3 个概念c_i,c_j,c_k,若c_i,c_k在语义关系树上无直接连接,其关联度可由它们与c_k的关联度复合,即

$$\mu(\overbrace{c_i,c_k})=\mu(\overbrace{c_i,c_j})\times\mu(\overbrace{c_j,c_k}) \tag{3}$$

关系映射的实现具有一定的条件.设本体O_i中的一个实例$x_i=(c_{11}\xrightarrow{r_i}c_{12})$映射为$O_j$中的实例$x_j=(c_{21}\xrightarrow{r_j}c_{22})$,此时,应有$r_i$到$r_j$的双射关系存在,即概念间的语义关系在翻译过程中是不能改变的.

定义 7(实例关联度). 不同本体内的实例语句x_1和x_2,从其语义的相似程度上来讲应包含以下 3 个方面的要素:x_1和x_2所属本体的关联度,构成x_1和x_2的概念之间的关联度,关系的一致性.基于此,定义实例语义关联度Sim为

$$Sim(x_1,x_2)=\kappa(O_1,O_2)\times\sum_{i=1}^{m}\theta_i\cdot\mu(c_{1i},c_{2i})\times\prod_{j=1}^{n}Cor(r_{1j},r_{2j}) \tag{4}$$

其中,x_1(x_1={(c_i,r_j)|1≤i≤m,1≤j≤n}),x_2由m个概念和n个关系组成.θ_i为组成实例的概念的权系数(根据主次关系).函数Cor(r_1,r_2)计算两个关系的一致性,当r_1和r_2可建立双射关系时,Cor=1;否则,Cor=0.

MedImGrid平台O_1和O_2间映射策略的产生过程实际上就是语义组件扫描两个本体,根据概念关联度在相似概念及其属性和关系间建立映射链接的过程,该过程对应的程序如图 6 所示.语义检索的过程是不同本体间通过实例/概念匹配产生映射策略、定位相关数据集以及数据集间数据模式转换的过程,其详细过程见第 5.4 节.

```
void createMapschema (ontology O1,O2)
 {xmlfile file.createfile ("schema.xml");
   for (all concept ci in O1)
     {if (ci can map to cj in O2)
        {file insert node (ci,cj);
         for (all properties of ci map to cj)
           file insert subnode (ri,rj);
         for (all subconcept of ci map to cj)
           {file insert subnode (si,sj);
                for (all relation of ci to si map to cj to
                sj)
                file insert subnode (ci,sj);}
        }
     }
```

Fig.6 Mapping schema creation program

图 6 本体映射策略件产生程序

5 MedImGrid 基于语义的信息集成框架

MedImGrid 基于语义的信息集成框架如图 7 所示,它为网格中各种数据源(包括医疗信息系统、医疗数据库、专家库等)的信息集成提供了透明的语义支持.基于该框架,MedImGrid 实现了各级本体间的映射和语义转换,以挖掘不同概念间的相互关联,并支持各级本体的扩展和更新.

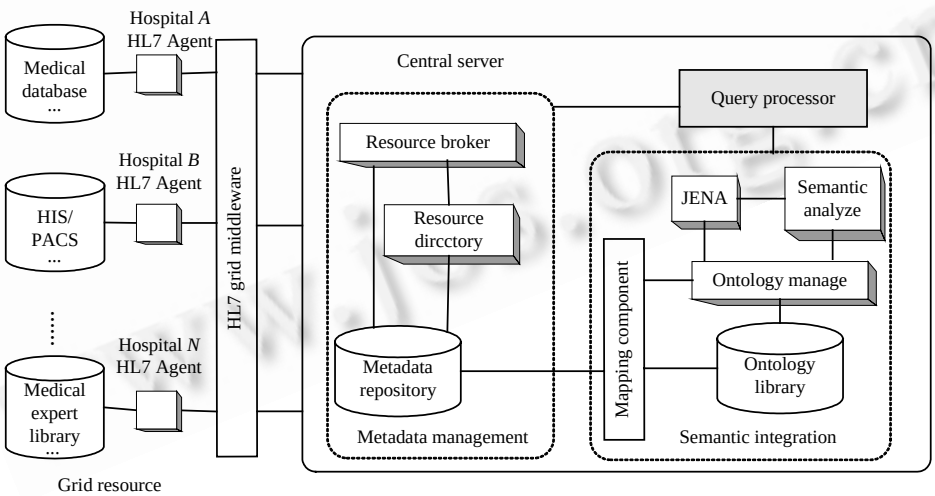


Fig.7 Semantic-Based information integration framework of MedImGrid

图 7 MedImGrid 基于语义的信息集成框架

5.1 基于HL7 Grid Middleware的智能代理

代理技术和中间件技术是实现分布式系统互操作和信息集成的有效方法.将两者有效地结合起来并在语义层次进行扩展,将使实现的代理具有一定的智能性.在MedImGrid中,每一个医疗单位可以将其信息源注册为

Grid Service,接受来自MedImGrid自治域Central Server的访问请求,并能够在HL7 网格中间件的支持下,利用HL7 智能代理访问网格中其他医院的信息资源.HL7 代理通过HL7 Message(SOAP消息)与其他医疗单位的HL7智能代理进行数据交换,实现信息的共享[2].

HL7 Message格式的消息能够被中间件转换成本地信息系统可以识别的格式,并根据需要使用本地信息系统支持的通信协议,以虚拟客户端的形式与信息系统交互.HL7 智能代理同时可以在 CGSP-DAI 支持下与开放的医疗数据库、专家库和知识库进行直接交互.该框架使得 HL7 标准可以与网格环境无缝地结合起来.众多的医疗信息系统经过网格互连在同一个虚拟组织内,在广大的地域范围内实现医疗数据的共享.基于该框架,通过医疗领域本体或应用本体的映射和转换创建本医院应用本体,并由 HL7 智能代理实现管理和更新. MedImGrid在 HL7 代理的支持下,由 Central Server 实现全局元数据的更新.全局本体能够从全局元数据、各应用本体、局部本体(local ontology)中获取语义信息.用户通过网格 Portal 提交的请求可以基于各层本体进行解析和转换提交到各数据源执行,并返回最优的匹配结果.HL7 智能代理拥有与 Central Server 相同的元数据管理组件和语义集成组件(图 7 中没有标记出来).

5.2 元数据管理组件(metadata management components)

本体的更新和维护在很大程度上需要各级元数据的支持.MedImGrid中各级元数据模型同样参照HL7-RIM建立,这使得本体和元数据有了互操作的基础.元数据管理组件管理数据源结构化与半结构化的语义信息,记录数据源间的属性和相互联系,并存储它们之间的上下文关联信息[22].

(1) 资源调度器(resource broker)用以管理和调度待集成的数据资源.查询处理器(query processor)接受用户的查询请求,并转发给语义分析器,通过语义匹配,确定哪一个数据源(或 HL7 Agents)需要访问,以满足用户或Grid Service 程序的请求.资源调度器接受语义分析器的处理结果,并查询元数据仓库,将相关资源的元数据返回给查询处理器.查询处理器根据元数据信息将用户请求分发到不同的数据源.相关语义参数由于与存储在元数据仓库中的上下文关联信息绑定,使得需要从用户请求中抽取的语义信息被精简.

(2) 元数据仓库(metadata repository)存储并管理着参照 HL7 词典从医疗数据源及其描述信息中抽取的信息所组成的元数据.它管理和维护各种元数据信息,并根据请求向资源调度器提供与查询相关的资源标识符(resource identifier)、资源属性以及上下文信息,以解释并确定语义信息中的有效部分.

(3) 资源目录(resource directory)中存放着各数据源的描述信息.使用目录的形式来管理资源,使得资源可以快速地根据需求进行索引和排序,加快资源定位和匹配的速度,如发现资源间的关系(parent-child或table-column)、属性描述以及对应的上下文信息等.目录中也存放一些语义无关的信息,如服务器地址以及网格服务的URL连接等[23].

5.3 语义集成组件(semantic integration components)

虽然元数据可以取代本体实现一些语义分析的功能,但非结构化的语义信息却不能得到有效的处理,必须将这些元数据进行进一步抽象,构建相应的本体以支持语义互操作.

(1) 本体仓库(ontology library)中存放了 MedImGrid 的全局本体、各局部本体和本体关联度表,它们采用HL7-RIM 作为父本体,其中也存放了一些医疗领域本体,以使平台能够支持一些特殊粒度的语义请求,使用最佳匹配的本体来进行语义操作.这些领域本体还有助于我们更好地理解用户的请求,因为我们可以用不同的本体来对它们进行解析,从多个语义环境解释用户请求.

(2) 本体管理器(ontology manage component)用以记录和管理各本体术语、概念之间的关系属性、推理属性、映射信息等,并缓存已经产生的本体映射策略文件,以支持本体之间的映射和转换.通常情况下,数据源的属性和数据模式都可以与 HL7-RIM 本体中概念的属性关联起来.

(3) 语义分析器(semantic analyze component)存放了用户请求使用本体进行解析的规则,以及在特定条件下数据源属性的解释和抽取方法.它储存的解析规则用以约束资源调度器向查询处理器返回元数据.该组件以JENA来实现[24],相关推理基于用户请求、映射信息以及MedImGrid本体间的关联关系.

(4) 语义映射组件(mapping component)实现了各种MedImGrid本体之间的映射,以将某一本体解析的结果映射到其他本体所支持的系统中(如概念到概念的映射),在语义层实现了数据模式的转换.由于MedImGrid中的各本体基于同样的词汇表,或拥有同样的父本体HL7-RIM Ontology,因而本体间的映射和转换的成功率 较高[23].

5.4 语义映射的实现过程

MedImGrid 支持以手工的方式建立本体间的映射关系,并能够以 XML Schema 的形式定义 MedImGrid 本体(概念关系模型)到底层数据源的本体(概念关系模型)或数据库存储模型(数据库-表-字段)的映射规则,包括映射方法、访问模式和关联规则等.例如本体到数据库的映射规则(ontology-to-database)定义了本体概念所对应的数据库表字段的位置以及建立表间相互关联的主关键字等信息,使得对概念实例的查询直接映射为到数据库字段内容的操作.

MedImGrid 语义查询映射的实现过程可以分为以下几个步骤:

(1) 请求发起者(用户或 grid service)向 MedImGrid 的 Central Server 的查询处理器发送请求(由概念和实例组成).

(2) 语义分析器利用全局本体对请求进行语义解析并将结果发送给资源调度器;查询处理器利用资源调度器返回的元数据信息将任务分配到不同的 HL7 智能代理上.

(3) HL7 智能代理获取需要进行翻译转换的实例语句,根据语义标记判断语句的源本体和转换后的目标本体.

(4) HL7 智能代理检索 MedImGrid 本体树对应本体关联度表,查找源本体和目标本体之间的关联度值,若为 0,则返回.

(5) 对源实例语句进行最大程度的分解,获取语句对应的语义网络的概念、关系和实例的集合.

(6) 查找映射的实现是否包含待转换的关系,如果不包含,则返回.

(7) 按照概念关联度最大的原则选取目标本体中的概念,检索选定的概念间是否存在与源本体概念间相同的关系,如果不存在,则返回(6),重新选取目标概念.

(8) 计算选定的源实例语句与转换后的目标实例语句的语义关联度,如果大于预设阈值,则创建映射策略文件,并将结果(也就是解析后的操作请求)发送至各医疗信息系统或数据库中进行数据操作.若存在下一级医疗信息系统的应用本体(如 RIS 等),则可以按照同样的转换步骤进行.

(9) 获取的结果根据映射策略完成数据模式的转换,按照全局本体表述的数据模式进行集成,并返回请求发起者.

6 测试环境与实验

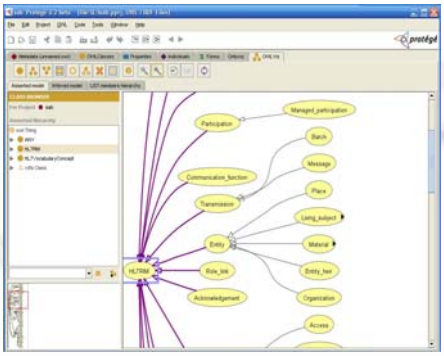
6.1 MedImGrid原型系统的实现

MedImGrid 医学图像网格原型系统基于我国第一个 IPv6 国家主干网(CERNET2),互联了域管理中心(domain center)和 3 个医疗数据资源服务节点.域管理中心包括 MedImGrid 门户服务器(portal server)、中心服务器(central server)和域信息服务器(domain information server).异构医疗数据资源服务节点由华中科技大学 CGCL 实验室、医学图像中心和同济医院信息中心组成,各节点的数据服务器中存储了由医疗数据库或信息系统管理的真实医疗数据.系统中各服务器具有接近的软、硬件配置(PIV1.2G XEON CPU/1G Memory/Red Hat Linux 9.0).

MedImGrid当前版本基于HL7 toolkits,JENA 2.4[24],JAVA和CGSP2(中国教育科研网格公共支撑平台v2.0)的API函数实现,支持IPv6 环境下的网格应用.为了简化语义数据管理和维护的操作,我们使用PostgreSQL作为元数据与本体库.当前平台存储了元数据(121MB)和本体(3MB).如图 8(a)所示,我们将HL7 RIM的部分信息导入 Protégé 2000 并进行扩展,建立了HL7-RIM Ontology,包含 80 个类、178 个关系属性、101 个推理属性和 110 个

实例等.各数据资源节点基于HL7-RIM Ontology建立了本地的局部Ontology.

在数据集成过程中,部署语义解析组件的中心服务器或 HL7 智能代理可以在语义映射完成后,根据匹配的语义网络子图来做进一步的查询优化和分解.如用户希望查找“武汉同济医院,CGCL 实验室的肺气肿治疗的DR 图片”,中心服务器通过概念/实例匹配,将查询分解为与同济医院和 CGCL 实验室的 HL7 智能代理对应的两个子查询.其中,同济医院的子查询又可以被 HL7 智能代理分解为 3 个子查询:“同济医院各个科室”、“每个科室诊断过的肺气肿病人”、“病人病例中诊断图片类型为 DR”,然后由智能代理按照预定的流程对不同的数据源进行操作,并将子查询结果进行整合,返回给用户.在以上流程中,包含着语义组件实现不同的本体标记对应的数据模式之间的转化过程.这与基于关键字的查询对数据源的整体扫描的方法有很大的不同.MedImGrid 的语义功能使查询的隐含语义能够被中心服务器和 HL7 智能代理识别并进行优化,降低了查询的总体开销.此外,MedImGrid 有效地支持 HL7 智能代理,实现了不同数据模式之间的转化.MedImGrid 实现的对分布在不同数据资源节点肺气肿病例检索的实例如图 8(b)所示.



(a) HL7-RIM ontology deployed in Protégé 2000

(a) Protégé 2000 中展开的 HL7-RIM 本体

(b) Semantic based information integration instance

(b) 基于语义的信息集成实例

Fig.8 Semantic based information integration of MedImGrid

图 8 MedImGrid 基于语义的信息集成

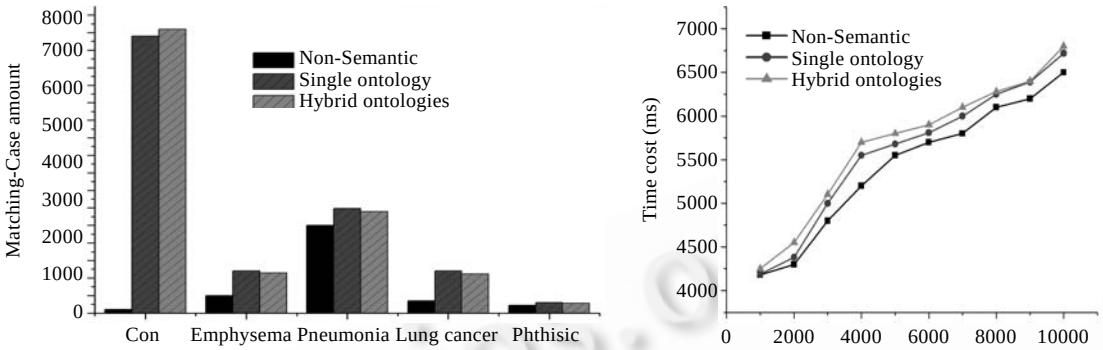
6.2 MedImGrid的性能评测

本文在相同的软、硬件资源的基础上,通过实验对比了 MedImGrid 在单一本体、混合本体、无语义支持条件下的性能.由于多本体策略在 MedImGrid 平台上不备具可行性,因此我们没有对该方法进行验证.无语义支持的方法应用了与 eDiaMoND 项目相同的中间件(OGSA-DAI),并采用基于关键字的方法对数据源进行检索.

对语义解析的支持使得 MedImGrid 能够深层理解用户的需求.用户在提交信息查询请求时只需关注概念的实际语义,而不是一个个具体的关键字.图 9(a)对比了 3 种方法对语义关系树(位于同一“肺部疾病”关系树上)不同层次的概念/实例查询的命中率.从中我们可以看出,基于语义的检索方法在对关系树高层节点概念进行检索的应用中的优势非常明显.基于单一本体和混合本体的方法在这一测试中表现出来的性能基本相同,而基于关键字(非语义)的方法则明显丢失了大量的相关病例.MedImGrid 的语义支持对医疗数据的挖掘及集成应用(如医疗数据建模、流行病研究),具有重要意义.

语义映射的实现一般都要经过多次叠代.图 9(b)对比了 3 种方法访问和整合不同的数量的 EPR(电子病历记录,average size=1K)所产生的时间开销,主要包括概念/实例匹配、查询路径产生、映射策略生成、数据查询及模式转换和网络传输 5 个部分.限于数据源的数量和性能,实验中基于语义的方法相对于查询路径优化的优势不是十分显著.3 种方法的数据查询及模式转换部分在时间开销中占有的比例最大.它们在分布的节点上并行完成,一般由耗时最大的节点来决定,对三者而言,这个时间接近相同.单本体方法的本体结构比较简单,其在概念/实例匹配、查询路径产生及映射策略生成过程中产生的时间开销要低于基于混合本体的方法,但两者在

这些过程中的时间差值不随 ERP 记录的增加而增加.非语义方法采用一对一映射的策略,不需要中间计算过程产生数据模式映射所需的 XML 文件;其概念/实例匹配以及查询路径产生采用基于关键字的方法,同样不随 EPR 记录的增加而增加.因此,实验中 3 种方法的时间差距始终保持接近.今后,对于 MedImGrid 语义映射和查询路径生成性能的优化将是研究的重点.



(a) Resource matching (b) EPR access and integrality

(a) 资源匹配 (b) EPR 访问和整合

Fig.9 Performance evaluation of MedImGrid

图 9 MedImGrid 的性能评价

MedImGrid 的信息集成框架和本体组织与 Infosleuth 系统有一定的相似性.eDiaMoND 基于 WSRF 框架的体系结构与 MedImGrid 十分相似.表 1 对 3 种原型系统的性能进行了对比分析.eDiaMoND 和 Infosleuth 绕过了原有的本地信息系统,通过直接的数据库操作(OGSA-DAI 和 JDBC)实现了对数据库的访问,牺牲了部分本地系统的安全性.它们对医疗数据库的相似性有一定的要求,系统的扩展和升级有一定的困难,且对动态增加的信息源无法进行有效的支持.MedImGrid 通过全局和局部本体的结合构建了全局资源视图;通过基于本体标记的语义操作实现了概念/实例对应的数据模式在不同信息系统间的转换和映射;通过 HL7 智能代理实现了与医疗信息系统的直接交互,并可以兼顾本地信息系统的访问控制模式.MedImGrid 基于语义的信息集成方法对于网格环境下海量异构信息集成的实现具有较大的优势和应用潜力.

Table 1 Performance contrast of three prototype-systems

表 1 3 种原型系统的性能对比

Evaluation type	eDiaMoND	Infosleuth	MedImGrid
Integration means	Structure	Semantic	Semantic
Efficiency	Low hit integration rate	Increased respond delay	Increased respond delay
Cost	Reprogram DAI interface	Reprogram JDBC interface	Directly with HL7 interface
Security	Low, access DB directly	Low, access DB directly	High, virtual client access
Auto-Adaptation	No	No	Yes
Manual degree	High, all manual	Low, semi-automatic	Low, automatic

7 总结和展望

医院数字化程度和健康保健社会关注度的提高,对当前医疗信息服务的质量提出了更高的需求.其中,跨医院、跨地区的医疗信息共享对医疗服务和医学研究具有重要意义,需要尽快建立起来.MedImGrid 基于语义的医疗信息集成方法,采用混合本体策略建立了全局本体和局部本体.在语义技术支持下,结合代理和中间件技术开发了 HL7 网格中间件,在对异构数据源进行 Grid Service 封装的基础上,实现了支持医疗语义解析的 HL7 智能代理;基于各级本体完成了数据模式的转换,实现了在统一的语义层次上的模式匹配和知识共享.MedImGrid 将在大规模测试和应用的基础上,建立安全控制模型、数据挖掘模型,以支持疾病建模与流行病预警等高级应

用,为全国范围的医疗卫生信息基础设施建设的相关研究做出贡献.

致谢 在此,我们向对本文的工作给予支持和建议的各位老师以及在系统实现中提供帮助的同学表示诚挚的感谢.

References:

[1] Foster I, Kesselman C, Eds.; Jin H, Yuan PP, Shi K, Trans. The Grid 2: Blueprint for a New Computing Infrastructure. 2nd ed., Beijing: Publishing House of Electronics Industry, 2004. 1−10 (in Chinese).

[2] Jin H, Sun AB, Zhang Q, Zheng R, He RH. MIGP: Medical image grid platform based on HL7 grid middleware. In: Yakhno TM, ed. Proc. of the Advances in Information Systems 2006. Berlin: Springer-Verlag, 2006. 254−263.

[3] Benetti H, Beneventano D, Bergamaschi S, Guerra F, Vincini M. An information integration framework for e-commerce. IEEE Journal on Intelligent Systems, 2002,17(1):116−122.

[4] InfoBus 1.1 specification. 1999. http://java.sun.com/products/archive/javabeans/infobus/infobus1.2.pdf

[5] Chawathe S, Garcia-Molina H, Hammer J, Ireland K, Papakonstantinou Y, Ullman J, Widom J. The TSIMMIS project: Integration of heterogeneous information sources. 2002. http://www.cise.ufl. edu/~jhammer/publications/tsimmis-overview.pdf

[6] Angle J. OntoBroker. 2002. http://www.zope.org/Members/ontoprise/obc/ontobroker_2002_english.pdf

[7] The InfoSleuth Agent system. http://www.argreenhouse.com/InfoSleuth/

[8] Kroeger W, Hasan A, Hanushevsky A, Martin L, Nief JY, Boutigny D, Petzold A. Babar data distribution using the storage resource broker. IEEE Trans. on Nuclear Science, 2004,51(4):1462−1464.

[9] Lloyd S, Simpson A. Project management in multi-disciplinary collaborative research. In: Davis M, ed. Proc. of the Professional Communication Conf. 2005. New York: IEEE Press, 2005. 602−611.

[10] Crompton SY, Matthews BM, Gray WA, Jones AC, White RJ, Pahwa JS. OGSA-DAI and bioinformatics grids: Challenges, experience and strategies. In: Tumer SJ, Lee BS, Cai WT, eds. Proc. of the Cluster Computing and the Grid. Washington: IEEE Computer Society Press, 2006. 8−16.

[11] McClatchey RH, Manset D, Solomonides AE. Lessons learned from MammoGrid for integrated biomedical solutions. In: Dillon TS, ed. Proc. of the 19th IEEE Symposium on Computer-Based Medical Systems. New York: IEEE Press, 2006. 745−750.

[12] Amendolia SR, Estrella F, Hauer T, Manset D, McClatchey R, Odeh M, Reading T, Rogulin D, Schottlander D, Solomonides T. Grid databases for shared image analysis in the Mammogrid project. In: Bernardino J, Desai BC, eds. Proc. of the Int'l Database Engineering and Applications Symp. (IDEAS 2004). New York: IEEE Press, 2004. 302−311.

[13] Goderis A, Li P, Goble C. Workflow discovery: The problem, a case study from e-science and a graph-based solution. In: Yan YH, ed. Proc. of the ICWS 2006. New York: IEEE Press, 2006. 312−319.

[14] Tang J, Liang BY, Li JZ, Wang KH. Automatic ontology mapping in semantic Web. Chinese Journal of Computers, 2006,29(11):1956−1976 (in Chinese with English abstract).

[15] Chen G, Lu RQ, Jin Z. Constructing virtual domain ontologies based on domain knowledge reuse. Journal of Software, 2003,14(3): 350−355 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/350.htm

[16] Cui W, Wu H. Using ontology to achieve the semantic integration and interoperation of GIS. In: Moon WM, ed. Proc. of the Int'l Geoscience and Remote Sensing Symp. 2005. Washington: IEEE Computer Society Press, 2005. 25−29.

[17] Wache H, Vögele T, Visser U, Stuckenschmidt H, Schuster G, Neumann H, Hübner S. Ontology-Based integration of information—A survey of existing approaches. In: Stuckenschmidt H, ed. Proc. of the IJCAI 2001. New York: IEEE Press, 2001. 108−118.

[18] HL7 v.3.0 introduction. 2007. http://www.hl7.org/

[19] The protégé 2000. 2001. http://protege.stanford.edu/plugins/owl/

[20] Deen SM, Ponnamperuma K. Dynamic ontology integration in a multi-Agent environment. In: Fu X, ed. Proc. of the Advanced Information Networking and Applications 2006. Washington: IEEE Computer Society Press, 2006. 6−18.

[21] Chen L, Han Y, Li SL. Dynamic integration and construct of Web services based on ontology in information grid. Journal of Software, 2006,17(11):2255−2263 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/2255.htm

[22] Amann B, Beeri C, Fundulaki I, Scholl M. Ontology-Based integration of XML Web resources. In: Horrocks I, ed. Proc. of the ISWC 2002. Berlin: Springer-Verlag, 2002. 117−131.

[23] Chong Q, Marwadi A, Supekar K, Lee Y. Ontology based metadata management in medical domains. Journal of Require and Practice in Information Technology, 2003,35(2):139−154.

[24] Jena semantic Web framework. 2006. http://jena.sourceforge.net/

附中文参考文献:

[1] Foster I, Kesselman C,编;金海,袁平鹏,石柯,译.网格计算.第 2 版,北京:电子工业出版社,2004.1−10.

[14] 唐杰,梁邦勇,李涓子,王克宏.语义 Web 中的本体自动映射.计算机学报,2006,29(11):1956−1976.

[15] 陈刚,陆汝钤,金芝.基于领域知识重用的虚拟领域本体构造.软件学报,2003,14(3):350−355. http://www.jos.org.cn/1000-9825/14/350.htm

[21] 陈磊,韩颖,李三立.信息网格中基于本体的 Web 服务动态集成和重构.软件学报,2006,17(11):2255−2263. http://www.jos.org.cn/1000-9825/17/2255.htm

金海(1966－)男,湖北武汉人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机系统结构,集群计算,网格计算,并行与分布式计算,对等计算,普适计算,语义网,存储与网格安全.

何儒汉(1974－),男,博士生,主要研究领域为图像网格,语义网,信息检索.

孙傲冰(1978－),男,博士生,主要研究领域为图像网格,信息集成.

章勤(1955－),女,教授,主要研究领域为图像处理,系统结构.

郑然(1977－),女,博士,讲师,CCF 学生会员,主要研究领域为网格计算,网格应用.

吴松(1975－),男,博士,副教授,主要研究领域为网格计算,网格存储.