文章编号:1001 - 9081(2008)09 - 2324 - 04

基于 RDF 的语义网格数据建模与检索

师雪霖,赵 英
(北京化工大学 信息科学与技术学院,北京 100029)
(shixl@mail.buct.edu.cn)

摘 要:语义网格所需要处理的信息通常为半结构化数据,如何以合理的模型表示这些半结构化数据并实现高效查询处理,是语义网格要解决的核心问题之一。提出了一种基于资源描述框架(RDF)的半结构化数据表示模型,并设计了相应的信息检索机制。最后介绍了一个基于化工计算网格平台的,实现了化工领域知识共享与检索的化工语义网格架构的设计与实现。

关键词:语义网格;资源描述框架;半结构化数据;信息检索
中图分类号: TP393 **文献标志码**:A

Data modeling and retrieval in semantic grid based on RDF

SHI Xue-lin, ZHAO Ying
(*School of Information Science and Technology, Beijing University of Chemical Technology, Beijing* 100029, *China*)

Abstract: The data processed by semantic grid are often semi-structured data. Therefore, how to model and retrieve these semi-structured data is a key issue of semantic Web. The presenting model for semi-structured data base Resource Description Framework (RDF) and its information retrieval mechanism were proposed in this paper. Then a semantic grid infrastructure was described, which implemented knowledge sharing and retrieval in chemical engineering domain based on a chemical computing grid platform.

Key words: semantic grid; Resource Description Framework (RDF); semi-structured data; information retrieval

0 引言

网格计算已经成为当今分布式计算和 Internet 应用的新趋势,它将分散的异构资源整合起来,为用户提供透明的、协作计算环境,从而使达到保证所有用户能最大限度地分享高性能计算服务的目的[1]。语义网是在 Web 的基础发展而来的,它的目的是在 Web 信息中加入可供机器处理的元信息,从而使 Web 提供更高质量的服务[2]。语义网与网格计算相结合产生了一个新的研究领域:语义网格。语义网格的目的是提供这样的一种架构,使得所有的资源和服务可以简单地发布、自动管理,从而实现更灵活的人机协作[3],如图 1 所示。

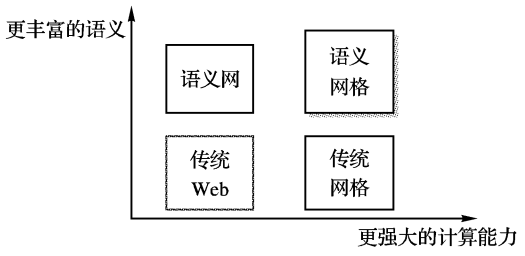

图 1 语义网格的发展

目前国外一些围绕语义网格的项目已经展开,如英国的 e-Science[3]项目等。对化学工程领域的研究来说,经常需要复杂的计算以实现分子结构设计模拟,同时还需要处理分析大量相关领域知识,以实现研究目标。因此,这不仅需要高性能计算(High Performance Computing, HPC)环境,更需要一个领域知识共享环境,能够运用和处理相关语义信息。

根据这种需要,在北京化工大学原有化工网格平台的基础上,初步搭建了一个化工语义网格架构(Chemical Engineering Semantic Grid Infrastructure, CESGI)原型系统。该系统主要针对量子化学和分子工程研究人员的需要,利用网格平台的高性能计算环境,为这些用户提供该领域知识的语义检索服务。系统架构基于网格中间件平台设计,易于部署、移植,能充分利用网格计算环境。为实现上述目标,我们设计了一种基于资源描述框架(Resource Description Framework, RDF)的领域数据表示模型,并在此基础上设计了相应的查询机制。

1 RDF 数据模型及检索机制

目前 CESGI 提供的化工领域知识主要来自于 Internet 上的相关 Web 站点以及部分开放化工数字资源库。这些数据的一个主要特点是半结构化:即主要为半结构化文本、HTML 页面、XML 文档,且具有一定的描述信息,即元数据(如主题、时间、分类等信息)。基于这个特点,在实现这些数据的整合、使之成为计算机可理解的语义信息时,适合采用 RDF 作为信息表示模型。基于 RDF 模型表示的数字资源中包含明确的语义描述,使计算机系统能够更准确地对信息资源进行过滤、分类和搜索。

RDF 是 W3C 继 XML 之后提出的一种新标准,XML 和 RDF 都是表示半结构化数据的标准。RDF 是为了描述 Web 元数据而建立的标准[4],它采用对象-属性-值三元组的结构来表示数据和数据之间的关系。RDF 与 XML 相比,其显著优势在于 RDF 对语义信息的支持。XML 的设计初衷是提供一

收稿日期:2008 - 03 - 11;**修回日期**:2008 - 05 - 29。 **基金项目**:北京化工大学"青年教师基金(QN0732)"和"化工网格"项目。
作者简介:师雪霖(1977 -),女,甘肃人,讲师,博士,主要研究方向:语义网、网格计算、Web 挖掘; 赵英(1962 -),男,天津人,教授,博士,主要研究方向:网格计算、计算机网络。

种灵活的文档表示结构,它关心的是文档的结构,而不是文档中数据所包含的语义信息。而 RDF 提供了灵活的机制使得用户可以根据实际需要,定制应用领域内资源、知识的表达规则、语义信息。涉及到语义互操作时,XML 的功能则远远比不上 RDF。因此,实现对半结构化数据源的整合、语义处理时,采用 RDF 作为数据表示模型,更有利于实现语义互操作。

1.1　模型层次结构

根据 RDF 描述半结构化数据的特点及其模式定义描述能力,划分了 RDF 数据模型的层次(如图 2 所示):RDF 数据层(RDF Data)、定制的 RDFS 层(Customized RDFS)和元 RDFS 层(Meta RDFS)。RDF 数据层是描述具体资源属性、关系的 RDF 语句。定制 RDFS 层定义了描述这些资源的类(class)。元 RDFS 层则定义了这些类之间的关系,提供了描述 RDF 的基本的模式。

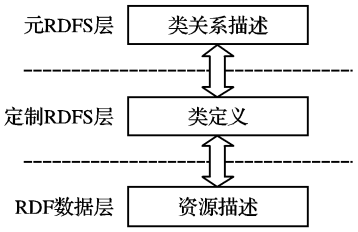

图 2　RDF 层次图

顶层的元 RDFS 预先定义了 RDF 的 3 个核心类:Resource、Property Type 和 Class;还定义了核心属性类型如 instanceOf、subClassOf、range 和 domain 等。根据这些基本的类和属性,派生出专门描述某一个领域资源的类和属性类型,形成第二层——定制的 RDFS。简而言之,RDF 数据层是定制 RDFS 层所定义的类的实例,而元 RDFS 层则是说明定制 RDFS 层的元数据。

1.2　数据模型(RDF/S-Graph)

RDF 的基本结构是对象-属性-值三元组:对象 O 具有属性 A,属性值为 V,通常记为 $A(O,V)$。还可以采用 *RDF* 图来表示这种三元组关系:对象 O 和值 V(即 RDF 资源)用节点表示,属性 A 为连接两节点的边,即 $[O]-A-[V]$。因为通常一个对象也可以作为其他对象的值,所以采用图来表示更为实用明了。

RDFS 定义了描述 RDF 资源的类的信息以及 RDF 属性的取值范围。RDFS 本身也是 RDF 文档,因此也可以使用图来表示:RDFS 图的节点代表类,边表示属性的类型。通过 RDF/S 图可以明确表示出 RDF 文档的语义信息。下面给出它们的形式化定义。

定义 1　RDF/S-Graph。RDF/S-Graph 是一个边加标签的有向图,可用四元组 $RG=(V, E, D, \lambda)$ 表示,其中:

1)V 表示节点的集合,即 RDF 所描述资源的集合或 RDFS 中定义的类的集合;

2)E 表示边的集合,即 RDF 资源的属性集合或 RDFS 中类的属性的集合;

3)所有节点来自域 D,D 是 RDF/S 所描述的信息适用的领域,在此领域之外,所描述的知识不再成立;

4)λ 是边到节点的映射,即 $\lambda: E \rightarrow V \times V$。

我们提出的这种简化的、边加标签的有向图作为表示 RDF 文档(包括 RDFS)的数据模型,可以明确表达 RDF 文档所包含的语义信息,更重要的是能够清晰地描述出类、属性之间的语义联系。RDF/S-Graph 模型对语义查询机制的构造具有两方面重要意义:首先,描述查询语言的谓词逻辑必须建立在模型所描述的逻辑关系基础之上;其次,查询机制在实现语义查询时,必须满足 RDF/S-Graph 所描述的形式和关系。在此模型基础上,通过对图的遍历,实现对 RDF 数据所包含的语义、语义联系的查询。

定义 2　RPath 路径。RPath 路径是形如:$\{X_0\}.att_1\{X_1\}.att_2\{X_2\}.\cdots.att_n\{X_n\}$ 的表达式。其中:$n \geqslant 0$;X_0,…,X_n 为对象标识、对象变量或对象常量;att_1,…,att_n 为对象的属性名或者属性变量。RPath 定义了查询语句中的查询范围。

1.3　基于 RDF 模型的半结构化数据查询机制

在上述 RDF 数据模型的基础上,我们设计了一种查询机制以实现对领域知识的语义检索。首先设计了一种融合了谓词逻辑的说明式查询语言 RQuery[5],它定义了基本查询谓词和逻辑操作符,通过操作符可以构造更复杂的查询语句。RQuery 不仅可以实现对 RDF 数据的查询,还可以实现对 RDFS 的查询,两种查询统一起来可以给用户提供超出普通信息查询更多的知识:信息之间的关系,相关信息等。下面给出了用 BNF 范式(Backus-Naur Form)描述的 RQuery 的词法和语法概要:

```
< RQuery > :: = < query >
< query > :: = SEARCH < item-list > FROM < path-expr > [ WHERE
    < cond-expr > ]
    |( < query > )
    |subClassOf[^]( < query > )
    |superClassOf[^]( < query > )
    |subPropertyOf[^]( < query > )
    |superPropertyOf[^]( < query > )
< cond-expr > :: = < var > < comp-op > < var >
   | < cond-expr > < bool-op > < cond-expr >
< item-list > :: = * |identifier{, identifier}
   | < query > {, < query}
< path-expr > :: = [ < data-var > ] [; ( < class-var > | < type-var > |
   identifier) ]
    | < class-var > | < type-var >
< comp-op > :: = < | < = | > | > = | = |! = |LIKE
< bool-op > :: = AND|OR
< var > :: = < data-var > | < class-var > | < type-var > | < property-
   var >
< data-var > :: = identifier
< class-var > :: = % identifier
< type-var > :: = % % identifier
< property-var > :: = @ identifier
```

目前,RQuery 仅仅定义了基本的查询语句 < query > 的语法。下面简要介绍各部分的含义,并举例说明了 RQuery 使用过程。SEARCH 是 RQuery 语句的开始,类似于 SQL 语言中的 SELECT,其后是所要查询的内容的投影列表。FROM 相当于 SQL 语言中的 FROM,指定了所查询的范围,在 RQuery 中,查询范围由 RPath 表达式 < path-expr > 来表示。WHERE 后面为条件语句 < cond-expr >,限定了查询条件。条件语句 < cond-expr > 可以是由比较操作符 < comp-op > 组成的单独表达式,也可以由布尔操作符 < bool-op > 连接多个条件语句。此外,RQuery 还支持嵌套查询,SEARCH 所包含的投影列表中可以嵌套其他完整的查询语句 < query >。当不确切某一 RDF 资源的模式描述时,不仅需要查询 RDF 数据,还需要查询 RDFS,此时的查询请求就需要使用嵌套查询。下面引入一个例子说明 RQuery 语言的使用情况。

q:查询类“chorography”的所有属性

q′：构造的 RQuery 查询语句，如下所示：

```
SEARCH @P
FROM {%C}.@P
WHERE %C = chorography
```

其中% C 为类变量 < class-var >，@ P 为属性变量 < property-var >，该查询将返回 chorography 的全部属性。

其次，为了实现语义查询，设计了分子资源元数据标引体制。所有的领域资源数据在使用 RDF 模型进行标准化转换存储中，自动提取元数据信息。我们采用了 DC（Dublin Core）元数据体制来描述所有数据的公共元数据（Common Metadata，CM），DC 包括 15 个基本著录项，足以表示数据来源、更新时间、语种等公共信息。在此基础上，根据量子化学和分子工程的领域特点，设计了一些字段表示专业元数据（Special Metadata，SM），如分子类型、物理性质、化学性质、分子结构等。通过 CM 和 SM 的共同标注，在原有信息中加入可供机器处理的元信息，从而可实现更准确的语义检索。此外，结合 RPath 表示方式，可以很清晰定义检索范围、实现准确检索。

2 化工语义网格架构（CESGI）

CESGI 是在化工网格平台的基础上构架的，它的目标为满足量子化学和分子工程研究人员的需要，利用网格平台的高性能计算环境，为这些用户提供该领域知识的语义检索服务。

2.1 化工网格平台

化工网格平台是基于 CGSP（China Grid Support Platform）搭建的网格计算平台，主要提供化工领域的高性能计算服务[6]。CGSP[7]是国际上第一个遵循 OGSA（Open Grid Service Architecture）架构的、参照 WSRF（Web Service Resource Framework）规范实现的网格中间件。CGSP 构建在 Globus Toolkit 3 core 之上，把网络上的硬件计算资源、存储资源、网络资源整合起来，并将这些资源包装为服务呈现给用户。

基于 CGSP 构架的化工网格平台的框架如图 3 所示。

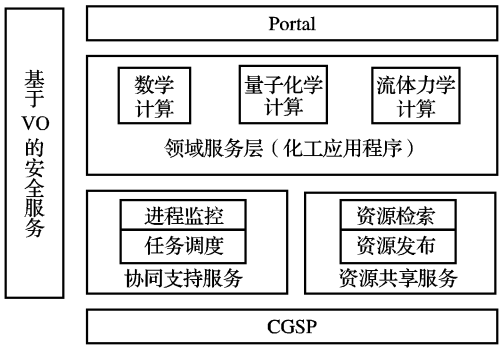

图 3 化工网格平台框架

化工网格平台由三层服务构成：底层为网格中间件层（即 CGSP）；中间层是在 CGSP 基础上开发的通用服务层，提供协作支持服务和资源共享服务；第三层为领域服务层，部署了化工领域常用的计算程序。在三层服务的基础上，为用户提供 Web 方式的 Portal 接口，供用户使用网格资源。为了保证资源和数据的安全性，设计了基于虚拟组织（Virtual Organization，VO）安全服务。

2.2 化工语义网格框架（CESGI）

在化工网格平台基础上，我们构架了 CESGI。从用户角度看，CESGI 为化工网格上提供的服务之一，用户可通过化工网格的 Portal 提交相应的检索请求，CESGI 接受请求后，查询领域知识库，获取查询结果。最后结果由 Portal 统一提供给用户。图 4 给出了 CEGSI 的框架。

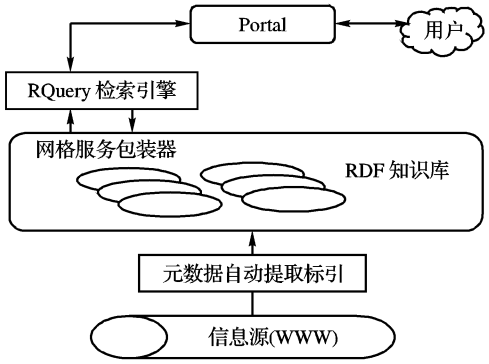

图 4 CESGI 框架

采用本文第二章描述的 RDF 数据模型和元数据标引机制，CESGI 首先实现来自不同数据源的资源的元数据提取标引，将其转换成基于 RDF 数据模型表示的语义记录，存放在 RDF 知识库中。用户提交的查询请求被转换成 RQuery 格式，由检索引擎实现对 RDF 知识库的语义检索。CESGI 作为化工网格上部署的服务之一，需要包装成可供网格平台集成的服务，因此所有功能由网格服务包装器（Grid Service Wrapper）包装成符合 WSRF 规范的、支持 SOAP（Simple Object Access Protocol）的 Web Service，使用 WSDL（Web Service Define Language）定义后发布在化工网格平台上。

3 应用工作流示例

CESGI 作为化工网格的应用之一，目前提供了化工领域专业知识的语义检索服务。下面以一次具体的用户请求服务执行的过程为例说明 CESGI 的具体工作流：用户通过 Portal 提交检索任务请求；Portal 将用户请求以 SOAP 信息的方式传递给网格平台中的任务调度模块；任务调度模块查询到 CESGI 提供的语义检索服务，调用该服务以完成任务。该服务具体运行的工作流入下：

1）RQuery 检索引擎接收服务请求——用户查询请求 q；

2）RQuery 检索引擎对 q 进行解析转换，将 q 翻译成 RQuery 语句 q'；

3）查询 RDF 领域知识库（异构数据源的信息进行元数据提取与标引后，以 RDF 模型存储在领域知识库中），查询结果以 RDF 格式表示，记作 $\{r1, r2, \cdots\}$；

4）查询结果返回给调用者；

5）此次查询请求处理完成。

执行完上述操作后，最终的信息结果返回 Portal，展示给用户。本文采用了 RDF 作为数据模型，设计了同时支持模式查询和数据查询的 RDF 查询机制 RQuery，因此在检索时，不仅能查询到用户所需信息，而且可以将其语义关系返回给用户。以炼油催化剂信息查询为例，来自网络的相关信息（如 HTML 网页）等，经过元数据提取标引，以串行化的 RDF 数据格式存放在 RDF 知识库中，如下所示：

```
< rdf: RDF >
  < rdf: Description about = "catalyst for pyrolysis oil production"
    < ch: molecule > Fe2O3 < /ch: molecule >
    < ch: physical_prop > dark-red sediment < /ch: physical_prop
    >
    < ch: chemical_prop >
      oil yield can be improved about 76.7% when using Fe2O3
      as catalyst
    < /ch: chemical_prop >
    < url_addr > http: //... < /url_addr >
```

</rdf: Description >
</rdf: RDF >

使用上述模型后,用户除了可以按照传统的关键词匹配检索信息,还可以根据催化效应、化学性质、物理性质等语义内容进行检索,从而提高了检索效率。图 5 为 CESGI 的检索 portal 页面,用户可先使用导航功能浏览语义关系,然后提交检索请求。

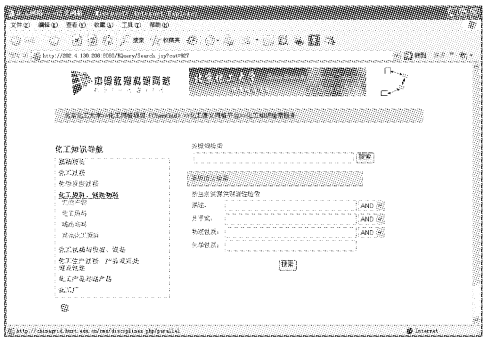

图 5　检索 Portal 页面

4　结语

本文提出了一种基于 RDF 的半结构化数据表示模型,并设计了相应的检索机制。以此为基础,基于化工计算网格平台构建了实现化工领域知识共享与检索的化工语义网格架构。该系统主要针对量子化学和分子工程研究人员的需要,利用网格平台的高性能计算环境,为这些用户提供该领域知识的语义检索服务。

本架构实现了异构数据源的集成信息检索,具有如下优势:首先系统架构基于网格中间件平台设计,采用 Web Service方式实现,易于部署和移植;其次RDF作为支持语义的 Web 元信息表示标准,已经成为构建下一代 Web——Semantic Web 的基础[8],使用 RDF 作为半结构化数据表示模型,利于对信息的语义处理加工,以便为用户提供更高层次的知识服务。

参考文献:

[1] FOSTER I, KESSELMAN C. The grid: blueprint for a future computing infrasturcture [M]. San Fransisco : Mogan Kaufmann Publisher, 1999.

[2] DAVIES J, FENSEL D, van HARMELEN F. Towards the semantic web: ontology-driven knowledge management [M]. New York: John Wiley & Sons, Ltd, 2003.

[3] De ROURE D, JENNINGS N, SHADBOLT N R. Research agenda for the semantic grid: A future e-science infrastructure. Technical Report UkeS-2002-02, [R/OL]. [2008 - 01 - 01]. National e-Science Center: Edinburgh, UK, 2001, http:// www. semanticgrid. org/v1. 9/semgrid. pdf.

[4] LASSILA O, SWICK R. Resource Description Framework (RDF) model and syntax specification [EB/OL]. [2008 - 01 - 05]. http://www. w3. org/TR/REC-rdf-syntax/.

[5] 师雪霖,牛振东,宋瀚涛. 一种基于 RDF 的半结构化数据查询语言 [J]. 计算机工程,2006,32(5):13 - 14.

[6] ZHAO YING, SHI XUE-LIN. Collaborative computational chemical grid based on CGSP [C]// 2007 IFIP International Conference on Network and Parallel Computing Workshops (NPC 2007). [S. l.]: IEEE Press, 2007: 199 - 202.

[7] HAI JIN, TAO YONG-CAI, WU S, *et al*. China Grid: Making grid computing a reality [C]// Proceedings of the 2008 conference on Computing frontiers. New York: ACM Press, 2005: 13 - 24.

[8] The semantic Web-on the respective roles of XML and RDF [EB/OL]. [2008 - 01 - 01]. http://www-db. stanford. edu/ ~ stefa/.

(上接第 2323 页)

假设 1 指出求和余弦方法的性能反比于 s 和 u 之间的距离,受 TREC 数据集所限,HQD 方法和相关度反馈方法所得集合 K 中真实相关文献的比重并不会太高,因此 s 并不会很靠近 u, 所以 HQD 第 3)的求和余弦并没有发挥多大作用,而只有前两步的 HQD 方法和相关度反馈方法很相似,所以在测评中两者表现出相近的性能。

鉴于此,通过采用其他自动选择规则,如关键词、聚类等[7-8],提高集合 K 中真实相关文献的比例,HQD 方法的性能还有进一步上升的可能。

5　结语

本文提出了一种结合数据融合及相关度反馈技术优点同时避免其某些弱点的混合型信息检索优化方法——HQD 方法。该方法有三个主要步骤,首先由初始结果集中自动生成若干替代查询,然后检索这些替代查询,最后采用求和余弦方法合并这些结果集。实验数据表明 HQD 方法能有效提高检索性能。理论分析表明 HQD 方法还可以采用更多的自动选择规则生成更好的集合 K, 其实验性能还能有进一步的提升。

参考文献:

[1] NG K, KANTOR P. Predicting the effectiveness of naive data fusion on the basis of system characteristics [J]. Journal of American Society for Information Science, 2000, 51(13): 1177 - 1189.

[2] LEE J. Combining the evidence of different relevance feedback methods for information retrieval [J]. Information Processing & Management, 1998, 34(6): 681 - 691.

[3] KANTOR P B. Two heads are better than one: The potential of data fusion concepts for improvement of online searching [C]// The 13th National Online Meeting. New Jersey: Learned Information, 1992: 147 - 151.

[4] MANMATHA R, FENG F, RATH T. Using models of score distributions in information retrieval[C]// Proceedings of the 2Jth ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2001: 267 - 275.

[5] BUCKLEY C, SALTON G, ALLAN J. The effect of adding relevance information in a relevance feedback environment[C]// Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: Springer-Verlag, 1994: 292 - 300.

[6] ROCCHIO J J. Relevance feedback in information retrieval[M]. Englewood Cliffs: Prentice-Hall, 1971.

[7] WHITE R W, RUTHVEN I, JOSE J M. A study of factors affecting the utility of implicit relevance feedback[C]// Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM Press, 2005: 35 - 42.

[8] JÄRVELIN K. An analysis of two approaches in information retrieval: From frameworks to study designs[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 971 - 986.

[9] WU S, McCLEAN S. Performance prediction of data fusion for information retrieval [J]. Information Processing and Management, 2006, 42(4): 899 - 915.