◎ 陈华钧 姜晓红 吴朝晖 著

DartGrid:

支持中医药信息化的语义网格平台实现

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

DartGrid:支持中医药信息化的语义网格平台实现

陈华钧　姜晓红　吴朝晖 著

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

DartGrid：支持中医药信息化的语义网格平台实现／陈华钧，姜晓红，吴朝晖著. —杭州：浙江大学出版社，2011.5

ISBN 978-7-308-08569-4

Ⅰ. ①D… Ⅱ. ①陈… ②姜… ③吴… Ⅲ. ①中国医药学—管理信息系统—研究 Ⅳ. ①R2-39

中国版本图书馆 CIP 数据核字（2011）第 077014 号

DartGrid：支持中医药信息化的语义网格平台实现

陈华钧　姜晓红　吴朝晖　著

责任编辑　许佳颖　黄娟琴
文字编辑　陈静毅
封面设计　刘依群
出版发行　浙江大学出版社
（杭州市天目山路 148 号　邮政编码 310007）
（网址：http://www.zjupress.com）
排　　版　杭州中大图文设计有限公司
印　　刷　杭州丰源印刷有限公司
开　　本　787mm×1092mm　1/16
印　　张　9
字　　数　187 千
版 印 次　2011 年 5 月第 1 版　2011 年 5 月第 1 次印刷
书　　号　ISBN 978-7-308-08569-4
定　　价　26.00 元

前 言

近年来，以 Web 2.0 为代表的网络新应用层出不穷，并推动着计算机各方面技术的进一步发展。这些应用的一个典型特征是数据驱动。例如博客、播客以及各种社会计算网站，如 Facebook、YouTube、Flickr、Twitter、Google Social Graph、Microsoft Connection 等，由用户产生的各种类型的 Web 数据已经远远超过了传统的 Web 网站。各种应用所产生的数据都已经达到或将要达到 Petabytes 甚至更高的规模。这些网络数据还具有跨域和异构的特征。数据驱动型网络应用的核心需求是如何有效地整合、聚合，乃至融合由多个管理域产生的多种异质异构数据，并支持大规模的并发访问。

语义网格，将以语义 Web 为代表的语义技术和以网格计算为代表的体系架构技术结合起来，通过规范化描述明确表达包括计算、存储、数据库、服务等各种信息资源的内涵语义，提供开放、安全、有序、可扩展的管理体系架构来解决和实现复杂网络环境下跨多个机构的大规模分布式协同计算和数据共享问题。语义网格为解决上述互联网新的挑战提供了新的思路和技术方法。

DartGrid 是由浙江大学计算机学院自主研发的语义网格平台软件系统。DartGrid 主要面向中医药信息化等数据驱动型网络应用领域的一些新需求，并结合语义 Web 技术和网格技术等新兴互联网技术研制。其主要技术包括数据的语义集成技术、语义搜索技术、流程服务的语义组合技术、语义网格中的分布式数据挖掘与知识发现等，旨在为语义网格提供一套综合信息管理平台。

DartGrid 平台的主要开发背景是中医药信息化。DartGrid 可以有效地支持数据与知识密集型领域中的知识表示、管理与问题求解，而中医药领域是一个典型的数据与知识密集型领域。本书结合对中医药领域的应

用需求分析，提出基于语义网格构建的、面向中医药领域的e-Science环境，并就基于语义网格的数据集成与共享、海量中医药数据挖掘与分析等，详细阐述了其体系结构、技术特征以及应用成果。

本书的组织结构如下：第1章介绍了语义网格的基本概念以及DartGrid功能与技术概述；第2章对DartGrid所涉及的关键技术进行了系统性的介绍；第3章结合中医药应用实例介绍了DartQuery语义查询系统；第4章针对中医药信息搜索的需求，介绍了DartSearch语义搜索系统；第5章介绍了DartMapping语义映射系统；第6章针对中医药海量数据的特征介绍了DartSpora海量数据挖掘系统；第7章介绍了一个基于语义的数据Mashup系统；第8章结合中医药的信息服务化需求，介绍了一个服务管理系统。本书既可以作为互联网技术领域的研究者和实践者的参考读物，也可以为从事中医药信息化工作的科研人员提供参考。

本书是浙江大学CCNT实验室的教师和学生多年的努力成果，以下老师和同学为本书的撰写或与本书相关的项目研发作出过贡献，在此一并表示感谢，他们是：邓水光、于彤、周春英、毛郁欣、封毅、张宇、王俊健、张小刚、付志红、秘中凯、密金华、刘森、盛浩、陶金火、杨克特、卢宾、郑耀文、梁欣颖、刘明魁、顾佩钦、张露、张湘豫等。本书还是浙江大学与中国中医科学院信息研究所多年合作的成果之一，在此向中国中医科学院的崔蒙研究员、尹爱宁、刘静、李园白、雷蕾等同仁给予我们的长期支持表示感谢！

与本书相关的研究内容受到如下项目的资助：科技部“973”语义网格专项(No. 2003CB317006)、国家杰出青年基金(No. NSFC60533040)、科技部现代服务业支撑计划(2006BAH02401)、科技部“863”计划(2006AA01A122，2009AA011903，2008AA01Z141)、国家自然科学基金项目(No. NSFC61070156，NSFC60873224)、浙江省科技重大项目(2008C03007)。在此一并表示感谢！

作　者

2011年1月

目 录

第 1 章

概　述

1.1　研究背景

1.1.1　背景概述

1969 年互联网在美国诞生，经过四十多年的快速发展，现在已成为人们生活、工作交流不可或缺的一部分。人们广泛地应用互联网查找信息、学习、娱乐、工作、购物、交友等。截至 2008 年年底，仅中国一个国家的网民数量就已超过 2 亿。随着互联网的不断发展和进一步的普及应用，在原有的搜索引擎、网络新闻、电子商务、在线交易、电子邮件等互联网应用继续保持快速发展的基础之上，不断出现了很多新兴的互联网应用模式，例如：P2P 数据共享下载，典型的如 BT 下载；软件即服务（SaaS），典型的如 Google Doc、Salesforce 等；社会网络应用（Social Networking Applications）及社会媒体应用（Social Media），典型的如 Facebook、LinkedIn 等。

这些新兴互联网应用模式的出现，不仅极大地改变和扩展了互联网的服务模式和应用范围，也大大推动了互联网技术本身的飞速进步。典型的新兴互联网技术包括：网格计算、语义 Web、服务计算、跨界混搭（Mashup）技术、Ajax 技术、富客户端技术等。这些技术相互交融，分别从个性化的人机交互、社会化的资源共享、广域范围的协同协作、跨领域的信息融合与系统集成等多个方面推动着整个互联网的快速进步和飞速发展。

本书介绍的 DartGrid 语义网格平台正是在这样的技术发展背景下被提出、研究和开发的。特别的，DartGrid 着重从网格计算和语义 Web 两个方面，从信息语义描述框架、语义搜索、语义映射与信息融合、网络数据挖掘与分析、网络数据混搭与拼接、网络服务发现与交互等多个方面，提出一系列新的技术思路和相关实现，并对这些相关技术的潜在应用进行了深入探讨。

1.1.2 网格计算的基本思想

网格计算是针对"大科学"(Big Science)在跨多个机构的协作共享方面的需求提出的。典型的大科学应用包括生命科学、地球科学、气象与天文、高能物理等。这些应用的典型特征可以归纳为:海量计算、高性能计算和多机构协作。网格计算的基本思想是利用互联网技术实现分散在不同地理位置的计算机和各种设备的动态协同管理,并把它们组织成一台"虚拟化的超级计算机",向用户提供透明的计算力服务。因此,早期的网格计算多指提供计算力服务的"计算网格",而这样的计算网格主要针对的是支持科学研究的 e-Science 应用。典型的如,支持欧洲高能物理研究的离子对撞机正是采用网格技术实现其底层的海量数据存储、传输和分析计算的功能。

随着网格计算的成功,网格的研究范畴也迅速扩展到对各种信息资源包括计算、存储、数据、软件、设备等的大规模协同共享。与此同时,网格的思想也逐渐渗透到企业应用领域,应用范畴也扩大到许多企业级应用。例如,Oracle 推出了数据库网格计划 Oracle 10g,旨在推动网格技术在企业界的应用。IBM、Sun、HP、Platform 等都推出了面向企业级应用的网格技术平台和服务。标志性的事件包括企业网格联盟(EGA)的成立,以及后期企业网格联盟(EGA)与全球网格论坛(GGF)的融合成立新的开放式网格论坛(OGF)等。

在技术层面,早期的网格技术强调对高性能计算的支持。随着网格应用范畴的不断扩展,网格技术的重点也转移到面向服务的网格体系架构技术的研究上。随着网格应用重点从典型的大科学应用朝着大企业应用的转变,网格的一些规范和标准也逐渐与实现企业信息集成的 Web Service 技术融合。标志性事件包括 WSRF(Web Service Resource Framework)的提出,从而使得网格计算在技术体系上统一到了面向服务计算的 SOA(Service-Oriented Architecture)架构下面。

总结起来说,网格计算所针对的核心问题是:实现跨多个机构的虚拟组织之间的动态资源共享与大规模协同。这是针对未来互联网应用的一个典型特征提出的,即未来互联网应用覆盖范围飞速增长,跨领域、跨机构之间的应用系统日趋融合。传统针对单一、封闭式的网络应用解决方案无法满足跨多域、大覆盖范围应用的需求,例如大规模数据共享、广域网范围内的协同协作等。在网格计算系统体系下,用户可以通过互联网享受一体化的、动态变化的、可灵活控制的、协作式信息服务。

从网格计算的角度而言,现有的互联网还缺乏一个面向共享与协同应用开发的统一分布式基础设施。网格技术从体系架构角度定义统一的标准与规范,提供了具有自主容错、动态可扩展特点的分布式信息系统体系架构。网格系统中的各个组成节点在部分节点运行失效、部分节点退出、新节点接入、需求变化等情况下,都可依据一定的策略或协议,动态地重新进行功能组合,动态生成新的虚拟组织,在总体上体现出自主容错、自我管理(Self-management)和自我配置(Self-configuration)等特征。

网格基础设施，如 Globus，依据网格体系架构规范，通过把一些分布式系统的共有特征实现为通用的、可复用的模块，使得新的资源或服务能够以“即插即用”的方式接入网格。这对于开放性的应用非常重要，也是系统可扩展性的必要保证。此外，网格对于用户来说，具备“按需服务”的特征。在网格中，更加容易对有序的信息资源进行筛选和过滤，更易于为用户寻找最能匹配用户需求的资源和服务，从而实现按需的服务。

1.1.3 语义 Web 的基本思想

1998 年，Web 发明人 Lee TB 先生指出：“Web 文档本身描述的是现实世界中的对象、概念和它们之间的关系，但这些信息都是用自由文本描述的，这虽然方便人浏览，但机器却无法理解 Web 文档中所蕴涵的语义。”这是导致当前搜索引擎无法对信息进行精确搜索的根本原因。基于这一考虑，他提出了对未来 Web 的一种设想，即：未来的 Web 会是一种“有意义的 Web”(Meaningful Web)。这样的 Web 中的信息不仅仅是给人来浏览的，而且可以被机器理解(Machine-understandable)。这种更加“聪明”的 Web 会极大地改变现有互联网搜索引擎的窘境。基于这些想法，他连同一些国际知名学者首次提出了语义 Web 的概念。并在他的直接推动下，国际万维网联盟组织 W3C① 制订了一系列的语义 Web 信息标准，这包括 Web 信息语义表达框架 RDF②，本体描述语言 OWL③ 等。

概括起来说，语义 Web(Semantic Web)是由 W3C 所倡导的一种新的 Web 技术。根据 W3C 的定义，语义 Web 与传统 Web 的主要区别在于要建立一个数据 Web (Web of Data)。有别于传统的 Web，语义 Web 主要是由超文本链接连接起来的文档 Web(Web of Document)，是更为细粒度的 Web。在语义 Web 中，Web 节点不再仅仅是一个 Web 文档，而可以是任何用户想描述的事物，比如可以代表一个人、一本书、一所学校等；而 Web 链接也不再是一种用于链接文档的简单链接，而是用于链接任何对象，并明确标明事物对象之间的关系(比如同学关系、雇员关系等)的语义链接。语义 Web 的最终目的是要通过建立各种数据对象之间的关联关系，使得用户更加容易浏览、查询、搜索、交换、管理他们的信息，解决传统单一依靠自然语言处理和统计分析技术的搜索引擎技术所面临的一系列瓶颈和难题，如图 1-1 所示。

① W3C 国际万维网联盟组织：http://www.w3c.org

② RDF 语义表达框架：http://www.w3.org/RDF/

③ OWL 本体描述语言：http://www.w3.org/TR/owl-features/

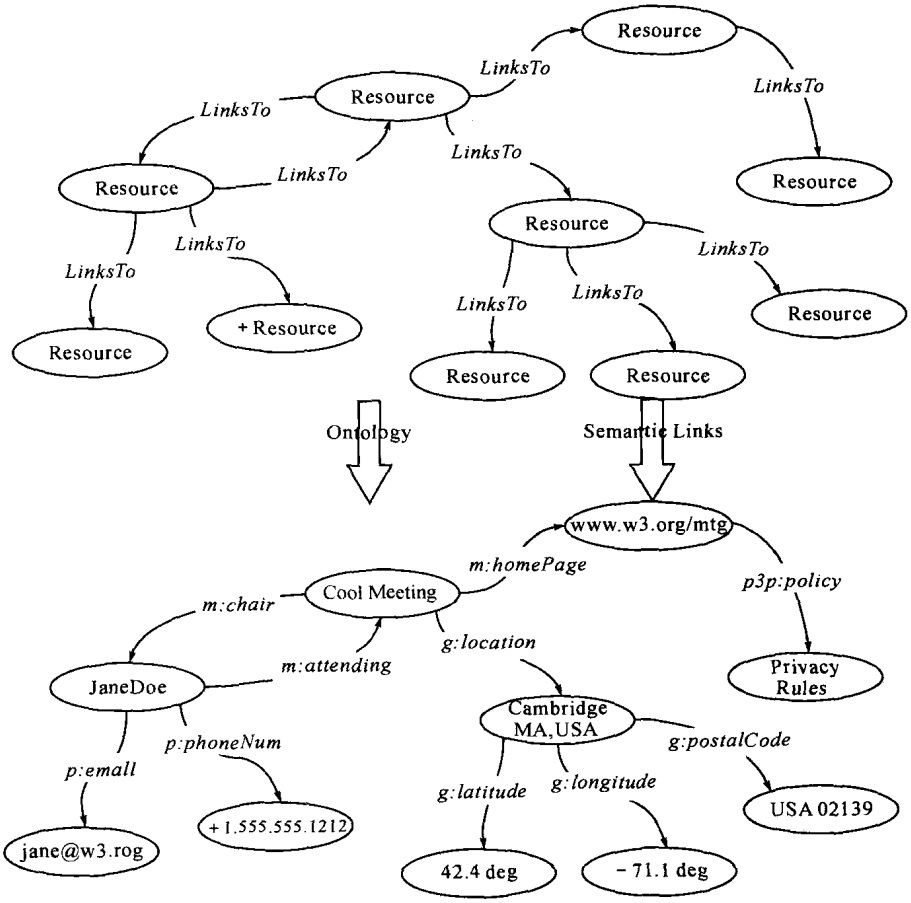

图 1-1　利用语义超链接描述对象的语义 Web

有关语义 Web[①] 的标准化工作由著名的 W3C 组织进行。2001 年，W3C 完成了 RDF(Resource Description Framework)的标准化工作。RDF 是有关 Web 语义表达的最基础的规范，它规定了如何对 Web 资源的语义进行规范化、明确化描述的基本方法和框架。RDF 吸取了关系数据库模型的经验，借鉴了人工智能领域的知识表达研究成果，并针对互联网的本质特征而设计，是专门面向信息集成的数据模型。相较 XML，其语义描述更加简单直观，语义约束更加明确和规范。从某种角度讲，XML 是资源描述的语法规范，而 RDF 是在 XML 之上，用于资源描述的语义规范。

但由于 RDF 在本体(Ontology)和术语规范(Terminology)定义上表达能力的不足，W3C 又启动了 OWL(Ontology Web Language)[②]的规范化工作。OWL 是具有更强表达能力的资源语义表达规范，它充分借鉴了知识表达领域有关语义网络、框架

① W3C 语义 Web 工作组：http://www.w3.org/2001/sw/

② W3C Web Ontology 语言工作组：http://www.w3.org/2004/OWL/

系统和描述逻辑的研究，能用于定义复杂的领域本体和术语规范，比如医学术语等。2004年，W3C完成了有关OWL的规范化工作。但是要实现语义Web的目标，仍然还有大量的规范化和标准化工作需要做。比如，W3C于2005年又启动了语义查询语言SPARQL[①]的标准化工作。此外W3C还成立很多兴趣小组支持对一些相关应用领域的研发工作，比如针对生命科学领域应用的兴趣小组旨在推动语义技术在生命科学领域的应用[②]。

借助于这些语义表达标准，语义Web可以通过语义把各种数据(Data)和程序(Program)互联起来，综合利用知识的方法解决信息资源的语义问题，从而解决资源的共享问题，使Web成为一个能提供知识服务的巨大知识库。

1.2 语义网格的基本概念

1.2.1 语义网格的提出

语义网格的最早提出可以追溯到2001年。英国的e-Science计划提出为全球化的协同科学研究提供分布式计算技术，支持科学家更高效地发布、共享、分析实验数据和协同科学研究。在当时，网格技术被认为可以为e-Science提供这样一个分布式共享网络环境。为此，欧盟的一些国家投入了大量经费到e-Science和网格相关的项目中。早期的e-Science项目主要侧重于中间件的研发，但在这些e-Science项目的研发中，人们发现网格的发展水平和e-Science的理想之间存在差距，要达到e-Science的易用性和无缝自动化要求，必须实现尽量多的机器自动处理特性和尽量少的人机交互介入。

De Rouge D等研究者发现e-Science的这一需求与语义Web的部分目标非常相似，于是考虑把语义Web的一些技术引入e-Science领域，并在2001年最先提出了语义网格的概念。2002年，一些研究者在全球网格论坛(GGF)成立了语义网格研究组，吸引了更多的研究者对语义网格的关键技术进行研究。该工作组对语义网格的概念进行了初步的界定，认为语义网格的目标是要通过语义实现系统高度的易用性和无缝自动化来支持全球化的协同工作和共享。

在语义网格的研究上，欧盟的资助力量是最大的。从2001年开始，欧盟陆续在其FP5、FP6框架中资助了一大批的语义网格相关项目。典型的比如：OntoGrid强调面向知识的网格服务和系统的构建；InteliGrid强调如何建设基于语义的虚拟组织；SIMDAT强调语义网格中的数据管理；Akogrimo强调普适计算与语义网格的结合；

① W3C Data Access 工作组：http://www.w3.org/2001/sw/DataAccess/

② W3C 语义 Web 与生命科学兴趣小组：http://www.w3.org/2001/swhcls

DataMiningGrid 强调语义网格环境中的数据挖掘;K-WF Grid 强调语义网格中的流程组合等。

中国的学者也较早开展了语义网格的研究和开发,并逐步形成了自身的特色。例如,2003 年,国家启动了"973"专项"语义网格的基础理论、模型和方法研究"。该项目从资源模型、流程组合、数据管理与存储、语义验证等多个方面对语义网格的研究进行了资助。

1.2.2 语义网格的定义

虽然当前从事语义网格相关研究的工作比较多,但是到目前为止还没有形成对语义网格概念的明确而统一的界定和定义。综合起来讲,语义网格是一种"基于语义的分布式计算技术"(Semantic-based Distributed Infrastructure for Internet),它建立在语义 Web 相关技术规范和网格体系架构相关技术标准基础之上,其目的是要支持构建基于语义的分布式网格系统。

具体来说,网格计算从开放系统的体系架构标准与规范的角度,研究如何实现一个足够灵活、支持共享的分布式计算基础设施。语义 Web 从规范化的资源语义表达角度,研究如何提供一个一致化的资源语义表达框架,以解决资源语义的异质异构问题。前者来源于计算机系统结构研究领域,后者来源于知识表达与信息表示领域。事实上,这两个领域对于解决互联网资源共享问题都同样重要,是同一个问题的两个方面。

语义网格,将以语义 Web 为代表的语义技术和以网格计算为代表的体系架构技术结合起来,通过规范化描述明确表达包括计算、存储、数据库、服务等各种信息资源的内涵语义,提供开放、安全、有序、可扩展的管理体系架构来解决和实现复杂网络环境下跨多个机构的大规模分布式协同计算和信息共享问题。

1.3 语义网格的研究范畴

语义网格所针对的核心问题可以概括为以下三个方面:互联网资源的语义异构、海量增长和泛在分布。一方面语义网格需要解决资源语义描述的不一致问题,另一方面也需要解决海量资源在组织形式和结构上的不一致问题。然而语义网格的核心技术框架仍然建立在传统的计算机和网络技术基础之上,因此我们把语义网格的核心研究范畴和基础技术框架总结为以下几个方面。

1.3.1 信息语义的描述框架与表达模型

语义网格需要研究适合开放网络环境应用需求的信息语义描述框架与表达模型。这种模型一方面为信息的语义表达提供严格的逻辑基础，另一方面又必须满足开放动态的松耦合知识集成的需要。在语义网格知识表示的层次结构中，目前较为成熟的部分（至底向上）有 XML（可扩展标记语言）、RDF/RDFS（资源描述框架及其模式）和 OWL（本体论语言），而在 OWL 之上的逻辑层、证明层（规则语言及其推理）的相关标准的制订工作仍处于需求征询阶段，信任层则还处于研究阶段。

1.3.2 异构网格资源的语义映射

网格旨在实现跨多个管理机构、多个领域的信息集成与管理。因此，网格计算所面临的一个核心问题即是异构网格资源的集成问题。这一问题对于数据资源尤为突出。语义网格与传统网格计算技术的典型不同之处即是采用语义 Web 技术解决这一问题。其中，如何基于语义建立各种异构网格资源之间的语义映射是一个突出的问题和难题。

1.3.3 Web 级的语义数据集成

传统的分布式数据集成技术只适合于局部范围内或单个企业范围内的信息集成，然而在互联网范围内的信息资源通常具有自治、异构、动态变化、开放扩展等特征。如何在这样一个开放式环境下实现 Web 级的语义数据集成是一个极大的挑战。语义 Web 技术为信息描述提供了一个标准化的语义描述和表达规范。语义 Web 语言 RDF/OWL 充分借鉴了传统数据模型，包括关系模型、XML 和各种知识表达系统，针对互联网的开放性、松耦合等特征设计，从而为异构信息的集成提供了非常理想的数据模型。另一方面，网格技术为分布式数据资源的管理提供了更加易于扩展的体系架构，使得在互联网范围内的资源组织和管理能有一个共同的规范和标准。

1.3.4 网格资源的语义搜索

语义网格中的资源都被标注为富含语义的资源之后，一个所需要解决的核心问题是如何基于语义来对这些资源进行搜索。这涉及对传统各种搜索技术的改造，最为典型的如排序技术。基于语义排序首先要考虑的是对象级的排序（而不是网页级的排序），其次还需要考虑如何利用语义信息的辅助进行排序。其他也发生变化的技术包括语义索引技术、语义分词技术等。

1.3.5 基于语义网格的数据挖掘与知识发现

语义在数据挖掘中所起的作用体现在以下两个方面。一方面,把分布式的、异质异构的数据资源通过语义集成为上层的数据挖掘提供大量、丰富的数据源,这就要求数据挖掘系统有效地利用这些数据富含语义的特性,并利用 Web 的访问方式提供数据挖掘服务。另一方面,语义本身的一个作用是建立资源之间的关联关系,通过语义建立的资源关系网本质上是以语义网络的形式存在的,这使得挖掘技术可以充分利用语义信息对语义网络进行复杂网络分析。

1.3.6 基于语义网格的服务组合与发现

语义网格的另外一项核心技术是异构系统之间的语义互操作和复杂流程的语义组合。跨多个机构的系统集成需要解决异构系统之间的互操作问题。尽管 WSDL/SOAP 等 Web 服务协议为互操作提供了实现基础,但是仍然忽略了异构系统交互过程中所需要解决的接口语义的不匹配问题。基于语义本体的方法可以通过将异构系统的接口定义映射到一个共享的语义本体之上,在语义层建立异构系统之间的逻辑连通性,从而实现异构系统之间的语义互操作。

1.3.7 基于语义网格的复杂工作流管理

分布式的工作流程日趋复杂化和易于发生改变,比如企业供应链的范围正在变得更加广泛和复杂。在语义网格中,采用语义本体定义抽象的流程,具体实现流程功能的实体(如 Web 服务或应用程序),以松耦合的方式映射到抽象的流程定义上。这样可以实现柔性的流程组合,在需求发生变化的时候,可以动态地对流程进行调整以适应变化,比如重新调整抽象的流程或映射新的功能实体。针对复杂流程的组合管理,语义网格所要提供的功能包括:复杂流程的语义建模、服务接口的语义匹配和映射、服务的能力描述、服务的语义组合、服务的自动发现等。

1.4 DartGrid 语义网格系统

1.4.1 DartGrid 简介

DartGrid 是由浙江大学计算机学院自主研发的语义网格软件系统。DartGrid 的研究和开发主要是围绕中医药信息化和智能交通系统等应用领域的一些新需求,并