

网格环境下信息集成框架的设计及其应用^{*}

张 慧 黄刘生 张国义 李牧维

(中国科学技术大学计算机科学与技术系 合肥 230027)

摘 要 网格是把整个因特网整合成一台巨大的超级计算机,实现资源的全面共享。由于网格跨越不同管理域,因此其信息是分布和异构的,如何快速准确地从这些分布和异构的信息中检索出所需的信息,则需要一个有效的信息集成机制。本文提出了一种基于本体论(Ontology)的信息集成框架,详细研究和分析了这种框架下的信息集成的原理和框架的构成,并给出了其在国家 863 项目“淮河水资源保护与管理应用网格”中的应用。

关键词 网格,本体,信息集成

Design and Application of Information Integration Framework under Grid Environment

ZHANG Hui HUANG Liu-Sheng ZHANG Guo-Yi LI Mu-Chu

(Department of Computer Science, University of Science and Technology of China, Hefei 230027)

Abstract Grid integrates the whole Internet as a very huge super computer to enable every resource to be shared. Because Grid span different management organizations, its information is distributed and heterogeneous, how to retrieve required information accurately and quickly from these distributed and heterogeneous information, this requires an efficient information integration mechanism. This paper proposes an ontology based information integration framework, studies and analyzes the theory of information integration and the framework's architecture in detail. The paper also presents the application of this framework in the National 863 project "The Application Grid of Huaihe Water Resource Protection and Management".

Keywords Grid, Ontology, Information integration

1 引言

网格(grid)^[1]被誉为下一代的 Internet,是一个集成的计算与资源环境,或者说是一个计算资源池。它把整个因特网整合成一台巨大的超级计算机,实现计算资源、存储资源、数据资源、信息资源、知识资源、专家资源的全面共享。网格是借鉴电力网(Electric Power Grid)的概念提出来的,它的最终目的是希望用户能像使用电力一样方便地使用网格上的所有资源。我们在使用电力网的时候,只需要插入一个插头,而不必知道电是从哪一个电站发出来的,利用哪种发电方式产生的,也即,对用户是完全透明的。而在网格环境下,由于其跨越不同的管理域,其信息是分布和异构的。如果按照以前基于关键字的检索方式,则会得到很多无关的信息,用户需要继续查看每一个信息来源来了解自己是否需要,这就相当于我们用电的时候需要知道具体的电站;而且由于信息是异构的,那么用户在处理这些信息时,还必须了解信息的结构方式,这相当于我们用电时还需要知道电是用哪种发电方式产生的,以上这两点显然都不符合网格概念。因此我们需要一个有效的信息集成机制,这个信息集成机制必须具备以下几个特点:

(1)能够给用户提供一个统一的接口,用户能通过这个单一的入口访问所有信息。

(2)屏蔽网格中信息的分布性和异构性。

(3)具有一定的智能性。这需要两个方面的配合:底层的信
息需要用结构化方式来描述,要能够描述出信息的含义以及与其他信息之间的关系;上层则提供赋予推理机制的智能主体(Agent),这些主体能够对底层有良好定义的信息进行操作。

本文提出了一种基于本体论的信息集成框架,该框架基

于本体技术,并建立了一个智能的信息检索中间件,相对于当前一些基于关键字的检索的信息集成的机制,我们的框架具有更快速、更方便、更准确的特点。论文的内容组织如下:第 2 节介绍本体的概念,信息集成框架在第 3 节描述,第 4 节介绍信息集成框架在淮河应用网格的具体应用,接下来讨论一些相关工作,最后是结束语。

2 本体的概念

本体论原本是一个哲学上的概念,表示客观事物存在的本质。在各种文献中与本体相关的概念和术语的用法并不完全一致。文[2]列举了一些本体的几个有代表性的定义。综合这几个定义,我们可以了解到本体是通过对一个领域的概念、术语及概念之间相互关系的规范化描述,从而能够基本表达出一个领域的基本知识体系。本体最显著的特点是通过将现实世界的事物抽象为一组概念以及对概念间关系的描述,使这些概念具有语义性,在人和机器之间搭起了一个通讯的桥梁。例如,在水资源领域,可以有以下概念{“河道、水闸、排污口、污染物”}及这些概念间的关系{“水闸位于河道上”、“污染物通过排污口排放到河道”}。

构造本体的目的是为了知识的获取、共享以及重用。当前本体已经应用在不同的方面,包括知识表示和获取、规划、进程管理、数据库框架集成、自然语言处理和企业模拟等等。在网格环境下,由于一个领域内的数据信息庞大,因此需要建立一个本体将领域内的数据信息组织起来,构建一个领域的概念语义网,这样底层的数据信息就具有了良好的定义。从而上层的一些主体对这些有良好定义的数据信息可以进行良好的操作,实现一定的智能性。这样机器就能够代替人做一些工作,用户的负担将大大减轻。

^{*}国家“863”项目(项目编号:2002AA104560)资助;安徽省教育厅重点科研项目(项目编号:2003kj049zd)资助。张 慧 硕士,主要研究领域为网格下的信息集成及数据处理。黄刘生 教授,博士生导师,主要研究领域为高性能计算,信息安全及信息融合,信息集成。张国义 讲师,主要研究领域为高性能计算。李牧维 硕士,主要研究领域为数据库。

在文[3]中,Uschold &Gruninger 提出了一个本体构造的方法学,包括以下内容:

①确定本体的目的和范围。②构造本体,分为以下子内容:本体捕获;本体编码;集成已有本体。③对本体进行评估,检验。④文档记录。⑤每一阶段的指导准则。

3 网络环境下的信息集成框架

为了屏蔽网络环境下信息的异构性,需要一个全局模式,用户只需要和这个全局模式交互,而不必了解底层各个数据源的具体模式。根据上述对本体的分析,我们发现其很适合作为全局模式。另外,为了给用户一个统一的接口,使用户不必了解底层数据的具体放置位置而能快速地获取需要的数据,我们建立了一系列多主体的检索服务器。这些服务器分布在网络环境中的一些节点上,通过相互间的交互,形成了一个智能检索中间件。通过以上分析,我们给出了网络环境下的信息集成框架(如图1所示)。框架分为三层:

(1)底层为数据库及文档。数据库包括各种异构数据库,例如 MSSQL, SYBASE, DB2 等等不同厂家的数据库管理系统。文档包括各种结构化、半结构化以及无结构化的文档。不同的数据源对数据的描述方式是不同的,例如对于一个污染物氨氮的含量,文档中可能直接表示为:氨氮含量:mg/l,在数据库中可能将其表示为一个属性 NH₃-N:numeric(7,3),而每种数据库的属性定义又各不相同。

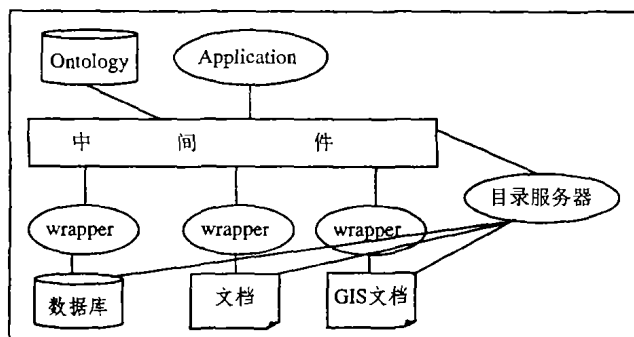


图1 信息集成框架

在本体中,污染物的含量可以描述为排污口的一个属性,有其固定的描述方式。这样,我们就可以把本体看作一个全局模式,而对每一类数据源根据它的数据描述方式利用 XML Schema^[4]建立一个本地模式,通过包装器(wrapper)在全局模式和本地模式之间进行映射。这个包装器可以对一类数据源进行转换,比如,用同一个包装器可以实现从本体模式到 SQL 数据库模式的转换。用户只要了解本体里数据的描述方式即可,而不必了解底层各个数据源各自的数据描述方式。另外,我们设立了一个目录服务器,每一个数据源与目录服务器中的一项对应。目录服务器中的每一项包括对应数据源的一些描述信息,例如这个数据源的内容信息(采用本体里的概念来描述其内容),物理存放地点,结构方式和查询方式等等,这样使得搜索主体能很快确定应该查询哪些数据源。

(2)中间层为一个智能检索中间件。它由一系列检索服务器组成,每个检索服务器包含了一系列的主体(Agent),主体是指在一定的环境下能够持续自主地运行的实体,是拥有自治性、社交能力、响应性和主动性的系统^[5]。在多个主体组成的多主体系统里,每个主体不仅可以动态响应外界环境的变化,解决自己局部需要解决的问题,而且能够和其他的主体进行相互之间的通信,获取自己需要的信息,解决自己局部无法解决的问题。

(3)最上层为应用层,主要由需要在网格中检索信息的用

户组成。

4 信息集成框架在淮河水资源保护与管理应用网格中的应用

淮河水资源保护与管理应用网格是在网格环境下,利用高性能计算环境来对淮河水资源进行防洪调度、防污调度、水资源管理。由于淮河流域面积有 27 万平方公里,流经四个省,大中小型水库有 5674 座,河网闸坝群 5427 座,因此其数据量非常大,而且流域内的四省水利厅水文局等管理部门有各自的数据库,文档以及 GIS 数据,各省以下的部分市级水利局等也拥有自己的数据源,因此数据是分布和异构的,特别需要用信息集成框架将其集成。

(1)首先建立水资源领域的本体。图2给出了本体中的一个例子。在图2中,椭圆表示概念,它们之间的边表示这些概念之间的关系。方框表示概念的属性(属性没有完全表示出来)。在这个例子中,包含的概念有:省份、城市、城镇污染源、河流、水系、入河排污口、水闸断面。这些概念之间的关系是:城市是省份的子类,城镇污染源位于城市,污染源通过入河排污口向河流排放污染物,河流是水系的子类,水闸断面位于河流上。

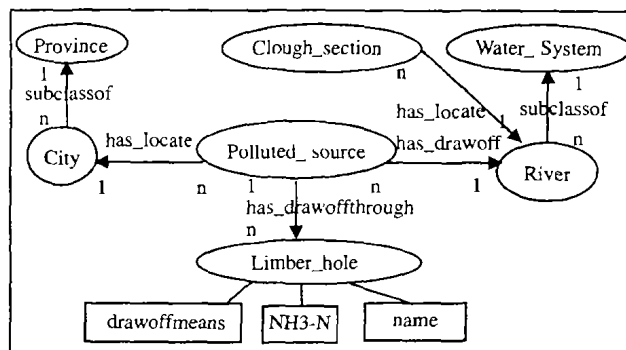


图2 水资源的本体的例子

(2)其次,建立检索服务器及构建多个主体。图3给出了检索服务器里主体的详细设计。一个检索服务器包含四个主体:用户交互主体,分析主体,本地搜索主体,远程搜索主体。其各部分的功能分别是:

A. 用户交互主体:这个主体的主要作用是和用户进行交互。它接受用户提交的检索请求,将用户提交的请求规范化后移交给分析主体;当查询结束以后,用户交互主体负责将查询结果提交给用户。

B. 分析主体:分析主体主要通过和 ontology 以及目录服务器的交互,来对用户的查询进行分解和定位。分析主体对用户交互主体提交过来的规范过的查询请求的每个关键字抽象到本体中的一个概念,并利用本体中这些概念的关系来进行查询分解。对于每一个子查询,通过在目录服务器里查询相关的信息,分析主体可以获知查询的具体的位置及对应的查询方法。另外,分析主体还必须对本地搜索主体以及远程搜索主体返回的结果结合用户的需求和本体进行整合。

C. 本地搜索主体:由于网格跨越不同的管理域,因此我们在一定范围内设定一个检索服务器。每个检索服务器管理属于它自己域内的数据信息。本地搜索主体就负责搜索属于自己检索服务器管理域的数据信息。

D. 远程搜索主体:由于用户检索的数据并不总在他提交查询的检索服务器管理的域内,因此需要一个主体能与远程的服务器交互。远程搜索主体就是用来完成这个工作的。它负责和远程的检索服务器里的本地检索主体交互,通过远程的检索服务器里的本体检索主体搜索位于远程管理域内的数据

信息。

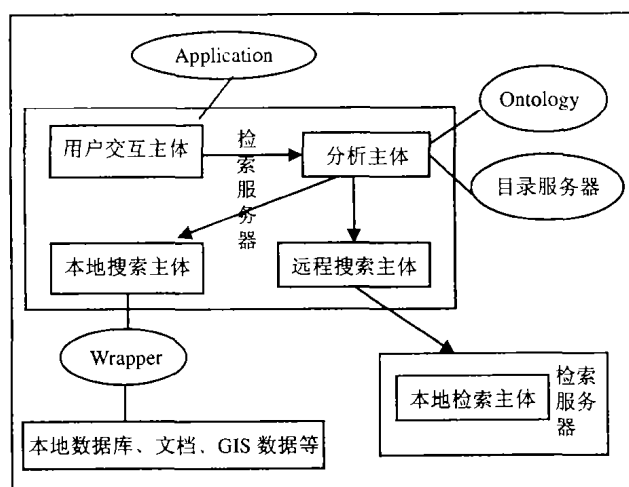


图 3 检索服务器及其包含主体

(3)智能检索过程。当用户提出查询时,首先将用户的关键字跟本体里的概念进行比较,将用户的关键字映射到本体里的概念或者属性,同时截取下这些概念以及与这些概念相关的一个语义网络图(一些节点和相关的关系边),并根据用户查询的目的并结合这些关系边进行查询分解。

例如:假设用户希望“查找安徽省,河南省的污水排放方式是连续排放的排污口的名称”,在用户的查询中,我们抽取“省”,“入河排污口”,“连续排放”,“排放方式”,“名称”五个关键字,其中“省”映射到“Province”概念,“安徽省”,“河南省”是“Province”的两个实例,“入河排污口”映射到“Limber hole”概念,“污水排放方式”则对应着“limber hole”的一个属性——“drawoffmeans”,“连续排放”则是这个属性的值。在这里,我们将“province”作为始节点,“limber hole”作为终节点,则通过图匹配,可以得到从整个本体中截取下来的语义网络图(图4)。

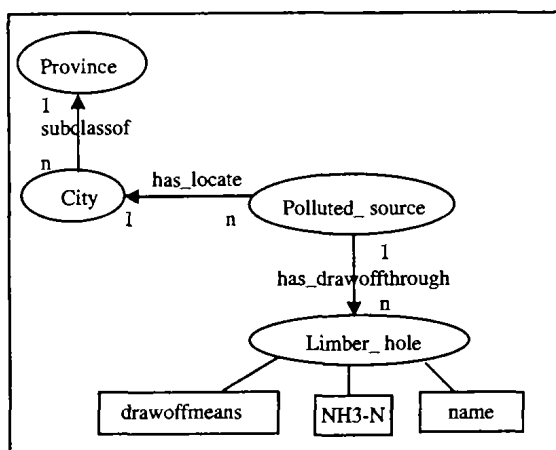


图 4 与查询相关的本体截取图

首先,河南省和安徽省都属于“province”这个类,它们是并列的,因此在这里可以将查询分为两个子查询:“查找安徽省的污水排放方式是连续排放的排污口的名称”和“查找河南省的污水排放方式是连续排放的排污口的名称”。对于每个子查询,从这个语义网络图看,province 和 limber hole 并不是直接相关的,但是我们注意到,每条边连接的两个概念都是直接相关的,因此,我们可以根据每条边来进一步进行查询分解,每条边做一个子查询。在此例中,对于安徽省的子查询,我们可以分解为三个子查询:“安徽省的各个城市”,“每个城市的污染源”,“每个污染源采用连续污水排放方式的入河排污口的名称”。对每个子查询,通过对目录服务器的访问,可以

确定具体可以在哪个数据库或者哪个文档中查找,然后通过 wrapper 将本体里描述的模式转化为与数据源对应的本地模式。对于子查询返回的结果,首先,wrapper 需要将各个数据源的 XMLschema 形式的结果转化为本体中采用的全局的 schema,这相当于两个 xml 文件的转换,我们采用 XSLT^[5]来实现。然后,分析主体必须结合用户的需求和本体来进行整合。在本例中,由于安徽省和河南省是一个类的两个实例,因此对两个省的情况可以给出并列的两个表,而由于 city 是 province 的子类,则省和市之间则需采用层次结构,city 和 polluted source, Polluted source 和 limber hole 都是 1 对多的关系,因此最后给出的形式可以如下:

```

province: 安徽
city: 蚌埠
polluted source: 造纸厂
limber hole
attribute name: 蚌埠一号码头
attribute name: 龙子河排污口
.....
polluted source: 怀远新啤酒厂
limber hole
attribute name: 蚌埠一号码头
attribute name: .....
city: 阜阳
polluted source:
:
```

5 相关工作

以前的信息发现和集成都是基于关键字检索,例如 google^[7],sohu^[8]等搜索引擎。搜索引擎一般都提供分类目录和关键字检索相结合的方式。虽然搜索引擎品种繁多,功能不一,但其总体结构和基本原理是一样的。大至可分为以下三个步骤^[9]:(1)派出网页搜索工具 Spider(蜘蛛)或 Robot(机器人)到互联网上搜索信息,并将它们带回到搜索引擎;(2)对文档信息进行分类索引,建立索引数据库;(3)通过 Web 服务器端软件,为用户提供浏览器界面下的信息查询。但是由于各种资源的描述方式、内部结构方式并不完全相同,并没有一种有效的机制对它们进行描述,因此通过搜索引擎搜索到的内容繁多,结构不一,用户很难快速获得自己需要的信息。在我们的框架中,通过建立本体,描述概念及概念之间的关系,形成了一个语义网,这样,给用户和机器搭起了沟通的桥梁,并且将本体作为一个全局的 schema。在底层则利用了 xml 内容和形式分离的特性,采用 xml 来描述信息,通过建立 wrapper 在本地 XMLSchema 和全局 schema 之间进行转换,屏蔽了底层信息的异构性,而各个检索服务器的合作屏蔽了资源的分布性。

结束语 本文针对网格环境数据信息的分布性和异构性,提出了一个网格环境下的信息集成框架。它能够通过网络将一些信息孤岛联系起来,并使用户能够准确、快速地获得自己需要的信息,而不是象一般的搜索引擎那样得到很多无关的信息,需要用户进一步去查找。

参考文献

- 1 Foster I, Kesselman K. The Grid: Blueprint for a Future Computing Infrastructure. San Francisco, USA: Morgan Kaufmann Publishers, 1999
- 2 王昕·综述:本体的概念、方法和应用·<http://www.prdm.net/papers/knowledge/Ontology-200overview.htm>
- 3 Uschold M, Gruninger M. Ontologies: Principles, methods and applications. The Knowledge Engineering Review, 1996, 11 (2)
- 4 W3C XML Schema. <http://www.w3.org/XML/Schema>
- 5 Wooldridge M, Jennings N. Intelligent Agents: Theory and Practice, The Knowledge & Jennings, N. Intelligent Agents: Theory and Practice, The Knowledge Engineering Review, 10 (2)
- 6 XSL Transformations (XSLT). <http://www.w3.org/TR/xslt>
- 7 google. <http://www.google.com>
- 8 sohu. <http://www.sohu.com>
- 9 姚国祥, 罗伟其, 沈镇林. 网上信息搜索技术与搜索引擎[]]. 计算机科学, 2000, 27(7): 35~38