网格的数据挖掘*

侯文国, 傅秀芬, 谢翠萍
(广东工业大学 计算机学院, 广东 广州 510090)

摘　要: 网格是网络计算、分布式计算和高性能计算技术研究的热点。随着科学计算领域中的数据剧烈增长以及未来网格计算环境下广域分布的海量数据共享成为现实, 数据挖掘技术将在挖掘有效的信息、发现新的知识和规律发挥着重要的作用。结合网格的特点, 概述了网格数据挖掘的特点和关键技术, 重点讨论了网格数据挖掘的体系结构和基本过程, 最后给出了基于 OGSA 的网格数据挖掘的例子。

关键词: 网格; 网格计算; 数据挖掘; 体系结构; OGSA; Globus
中图法分类号: TP274　　**文献标识码:** A　　**文章编号:** 1001-3695(2004)10-0241-03

Data Mining on Grid

HOU Wen-guo, FU Xiu-fen, XIE Cui-ping
(College of Computer, Guangdong University of Technology, Guangzhou Guangdong 510090, China)

Abstract: Grid is now a hotter research field of grid computing, distributed computing and high performance computing technologies. As the number of data in the scientific computational field is enormously increasing and in the future a mass of shared data distributed geographically will be realized in the grid computing environment, data mining technology will play an important role in discovering value information, new knowledge and laws. Summarize the characteristics and key technologies of data mining with grid characteristics. Then we mainly discuss the architecture and basic processes of data mining on grid. Finally we give an example of data mining based on OGSA.

Key words: Grid; Grid Computing; Data Mining; Architecture; OGSA; Globus

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[1]。数据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、经济、医疗等领域的重要资源。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是无法胜任的。建立在数据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。

1　网格的数据挖掘

1.1　网格数据挖掘的基本概念

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[5]。

收稿日期: 2003-12-23; 修返日期: 2004-02-13
基金项目: 广东省自然科学基金项目资助(32497)

1.2　网格数据挖掘的特点

(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。网格的数据挖掘系统建立在网格计算的技术基础上, 数据的传输具有高效的并行性特点, 而且数据处理能力超强。

(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, 能够灵活调整数据搜索的范围; 另外, 由于数据的分布性特点, 数据挖掘算法是以分布计算的方式并考虑数据流通负载来分析数据的。

(3) 具有高性能的 I/O 负载平衡能力。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。网格上的资源是动态连接到 Internet 上的, 不同网格节点之间的通信和用户向网格计算环境提交任务和监控管理任务也是通过 Internet 连接的。同时,

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。因此,网格数据挖掘系统的安全性能要求更高,解决的方案也更复杂。

2 网格数据挖掘的关键技术

2.1 Globus 中间件系统

Globus[1]是有效支持网格计算环境与网格应用的中间件系统,为网格的数据挖掘提供一个运行的环境。Globus 在网格计算逻辑结构中的组成部分包括:网格安全基础设施、Globus 资源分配管理、元计算目录服务、全局二级存储服务、网格 FTP 服务和 Globus 复制管理[3]。

网格安全基础设施负责在广域网络下的安全认证和加密通信,提供单点登录功能、远地身份鉴别功能、数据传输加密功能等,是保证网格计算环境安全性的核心部分,为广域分布下的数据库系统提供了安全可靠的数据访问和数据传输。资源分配管理负责远程应用的资源请求处理、远程任务调度处理、远程任务管理等工作,是网格计算环境中的任务执行中心。网格 FTP 服务 GridFTP[1]是一个高性能、安全、可靠的数据传输协议,并针对高宽带的广域网络环境进行了优化,具备支持第三方传输、断点续传、并行传输、与结合的取暖器认证、缓存等特征,为网格的数据挖掘提供高效的数据传输服务。数据复制管理负责产生一个数据包的全部或部分数据的拷贝,将新的拷贝注册到复制文件目录中等功能。在网格中实现数据复制功能,是为了获得更好的数据访问效率和容错性。

2.2 网格数据库管理技术

网格数据库管理为用户提供了一组网格服务接口,包括网格数据库服务注册、网格数据库服务工厂、网格数据库服务和网格数据传输,以供用户访问广域网络上互连的各种异构数据资源。网格数据库服务是以基于服务的方式访问已有的数据库,它独立于各种数据库,与其他的 Web 服务和网格服务共存,与网格认证和授权机制相衔接,采用 XML 文档方式进行服务描述 WSDL。因此,网格数据库服务构建了一种可靠性高,可用性强的数据库服务和网格服务,使得用户共享使用网格数据库和访问广域网络上互连的各种异构数据库系统,为海量数据的存储、传输、复制和管理提供了很大方便。

2.3 XML 技术

XML 是一种用来描述数据的语言,提供了独立的运行程序的方法来共享数据。XML 能够使不同来源的数据很容易结合在一起,解决了数据的统一接口问题。Globus 用 XML 来描述各种信息,并且与 Web Service 技术中的 SOAP, WSDL 和 WS-Inspection 结合,提供更加方便和有效的元数据访问服务和元计算信息服务,有效解决了半结构化的数据源问题。

2.4 网格数据挖掘的体系结构和算法

网格的数据挖掘是面向广域分布的大型、异构数据库的。因此,设计合理的数据挖掘的体系结构对整个数据挖掘系统的影响重大。另外,处理的数据非常庞大及数据挖掘方法的计算复杂性,高效的网格数据挖掘的算法必须具备分布计算和并行计算的能力。

3 网格数据挖掘系统

3.1 网格数据挖掘系统的体系结构

网格的重要特点是分布性和动态性。网格的各种资源不是集中在一起,而是分布在地理位置互不相同的各个地方,各种资源也是动态变化的。正是网格的这些特点及其分布式环境,使得网格的数据挖掘系统不再局限于传统的集中式数据挖掘系统,而是和网格一样具有分布性、动态性和自适应性。

网格的数据挖掘系统采用分布式的组件架构和自适应的分布技术,由一系列的组件集成,组件之间可以实现互相通信和数据交换。这种基于分布式组件技术的体系结构允许更大的弹性,包括集成不同的协议、应用程序接口、应用程序、操作系统和硬件,能够提供多级的抽象能力、高可靠性、可扩充性和安全性。其体系结构如图 1 所示。

图 1 网格数据挖掘系统的体系结构

3.2 网格数据挖掘系统的基本模型

按照这种分布式的组件构架技术和网格的可扩展性和自相似性的特点,基本的网格数据挖掘系统的模型可以分成三类:①模型的本地数据分析遵循下一代的全局数据模型 - 自适应的分布式数据挖掘技术。这种模型没有数据的复制过程;②数据挖掘模型的组件有选择性地分布在网格计算环境中,系统中没有动态的数据复制;③组件同样是选择性地分布在网格计算环境中,而且系统存在动态的数据复制过程。

3.3 网格数据挖掘的基本过程

根据上述网格数据挖掘系统的体系结构,网格数据挖掘的基本过程(图 2)分为以下三种:

(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。这阶段由于数据的归档处理,存在相对静止的索引数据,数据的处理功能比较稳定,而且对归档的数据进行周期性的数据再处理。

(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。

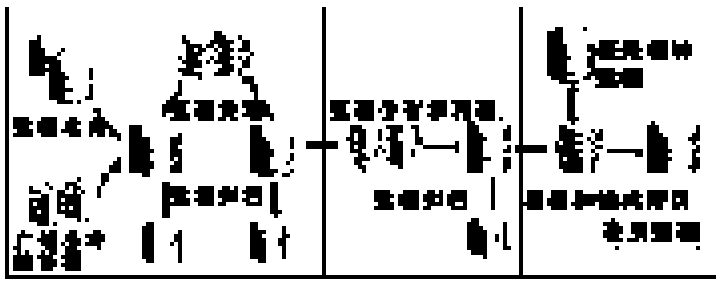

图 2　网格数据挖掘的基本过程

4　基于OGSA 的数据挖掘

4.1　基于 OGSA 的数据挖掘的体系结构

开放网格服务结构[3] (Open Grid Service Architecture, OGSA) 是一个面向服务的网格体系结构, 它建立在网格服务 (Grid Service) 的基础上, 将一切都抽象为服务, 包括计算资源、存储资源、网络、程序、数据库、仪器设备等。这种观念有利于通过统一的标准接口来管理和使用网格, 容易构造出具有层次结构的、更高的服务, 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, 确保异构系统间的互操作性。网格服务是一种扩展的 Web Service , 该服务提供一组遵守特定约定的定义明确的接口, 这些接口解决服务发现、动态服务创建和删除、生命周期管理、通知和可管理性等方面的问题, 所有的服务都遵循指定的网格服务接口和行为。

数据挖掘应用程序属于 OGSA 的应用层, 并由网格服务提供一组接口和管理。图 3 基于 OGSA 的数据挖掘的体系结构说明了数据挖掘应用程序与 OGSA 之间的关系。

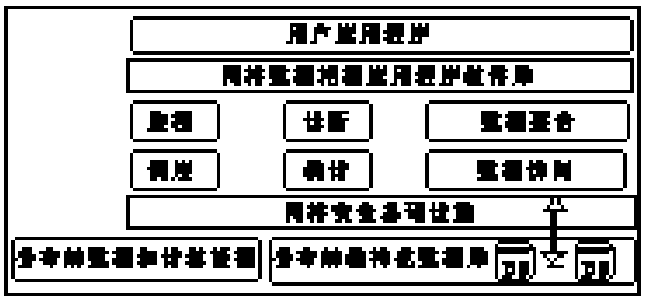

图 3　基于 OGSA 的数据挖掘的体系结构

4.2　基于 OGSA 的数据挖掘实例

下面给出一个基于 OGSA 数据挖掘的例子[2], 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 4 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。

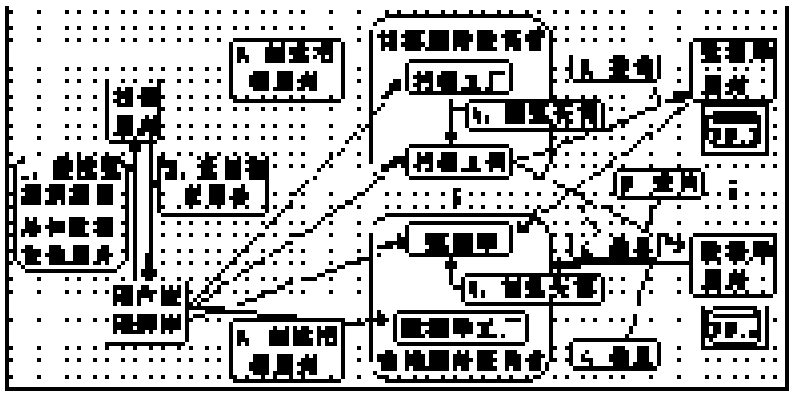

图 4　基于 OGSA 的数据挖掘实例

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。

5　结论

本文介绍了网格的数据挖掘。在结合网格计算、Globus, 网格数据库管理、XML 等技术下, 提出了网格数据挖掘的体系结构及其基本过程。这种结构的最大特点是将网格数据库服务和网格中间件系统紧密结合起来, 并将一切都看作服务, 极大地扩展了网格计算的应用领域和数据挖掘的有效性。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。

参考文献:

[1]　an Foster, et al. The Anatomy of the Grid[Z]. 2001.

[2]　Ian Foster, et al. Grid Services for Distributed System Integration[Z]. 2002.

[3]　都志辉, 等. 网格计算[M]. 北京: 清华大学出版社, 2002.

[4]　Malcolm Atkinson. Future Infrastructure for Scientific Data Mining, Integration and Visualisation[Z]. 2002.

[5]　Peter Brezany. Grids, Grid Technologies and Data Mining[Z]. 2003.

作者简介:

侯文国 (1979-), 男, 硕士研究生, 研究方向为网络软件及网格技术; 傅秀芬(1957-), 女, 教授, 研究方向为多媒体网络软件、中间件技术等; 谢翠萍(1980-), 女, 硕士研究生, 研究方向为网络软件及网格技术。

(上接第 240 页)

卷) [M]. 北京: 清华大学出版社, 1999.

[2]　美] W Richard Stevens. TCP/IP 详解 卷一: 协议[M]. 北京: 机械工业出版社, 2000.

[3]　卢显良. UNIX 系统管理[M]. 北京: 清华大学出版社, 1993.

[4]　周明天, 汪文勇. TCP/IP 网络原理与技术[M]. 北京: 清华大学出版社, 1993.

作者简介:

王涛(1978-), 男, 博士研究生, 主要研究方向为计算机网络、计算机操作系统; 卢显良(1943-), 男, 教授, 博士生导师, 主要研究方向为计算机系统软件、计算机网络、操作系统。