

基于网格服务的数据挖掘应用研究

郭小雪

(茂名学院 理学院, 广东 茂名 525000)

摘 要:建立在网格基础上的数据挖掘结合了网格技术的优点,能够对 Internet 上广域分布的海量信息进行高效的处理、分析和挖掘。分析了网格与数据挖掘的特点,并结合网格与数据挖掘的过程和关键技术,详细介绍了开放网格服务体系结构、层次功能、网格服务及其接口,基于 OGSA 的网格数据挖掘的例子和应用验证了数据挖掘网格系统的可行性和高效性。

关键词:网格服务;数据挖掘;开放式网格服务结构;服务接口

中图分类号:TP311.13

文献标识码:A

文章编号:1673-629X(2008)12-0224-04

Research on Application of Data Mining Based on Grid Service

GUO Xiao-xue

(College of Science, Maoming University, Maoming 525000, China)

Abstract:Data mining based on grid environment which integrates the merits of grid, can transact and analyze the vast information, and discover new knowledge. The features of grid and data mining are analyzed. Summarize the procedure and key technologies of data mining with grid characteristic, mainly discuss the open grid services architecture, layer functions, grid service and service interface of data mining on grid. Finally give an example of data mining based on OGSA, the feasibility and the efficiency are proved.

Key words:grid service; data mining; OGSA; service interface

0 引言

随着 Internet 的普及和计算机软硬件技术的发展,网格技术越来越得到人们的重视,网格已经被认为是下一代的互联网^[1]。网格是构筑在 Internet 上的一组新兴技术和基础设施,其目标是在动态变化的、广域分布的异构虚拟组织间实现协同资源共享、多领域的科学和工程的问题求解。网格技术的兴起就是为了突破计算能力和地理物理位置的限制,节约资源,实现世界范围的资源共享与服务协作^[2]。网格计算技术是解决复杂海量科学数据的访问、存储、组织和管理的一种有效技术。

未来的科学计算以数据为中心。数据已成为科学、军事、电信、医疗等各个领域的重要资源。在网格计算环境下,许多科学与工程计算问题,如高分子材料分析、生物计算、数字地球等,以及信息服务、大型跨国企业、远程医疗合作将产生大量的数据。要分析和挖

掘这些广域分布的海量数据,以获取新的科学知识、规律和决策支持信息,传统的数据挖掘模式和技术已经无法胜任。建立在网格基础上的数据挖掘结合网格计算的思想及其技术的优点,能够对广域分布的海量数据进行高效的处理、分析和挖掘,给科学研究领域、经济领域和社会生活带来新的发现和巨大的价值^[3]。

1 数据挖掘和网格

数据挖掘(DM, Data Mining)是一个利用各种分析方法工具对海量数据进行分析,建立模型和发现数据间联系,并在商业、科研等领域进行应用,辅助做出基于知识预测、决策的过程。数据挖掘指“从数据库或数据仓库中发现隐藏的、预先未知的、有趣的信息的过程,该过程可以看作是知识发现中的一个核心的步骤”。这门新兴的科学研究领域自从诞生后就成为研究的热点,至今方兴未艾^[4]。数据挖掘的范围非常广泛,可以是经济、工业、农业、军事、社会、商业、科学的数据和卫星观测得到的数据。数据的形态有数字、符号、图形、图像、声音等。数据组织方式也各不相同,可以是有结构、半结构、非结构的。数据挖掘的结果可以表示成各种形式,包括规则、法则、科学规律、方程和概

收稿日期:2008-06-22

基金项目:2006 佛山市产学研专项资助项目(2006A018);茂名市科技计划项目(2007029)

作者简介:郭小雪(1979-),女,广东高州人,硕士,讲师,研究方向为分布式计算及网络应用。

念网。数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程。

数据挖掘就是从大量的数据中发现或“挖掘”知识,而网格上含有丰富的数据,是数据挖掘的理想目标^[5]。网格的数据挖掘建立在数据网格的基础设施和相关技术的基础上,在广域分布的海量数据和计算资源的环境中发现数据模式,获取新的科学知识和规律。这个网格计算环境提供特殊的数据管理、数据存储、数据复制和安全监控等功能。

2 基于网格的数据挖掘

2.1 网格数据挖掘过程

网格数据挖掘的基本过程分为以下三种^[6]:

(1)数据的处理。数据的处理阶段主要完成从数据网格环境中收集广域分布的数据和计算资源,并对原始数据进行归档处理,更正校对,过滤清理和数据的转换、合并。最后再对经过处理后的数据进行归档。

(2)数据的分析与挖掘。这阶段主要完成对处理后的数据进行分析、概括和挖掘,生成关联的规则,发现新的数据关系等,并归档概括出来的数据。

(3)模式的评价。这阶段对处理后的数据和归纳后的数据再次进行分析,得出一些数据模式,并评价数据挖掘结果的有效性和可靠性,提交得出的结论或新的关系、趋势。

2.2 网格数据挖掘特点

(1)超级计算能力。网格计算能够为科学计算领域和社会经济生活领域提供超级的计算能力。

(2)具有分布性和动态性,数据分布范围广。在网格计算环境中,广域分布的各种资源都是动态创建和删除的,因此,网格的数据挖掘系统具备分布性和动态性,并以分布计算的方式并考虑数据流通负载来分析数据。

(3)具有高性能的 I/O 负载平衡能力。在对广域分布的海量数据处理的过程中,无论是数据的远程传输还是挖掘过程中的数据处理、分析挖掘、模式评价等过程,数据的工作流都是很大的。这需要网格提供网络负载调度、管理和高性能的 I/O 负载平衡能力。

(4)高效的数据存储服务、传输服务和复制管理。在数据挖掘过程中要进行大数据集存储、复制的时候,网格能够提供高效的广域网数据高速缓存服务以解决网络带宽管理的问题;数据传输策略能够支持多种存储系统,并行数据传输,部分文件传输和数据重传、容错能力;数据复制策略能在不同站点之间高速移动和复制数据,保持远程数据拷贝的一致性。

(5)网络安全性要求更高。数据挖掘涉及广域分布的属于不同虚拟组织的数据源,数据的安全性和访问权限问题至关重要。在数据存储、传输、复制管理和网络通信过程中,网格操作系统必须具有抗拒各种非法攻击和入侵的能力,保证系统正常高效运行和各种信息的安全。

3 网格环境数据挖掘体系结构

3.1 开放网格服务体系结构

数据挖掘是一个复杂的处理过程,可通过多种方法加以实现。数据的分布式特征和信息共享的可扩展性使网格成为数据挖掘应用较为合适的方案。开放网格服务体系结构(Open Grid Services Architecture, OGSA)为网格环境数据挖掘的实施创造了条件^[7]。开放网格服务体系结构是 Globus 标准与以商用为主的 Web Services 的标准相结合的产物,是目前最新也最有影响力的一种网格体系结构,其目的就是将 Grid 尤其是 Globus 的一些功能融合到 Web Service 框架中。OGSA 是面向服务的结构,即将所有事务都表示成一个 Grid 服务,计算资源、存储资源、网络、程序、数据等都是服务,所有的服务都联系对应的接口,通过标准的接口和协议支持创建、终止、管理和透明的服务。

相对于五层沙漏结构的以协议为中心的协议结构,OGSA 是以服务为中心的服务结构^[8],如图 1 所示。这里的服务所指的概念更为广泛,包括各种计算资源、存储资源、网络、程序、数据库等,一切均是服务。在 OGSA 中实现的是对服务的共享。它将资源、信息、数据等统一起来,十分有利于灵活、一致、动态共享机制的实现,使得分布式系统管理有了标准的接口和行为。为了使服务的思想更加明确和具体,OGSA 定义了网格服务的概念,它是 Web 服务的一个扩展。它把 Globus 标准与面向商业应用的万维网服务结合起来,把网格计算从科学与工程计算应用扩展到更广泛的以分布式系统服务集成为主要特征的商业应用领域。OGSA 将一切都看做是网格服务。网格服务可以不同的方式聚集起来满足虚拟组织的需要。虚拟组织自身也可以部分地根据它们操作和共享的服务来定义。

与五层模型一样,在 OGSA 中也非常重视互操作性。从服务的观点看,OGSA 将互操作性问题转换为两个子问题,即定义服务的接口和识别激活特定接口的协议。以网格服务为中心模型具有如下优点:

1)由于网格环境中所有的组件均是虚拟的,通过提供一组相对统一的核心接口,所有的网格服务均基于这些接口实现,就可以很容易地构造出具有层次结构的、更高级别的服务。这些服务可以跨越不同的抽

象层次,以一种统一的方式来看待。

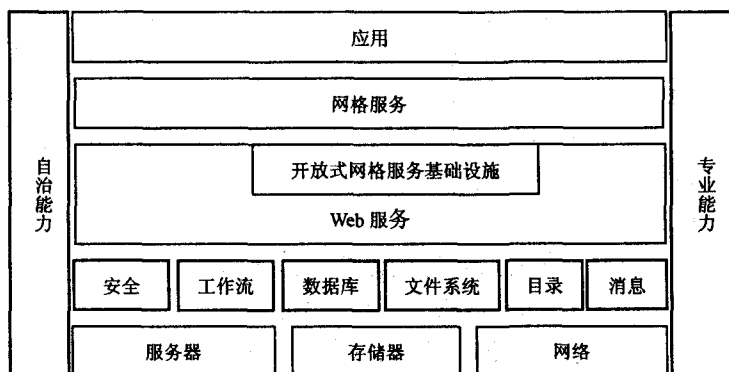


图 1 开放网格服务结构

2)虚拟化也使得将多个逻辑资源实例映像到相同的物理资源上成为可能,在对服务进行组合时不必考虑具体的实现,可以在底层资源组成的基础上,在虚拟组织中进行资源管理。通过网格服务的虚拟化,可以将通用的服务语义和行为,无缝地映像到本地平台的基础设施之上。

Web 服务是网格服务及 OGSA 的基础,因此了解 Web 服务的体系结构能够更好地理解和部署网格服务。Web 服务是一种分布式计算技术(类似 CORBA、RMI、EJB 等),它能在所有支持 Internet 通信的操作系统上实现,目前被大量部署于商业应用的 C/S 模式中。Web 服务的基本结构是基于服务提供者(Web 服务容器)、请求者(客户机)和中介者(UDDI 注册中心)三个角色之间的交互、交涉及服务的发布、发现和服务请求者与提供者之间的绑定三个动作。可简单地归结为客户根据需求向 Web 服务发送服务请求,Web 服务向客户返回服务结果。Web 服务的所有协议均基于标准的 Web 协议,如 HTTP、XML、SOAP、WSDL、UDDI 等。这些协议组成堆栈的形式,每一个下层提供对它上层的支持;同时每一个上层均必须基于所有的下层协议之上。开放网格服务结构基于统一的 Web 服务框架。一个 Web 服务就是一个可以被 URI 识别的软件应用。其接口和绑定可以被 XML 语言描述和发现,并且可以通过基于 Internet 的协议直接支持与其他基于 XML 的软件应用进行交互^[9]。

OGSA 架构从下到上依次是:

(1)资源层。它包括物理资源和逻辑资源。物理资源包括存储器、网络、计算机、显示设备、服务器和其他相关的本地服务。逻辑资源通过虚拟化和聚物理层的资源来提供额外的功能和通用的中间件,如文件系统、数据库、目录、工作流管理和安全认证等,在物理网络上提供这些抽象服务。

(2)Web 服务层。在这一层所有的网格资源(物理

的和逻辑的)均被建模为服务。OGSI(Open Grid Services Infrastructure)规范定义了网格服务并建立在标准 Web 服务技术之上。OGSI 进一步扩展了 Web 服务的定义,利用如 XML 和 WSDL 这样的 Web 服务机制,为所有网格资源指定标准的接口、行为和交互,提供动态的、有状态的和可管理的 Web 服务的能力。

(3)网格服务层。基于 OGSI 架构的网格服务层是 Web 服务层及 OGSI 扩展为上一层提供的基础设施。目前,研究人员致力于在程序执行、数据服务、核心服务等方面定义基于网格架构的服务。定义这些核心网格服务,主要是因为它们最有可能得到大多数高级服务的利用。实现这些高级服务或者是为了支持程序执行,或者是支持数据访问,或者是将它们实现为特定领域的服务。

(4)网格应用层。随着基于网格架构的服务不断被开发出来,使用一个或多个基于网格架构的服务的新网格应用程序亦将大量出现,构成网格应用层。

3.2 网格服务

在网格应用环境中,大量的服务是临时性的、短暂的服务,而非 Web Service 中的永久性服务。为了使服务的思想更加明确和具体,OGSA 在 Web Service 服务概念的基础上,定义了“网格服务”的概念。网格服务是特殊的 Web 服务,具有特定功能的网络化实体。它定义了一组接口,这些接口的定义明确并且遵守特定的惯例,用于解决服务发现、动态服务创建、服务生命周期管理、通知等与临时服务有关的问题。基于网格服务的概念,网格就是可动态扩展的网格服务的集合,可以以不同的方式聚集起来满足虚拟组织的需要,虚拟组织自身也可以部分地根据他们的操作和共享的服务来定义。

3.3 服务接口

网格服务是由它们提供的能力来刻画的,包括服务接口和服务数据。一个网格服务实现一个到多个接口,每一个接口定义了一些操作,这些操作通过交换定义好的一系列消息来激活,并完成不同的功能。服务数据是关于网格服务实例的信息。

网格服务可以表示为“网格服务=接口/行为+服务数据”。在 OGSA 中,目前已提供服务生命周期管理、创建临时服务、注册服务、主键服务、消息发布服务、消息接受服务、句柄映射 7 个服务接口^[10]。其中服务生命周期管理服务接口是必需的,每一个服务接口提供了相应的操作。网格服务通过提供一组相对统一的核心服务接口,所有的网格服务都基于这些接口实现,这样就可以很容易地基于简单的、基本的、具体

的服务构造出具有层次结构的、更高级别抽象的服务。这些服务可以跨越不同的抽象层次,以一种统一的方式来看待,有利于通过统一的标准接口来管理和使用网格,具体如表 1 所示。

表 1 网格服务的接口

接口	操作	描述
GridService	FindServiceData	查询网格服务实例的各种信息
	SetTerminationTime	设置并得到网格服务实例的终止时间
	Destroy	终止网格服务实例
NotificationSource	SubscribeToNotificationTopic	向通知发送者进行登记
	UnSubscribeToNotificationTopic	取消登记
NotificationSink	DeliverNotification	异步发送消息
Registry	RegisterService	网格服务句柄的软状态注册
	UnRegisterService	取消注册的网格服务句柄
Factory	CreateService	创建新的网格服务实例
PrimaryKey	FindByPrimaryKey	返回根据特定键值创建的网格服务句柄
	DestroyByPrimaryKey	撤销特定键值创建的网格服务实例
HandleMap	FindByHandle	返回与网格服务句柄相联系的网格服务实例

4 网格服务的数据挖掘应用

下面给出一个基于 OGSA 数据挖掘的例子,它展示了基本的远程服务发现、激发、生命周期管理等功能。

(1)查找符合要求的服务。用户在虚拟组织所维护的注册表中查找数据挖掘功能(包括相应存储空间能力)的提供者。

(2)得到服务句柄。注册服务根据用户提出的要求,在众多的服务提供者中进行筛选,最后返回满足要求的服务提供者。

(3)创建服务实例请求。用户根据返回的服务句柄,向服务方提出请求,创建特定的服务实例,指定相关的参数,如实例存活的时间(服务生命周期),进行何种类型的数据挖掘操作等,这些请求需要与服务方进行协商。

(4)服务方创建满足要求的实例。图 2 中是数据挖掘方和存储能力提供方都创建了应用方要求的服务实例。

(5)新创建的数据挖掘服务实例以“用户”的身份,在不同的数据库中执行查询任务,这种基于用户身份的代理策略是由 OGSA 的安全机制支持的。

(6)得到结果。将查询结果存放在(4)申请到的存

储空间中。

以上是一个简单的基于 OGSA 框架的应用例子,基本描述了应用在 OGSA 框架下数据挖掘的工作过程和执行机制。

网格服务的数据挖掘可以应用到不同的领域中。现代商业竞争非常的激烈,任何一间大型的零售企业如果要想在市场上占得先机,就必须能够选择合适的地点开设售卖自己商品的商铺,以抢占更大的市场份额。因此如何更方便地选择最合适地点和方案来开设新的商铺,便成为企业最关心的问题。因此可以利用网格服务的数据挖掘的原理和流程为企业提供模拟的商业决策支持。

5 结束语

文中对网格数据挖掘理论进行了研究,在广域分布的海量数据和计算资源的环境中发现数据模式,获取新的科学知识和规律。目前,网格计算、网格数据库服务和网格的数据挖掘技术还不成熟,随着研究的深入和不断发展,数据挖掘的工具及其算法也必须在分布性、并行性和灵活性方面得到进一步发展和提高。随着网格和数据挖掘的技术不断提高,网格的数据挖掘将得到广泛的应用。

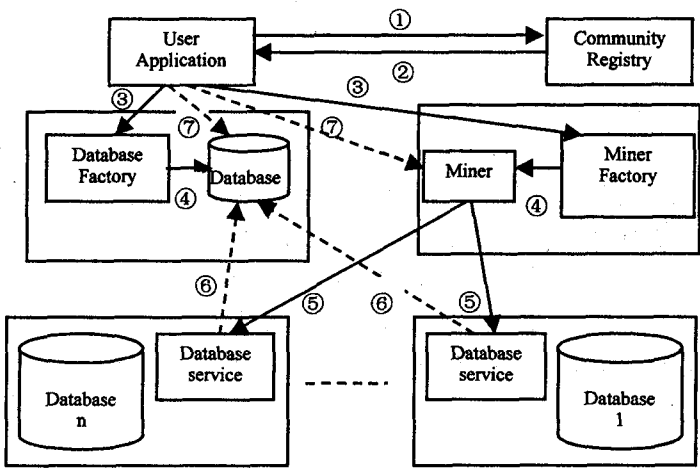


图 2 基于 OGSA 的数据挖掘

参考文献:

- [1] 王 铮,吴 兵. GridGIS——基于网格计算的地理信息系统[J]. 计算机工程,2003,29(4):38-40.
- [2] 胡 蓉,肖基毅. 基于知识网络的分布式数据挖掘[J]. 计算机技术与发展,2007,17(10):99-101.
- [3] 佟 强,周园春,吴开超,等. 科学数据挖掘网格服务框架[J]. 计算机应用研究,2007(6):26-29.
- [4] 庄新鹏,赵建民,朱信忠. 基于 Multi-Agent 的数据挖掘模型的研究[J]. 计算机技术与发展,2006,16(7):129-131.

(下转第 234 页)

(1)查全率 R (Recall)。Recall = 正确过滤掉的邮件数/应该过滤掉的垃圾邮件。数值越高,表示漏网的垃圾邮件就越少。

(2)准确率 P (Precision)。Precision = 正确过滤掉的邮件数/实际过滤掉的邮件。数值越高,表示将合法邮件误判为垃圾邮件的可能性越小。

(3) F 测试值。 $F = (2 * R * P) / (R + P)$ 。

实验结果见表 1。

表 1 邮件过滤系统实验测试结果(%)

数据集	查全率	准确率	F 值
KeyLib (1000)	96.2	90.3	93.156
CDSCE (1830)	94.5	87.1	90.649

实验显示,训练样本的选择的典型性直接影响着最终的过滤效果,与之前在采用朴素贝叶斯方法进行邮件分类相比较,在查全率和准确率上都有一定的改善,基本上可以满足垃圾邮件的过滤任务。

3 结束语

在目前广泛取得应用的大多数客户端垃圾邮件过滤系统中(如 FoxMail 6.0 等)都是采用的朴素贝叶斯方法^[8],朴素贝叶斯方法的优点是算法简单、执行速度快,但对样本的要求很高,而且分类精度上提升空间有限。文中给出了一种基于统计理论的垃圾邮件过滤方法——支持向量机进行垃圾邮件分类,它具有严格的理论基础,又能较好地解决小样本、非线性、高维数和

局部极小点等实际问题。最后通过实验证明这种方法在垃圾邮件分类中已经可以很好地完成分类任务,确定系统中的过滤精度,在查全率与准确率上较之朴素贝叶斯分类有了很大的提升。但是支持向量机在计算时间和空间要求较高,这也从一定程度上影响了它的应用推广。如何在保持分类精确度的前提下,提高算法的效率,这是下一步需研究的课题。

参考文献:

- [1] Christopher J, Burges C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998(2):121-167.
- [2] Hsu Chih-Wei, Lin Chih-Jen. A comparison of methods for multi-class support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2):415-425.
- [3] 史忠植. 知识发现[M]. 北京:清华大学出版社, 2002.
- [4] Ronnie, Rifkin R. Improving Multi-class Text Classification with the Support Vector Machine[DB/OL]. 2002-05-23. www.ai.mit.edu/research/abstractabstracts2001/machine-learning.
- [5] 安全龙,王正欧. 一种新的支持向量机多类分类方法[J]. 信息与控制, 2004(3):262-267.
- [6] 王斌,潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报, 2005, 19(5):1-3.
- [7] 曹麒麟. 反垃圾邮件的研究[D]. 北京:清华大学电子工程系, 2002.
- [8] 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):241-243.
- [9] 林雯,段小斌,谢晓兰. 数据挖掘技术在中小型制造企业 CRM 中的应用[J]. 计算机技术与发展, 2007, 17(11):247-250.
- [10] 刘天鹏,周娅. P2P 网络中的数据挖掘[J]. 计算机应用, 2008(1):162-164.
- [1] 王永蛟,莫国良. 植物的三维建模研究进展[J]. 计算机应用研究, 2005(11):1-4.
- [2] Weber J, Penn J. Creation and Rendering of Realistic Trees[J]. SIGGRAPH, 1995, 64(8):119-127.
- [3] 康孟珍, De Reffye P, 胡包钢, 等. 快速构造植物几何结构的子结构算法[J]. 中国图形图像学报, 2004, 9(3):79-86.
- [4] 李庆忠, 韩金妹. 基于 IFS 的树木形态模拟真实感的研究[J]. 微机发展(现更名为:计算机技术与发展), 2005, 15(7):86-88.
- [5] 孙家广. 计算机图形学[M]. 北京:清华大学出版社, 2002:369-373.