Web服务合成技术在网格数据挖掘中的应用研究

陈增科，肖基毅，邵明前
（南华大学 计算机科学与技术学院，湖南 衡阳 421001）

摘 要：在网格环境下，存在大量的数据挖掘服务，但传统数据挖掘系统难以满足用户实际应用的需求。提出把Web服务合成技术运用于网格数据挖掘中，对已有的服务进行合成，形成新的、方便用户使用的数据挖掘服务，探讨把传统的数据挖掘系统与OGSA和Web合成技术结合，构建一个开放数据挖掘系统，满足不同领域、不同层次的知识发现。

关键词：Web服务；数据挖掘；合成；网格
中图分类号：TP393 **文献标识码**：A **文章编号**：1673－629X(2009)01－0234－03

Research on Web Service Composition Techniques Using in Grid Data Mining

CHEN Zeng-ke, XIAO Ji-yi, SHAO Ming-qian
(School of Computer Science and Technology, University of South China, Hengyang 421001, China)

Abstract: Under gird environment, there are a lot of data mining service. Traditional data mining service just provides limited functionality. Presents Web service composition use to data mining system based on grid, compose existing data mining service to obtain new functionality service, and use them directly. Discuss that is the combination of OGSA and Web service composition and architecture of traditional data mining system. An open data mining system is designed and can meet the user requirement of knowledge discovery in different domains and different hierarchies.

Key words: Web service; data mining; composition; grid

0 引 言

随着计算机、互联网等技术的发展，各种数据信息呈爆炸式增长。网格[1,2]环境下存在大量的分布的、自治的、异构的、动态的数据资源，在这种分布的环境下要分析和挖掘这些异构、动态的数据资源，以获取新的科学知识、规律、模式和决策技术支持信息，运用传统的数据挖掘[3]的模式、方法和技术是难以实现的。目前的数据挖掘系统是封闭的，系统提供的挖掘功能是固定的，无法满足特定用户的实际需求，难以动态有效地管理和维护多个挖掘算法；系统挖掘结果格式是封闭的，提供的挖掘结果往往无法为其他应用系统使用；不同的挖掘算法开发语言不同，对外交换的数据格式也有着很大的差异等等，这些都是传统数据挖掘自身的重大缺陷。

收稿日期：2008－04－28
基金项目：湖南省教育科研项目(06C724)
作者简介：陈增科(1972－)，男，湖南衡阳人，讲师，硕士研究生，研究方向为网格、Web服务和数据挖掘；肖基毅，副教授，硕士生导师，研究方向为网格信息资源共享与数据挖掘。

Web服务技术[4]是一种基于网络的，组件式的软件集成技术，Web服务合成技术[5]作为Web服务一项重要的增值功能，为服务的重用与自动化集成提供应用的基础。把挖掘系统构筑在Web服务技术之上，运用Web服务合成技术构建开放式数据挖掘系统能从网格环境中动态更新算法、挖掘结果表示方式、数据预处理工具等；Web服务合成技术能合成各种语言编写的算法，完成特定的挖掘功能，满足不同用户的实际需要。这些都无疑能够大大提高挖掘系统灵活性，增强挖掘性能。

1 Web服务

Web服务是一种部署在Web上的对象，是一个基于标准的、广泛部署的分布、开放的计算模式。可以对Web服务加以定义：

定义1：(Web服务) 一个Web服务就是一个二元组：$W = \{d, o \mid d \supseteq (\text{data}_1, \text{data}_2, \cdots, \text{data}_N), o \supseteq (\text{oper}_1, \text{oper}_2, \cdots, \text{oper}_N)\}$，其中 $d \supseteq (\text{data}_1, \text{data}_2, \cdots, \text{data}_N)$ 表示数据，这些数据包括简单类型和复杂类型

的数据,$o \supseteq (oper_1, oper_2, \cdots, oper_N)$表示操作,包括消息的交互和操作。

定义 2:(Web 服务的发现)是指通过一定方式找到满足特定要求 Web 服务的过程。描述为:$W = F(W_i), (i = 1,2,\cdots,N)$,其中 F 表示某一特定的关系运算,W_i 表示 Web 服务集,W 表示特定的 Web 服务。

定义 3:(Web 服务的合成)是指通过一定的方式方法把较小粒度和功能简单的 Web 服务组合成较大粒度和功能复杂 Web 服务的过程。可以描述为:$W = W_i \otimes W_j (i = 1,2,\cdots,N; j = 1,2,\cdots,N)$,其中 $\otimes$ 表示 Web 服务合成的运算符,W_i, W_j 表示较少粒度和功能简单的 Web 服务,W 表示大粒度和功能复杂的 Web 服务。

2　合成数据挖掘算法

在网格环境下,各种共享的软件、数据、设备都可定义为服务。数据挖掘算法是一种服务,但单个数据挖掘算法服务功能有限,只有对已有的数据挖掘算法服务根据不同资源环境和不同用户的需求进行合成,来满足不同用户的要求。针对数据挖掘算法服务的特点,设计一种基于本体的服务合成方法。

2.1　数据挖掘算法本体

数据挖掘本体[5,6]通过对数据挖掘领域内概念及概念间关系的精确描述,提供一个人机之间、机器和机器之间互相理解的语义平台,能让机器相互了解数据挖掘算法的功能、输入、输出、数据的流动和接口的参数等。基于本体论的数据挖掘方法使用本体来表示领域知识是为已经存在的、被证明可以有效使用的数据挖掘算法建立本体,通过数据挖掘算法本体,协助数据挖掘领域的用户在实施数据挖掘过程中对众多可供选择的算法和方法进行选择和合成。在挖掘过程中定义已经存在的数据挖掘技术以及特征属性,算法本体主要包含以下信息:1)算法的创建时间、版本、其执行环境。2)算法的可访问性说明信息。3)算法的响应时间、抖动、响应时间、错误处理信息。4)算法的容量信息。5)每个操作的可读信息。6)对于每个操作,包括前提条件下以及该操作和前驱操作的兼容性。7)算法执行说明的结果的详细说明。8)说明阈值情况。9)对影响操作属性如速度、精度、模型复杂性的估计。

2.2　数据挖掘算法服务合成思想

基于本体数据挖掘算法服务组合方法的基本思想:首先确定一个数据挖掘算法服务,然后通过采用关联度和匹配技术找出该服务的后继服务,以此类推,直至找出所有的后继服务,形成了一个服务序列,然后对这个服务序列的服务进行合成。

2.3　数据挖掘算法服务合成过程

数据挖掘算法服务合成[7~9]步骤按实用性原则设计,基于用户输入的需求和现有的可用 Web 服务展开,服务合成具体步骤描述如下:

1)输入服务需求:包括输入数据挖掘服务领域本体和数据挖掘算法本体信息,及数据挖掘算法服务所能提供的输入参数信息、预期的输出结果、服务质量要求及比例和本体匹配的相似度阈值。

2)服务获取:从算法库和网格中获取算法服务。

3)分类:用 WSDL 对服务描述,根据数据挖掘领域对数据挖掘算法本体进行分类,符合要求的加入可用 Web 服务队列。

4)同等服务合一:比较可用 Web 服务队列中的服务对象,把输入和输出参数分别完全匹配的服务归入同一服务组别,每一组服务在可用服务队列中只保留一个服务对象;

5)接口参数匹配:根据输入输出接口,使用接口匹配算法在可用 Web 服务队列中发现、匹配。

6)生成合成方案:根据接口参数的匹配和用户的需求设计合成方案,并计算每个方案的执行代价,选择一条执行代价最小的合成方案。

7)转换:把合成方案转换 BPELAWS 代码。

8)合成失败处理:如果不能找到满足服务需求的服务组合,则把当前的合成工作状态反馈给管理员。

3　数据资源合成

数据资源合成是将多个数据库集成为一个统一的数据库视图,构造一种虚拟的数据库,它包括了多个实体的物理数据库。数据合成利用通用工具和其他服务提供商开发的专用接口对原有系统数据进行抽取、转换、清洗和装载,并对处理的数据用 XML 把数据封装到 XML 结构中。WSDL 描述 Web 服务接口规范的标准格式和服务的细节内容,告诉用户如何使用 Web 服务,利用 UDDI 在服务注册中心处将这些服务进行注册,服务注册中心接收请求者的查询,服务请求者通过使用 UDDI 在服务注册中心进行查询,找寻自己需要的服务,然后利用 Web 服务合成技术合成数据。

Web 服务技术提供了一个分布式计算技术,通过开放的 Internet 标准,利用 Web 服务技术进行数据合成的主要步骤如下:

1) 数据的抽取、分割、另存为基本服务。

2) 对服务提供端的数据用 Web 服务接口给予暴露,提供了可调用的标准化接口。

3)"异构数据语义描述"建立源数据与目标数据的映射关系。

4) 通过"基于接口合成算法"建立各个基本服务组合的有序序列。

5) 生成合成方案,把合成方案转换 BPELAWS 代码进行合成。

6)合成失败处理。

4 基于 Web 服务合成技术的网格数据挖掘系统设计

4.1 基于 OGSA 的数据挖掘的体系结构

开放的网格体系结构[10] OGSA 的概念是结合现有的网格标准、面向服务的体系结构以及 Web 服务技术。OGSA 是一种面向服务的网格结构,它建立在网格服务的基础上,把一切资源定义为服务。网格服务是一种扩展的 Web service,由 Web service 服务体系结构[11]可以看出,它提供了服务的发现者、服务的注册、服务的请求者。

该体系结构基本特点是:

1)开放性,能实时从网格环境中获取算法和数据处理工具,并对组装好的算法和工具实时发布。

2)组装性,能对选择的挖掘算法和工具按需组装。

3)分布性,算法的本体的构成、算法的描述、算法库、方案库都可分布在网格环境中。

4)分时性,各阶段工作都可分时进行。

5)自治性,能对各阶段工作进行实时评估,对发生错误实时处理。

文中所设计数据挖掘系统[12,13]是基于网格环境的,采用开放式体系结构,挖掘系统能与外界进行数据资源和算法资源进行交互。其体系结构图如图 1 所示。

4.2 网格数据挖掘系统

数据挖掘系统主要是由五大部分组成:门户、服务池、合成器、数据挖掘引擎、评估器。

1)门户:通过提供友好的人机界面帮助用户更准确提交数据挖掘服务合成请求,在领域和服务功能两个层次上对用户的数据挖掘请求进行限定,避免模糊性和歧义性。

2)服务池:主要用存取挖掘算法、挖掘方案和知识库。服务池是由 UDDI 注册中心、算法服务库和合成知识库构成。算法库负责管理数据挖掘算法,所有的挖掘算法均以 Web 服务的形式封装。UDDI 注册中心注册,当算法库中的某个挖掘算法被请求调用时,接受数据挖掘系统的挖掘请求,向挖掘系统提供合适的算法服务。数据挖掘服务合成知识库存放着领域知识、用户记录和组装方案。

3)合成器:根据挖掘用户的需要从知识库调用一些方案推荐给用户,并给出挖掘方案的评估报告,对用户进行导航,用户也可自定义一些方案。对选定的数据挖掘算法服务进行合成,在服务合成过程中使用推理机进行推理工作,最主要是完成服务匹配,使用户在服务合成过程找到合适的服务。在合成过程还可根据用户的具体需求调整和设置各个服务的可定制属性,完成用户所要求的服务功能。

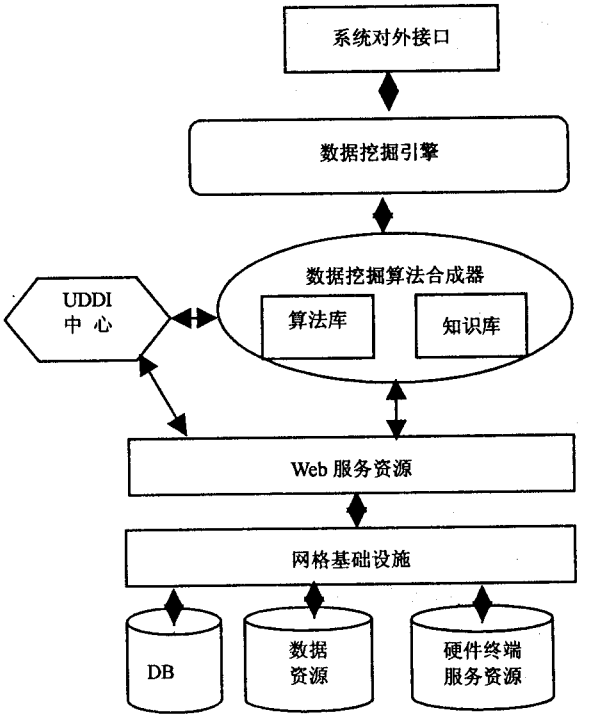

图 1 基于 OGSA 的数据挖掘体系结构

4)数据挖掘引擎:数据挖掘引擎是整个挖掘系统的核心部件。在系统中数据挖掘引擎主要是负责挖掘流程的定义、挖掘方案的设计、构建数据挖掘服务和挖掘的执行,还要负责与算法合成中心进行交互,以及与外界进行数据交换。

5)评估器:当服务合成产生多个方案时,对各个服务方案进行评估,选出最优的一个。

5 结束语

为了使数据挖掘用户方便和有效地使用数据挖掘服务,提出对数据挖掘服务进行合成。文中对数据挖掘服务合成进行分析,并探讨了用户利用数据挖掘服务合成系统根据实际需要选择合成方案进行合成。但由于网格技术、Web 合成技术和数据挖掘技术的复杂性,还有诸多问题仍待研究。

参考文献:

[1] Foster I,Kesselman C,Nick J,et al. Grid service for Distri-

(下转第 240 页)

3.1 编写 Qt/Embedded 应用程序

在宿主机上编译 Qt for X11 时生成的 Qt Designer 软件是一个非常流行的快速应用程序开发工具,用来设计界面和编制代码。在 Qt Designer 中,可以通过拖拉或点击的方式,在一张空白表单的适当位置上添加一些输入框和按钮等窗口组件。这时 Designer 工具会自动编写和维护代码。

使用 Qt Designer 进行 C++ 程序编制的基本步骤是:首先建立窗体,并根据应用的需要在窗体中添加控件。Qt 会将建立的窗体保存为.ui 文件,使用 Qt 提供的 uic 工具将文件转换为.h 和.cpp 文件,对于控制动作是需要手动添加不同的操作函数。之后使用 progen 工具为该应用程序建立.pro 工程文件,并通过 tmake 工具为该工程建立 Makefile 文件。最后,只需要运行 make 即可生成可执行文件。

在编写操作函数时,本系统通过计算相邻监测点的观测值确定每个电池芯的电压值。通过查询选用电池芯的放电结束电压查找表[6],得出每个电池芯的剩余电量。当电池组电量降至一定阈值或某个电池芯的储电量低于 3%时[7],利用应用程序报警,通知用户予以充电或按照指示的电池编号及时进行个别更换。这样,既可保证电源的正常工作指标,又能够提高所有电池的利用率。

3.2 将应用程序植入 Qtopia

在 Qtopia 平台上发布自己的应用程序,需要三个文件:一个执行文件、一个启动器文件、一个图标文件。执行文件就是前面讲到编写并编译生成的可执行文件,需要将该可执行文件保存在 qtopia/bin 目录下。图标文件就是为应用程序制作 48 * 48 大小的 PNG 格式的小图像,它一般存放在 qtopia/pics 目录下。同时还需要建立应用启动器(.desktop)文件,把它保存在 qtopia/apps/Applications 目录下。将上述文件分别复制好以后,重新运行 Qtopia,就可以看到添加的应用图标,点击此图标便可运行该应用程序了。

4 结束语

降低系统能耗需要嵌入式系统硬件环节提供可靠支持,同时也离不开嵌入式操作系统和应用程序对硬件资源的合理管理[8]。现有的 Linux 提供了电源管理的部分功能,但由于具体进行嵌入式系统设计时的灵活性,一般需要根据特定需要而定制应用程序对电源电量进行精确的监测与提示。通过对手持式终端设备中电源监测技术的研究和实现,将有助于其它类型嵌入式系统的电源监测、管理工作的完成,亦有望对类似系统的建立具有一定的参考和借鉴价值。

参考文献:

[1] 朱玉军,王香凤.笔记本电脑电池简介[J].化学教学,2006(4):30-32.

[2] 原 亮,王盼卿,陈立云,等.装备信息化工作中嵌入式系统的研究实现[J].军械工程学院学报,2007,8(1):79-81.

[3] 潘巨龙,黄 宁.Arm 嵌入式 Linux 系统构建与应用[M].北京:北京航空航天大学出版社,2006.

[4] Bovet D P,Cesati M.深入理解 Linux 内核[M].北京:中国电力出版社,2004.

[5] 倪继利.Qt 及 Linux 操作系统窗口设计[M].北京:电子工业出版社,2006.

[6] 陈金舟.储氢合金测试系统及镍氢电池管理系统的设计[D].北京:中国科学院研究生院,2004.

[7] 华宏懿.镍氢蓄电池的数学建模及其电池管理系统实现[D].北京:北方工业大学,2006.

[8] 阳富民,梁 晶,张 杰,等.嵌入式 Linux 电源管理技术的研究与实现[J].计算机工程与科学,2004,26(12):92-93.

(上接第 236 页)

buted System Integration[J]. IEEE Computer,2002,35(6):37-46.

[2] 侯文国,傅秀芬,谢翠萍.网格的数据挖掘[J].计算机应用研究,2004(10):241-247.

[3] HAN Jiawei, Kamber M. Data Ming Concept and Techniques [M].[s.l.]:Morgan Kaufmann Publishers Inc,2001.

[4] 顾 宁,刘家茂,柴晓路.Web Services 原理与研发实践[M].北京:机械工业出版社,2006.

[5] 邱莉榕,史忠植,林 芬,等.基于主体语义 Web 服务自动组合研究[J].计算机研究与发展,2007,44(4):643-650.

[6] 邹力鹏.数据挖掘方法本体研究[J].计算机科学,2005,32(3):197-199.

[7] 张佩云,孙亚民.动态 Web 服务组合研究[J].计算机科学,2007,34(5):4-7.

[8] 袁 琴,杨小虎.基于本体分类的 Web 服务合成的研究及应用[J].计算机工程,2007,33(2):79-81.

[9] 李 曼,王大治,杜小勇,等.基于领域本体的 Web 服务动态组合[J].计算机学报,2005,28(4):644-650.

[10] 蔡 刚.基于网格的分布式数据挖掘体系结构研究[D].重庆:重庆大学,2007.7.

[11] Box D. Simple object Access Protocol(SOAP) 1.1[EB/OL]. 2001. http://www.w-3.org /TR /SOAP/.

[12] 侯敬军,曾致远,向 凌.一种基于 Web 服务的分布式数据挖掘体系结构[J].微机发展,2004,14(6):48-51.

[13] 吴小竹.利用 Web service 技术构建开放式数据挖掘系统[J].计算机工程与设计,2007,28(15):3563-3565.