基于知识网格的分布式数据挖掘

胡 蓉,肖基毅
(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要:科学和工商业应用需要分析分布在各异构站点的海量数据。这就需要合适的分布式并行系统来存储和管理数据。网格为分布式数据挖掘和知识发现提供了有效的计算支持。文中在讨论知识网格体系结构的基础上,利用可视化网格应用环境VEGA实现了基于网格的分布式数据挖掘过程。

关键词:知识网格;VEGA;数据挖掘;知识发现
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2007)10-0099-03

Distributed Data Mining Based on Knowledge Grid

HU Rong, XIAO Ji-yi
(School of Computer Science and Technology, Nanhua University, Hengyang 421001, China)

Abstract: In many scientific, industrial and commercial applications, it is often necessary to analyze large data sets over geographically distributed sites. An appropriate distributed and parallel system is needed to store and manage the large data sets. The grid plays an important role in providing a computational support for distributed data mining and knowledge discovery. Based on the discussion of knowledge grid architecture, the visual environment of grid application—VEGA is applied to fulfill the process of distributed data mining on the grid.

Key words: knowledge grid; VEGA; data mining; knowledge discovery

0 引 言

随着科学、工业、商业等领域的发展,出现了大量的TB级甚至PB级的大规模数据集,在这些数据集中包含了大量的对生活、生产、科学研究等具有决策性作用的有用信息,那么如何从这些海量数据中提取信息是人们面临的一个重大的问题[1]。显然,原先的集中式数据挖掘模式已无法满足人们的需求,这就需要探索出面向分布式数据挖掘的体系结构和工具。

1 网 格

网格是建筑在Internet上的新兴技术和基础设施。现在还无法给出网格的基础定义。Lan Foster认为网格[2]是一个集成的计算资源环境,或者说是一个计算资源池(Computing Pool)。通过网络共享机制,将地理上分布的计算资源、存储资源和网络资源组织成一个虚拟的超级计算机。

网格提供了强大的计算能力。用户可以方便地使用网格所提供的集成的、动态的、可扩展的、可控制的、智能协作的服务。网格的作用是将分散在网络上的信息及信息存储、处理能力以合理的方式"粘合"起来,形成有机的整体,以提供比任何单台高性能计算机都强大得多的处理能力,实现信息的高度融合和共享。它与传统的分布式计算不同,主要关注大规模的资源共享和高性能计算。目前网格已用于协调资源共享和解决工商业领域动态、多机构虚拟组织间的问题。故网格可以作为分布式高性能计算和数据处理的有效架构。

数据网格可以存放大规模数据集,支持单点登录,避免重复验证,支持分布式密集型数据应用的实现。典型的数据网格有欧洲数据网格。Globus[2]数据网格项目也定义和开发了持久的数据网格中间件。数据网格中间件对于网格中数据的管理是非常重要的,但是仍然需要有专门的工具和环境来支持科学和商业领域的数据分析和发现。

知识网格[3]代表了数据网格的发展,为网格中分布式数据挖掘和抽取提供了高级工具和技术。知识网格是设计和实现分布式高性能知识发现应用环境的体系架构,用于执行网格中的数据挖掘,进行科学发现,发现有用的商业信息。

收稿日期:2006-12-14
基金项目:湖南省教育厅科研项目(06C724)
作者简介:胡 蓉(1982-),女,安徽绩溪人,助教,硕士研究生,研究方向为网格数据挖掘;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。

2 知识网格体系结构

知识网格体系结构[4]是在 Globus toolkits 网格工具集和服务的基础上定义的。在 Globus 中,知识网格集成局部服务以提供全局服务。知识网格体系结构保证了数据挖掘工具和底层的网格机制和数据网格服务兼容。

知识网格服务由两层构成:核心知识网格层和高级知识网格层。核心知识网格层是在基本网格服务的基础上直接实现的服务;高级知识网格层用于描述、开发和执行分布式知识发现计算。如图 1 所示。

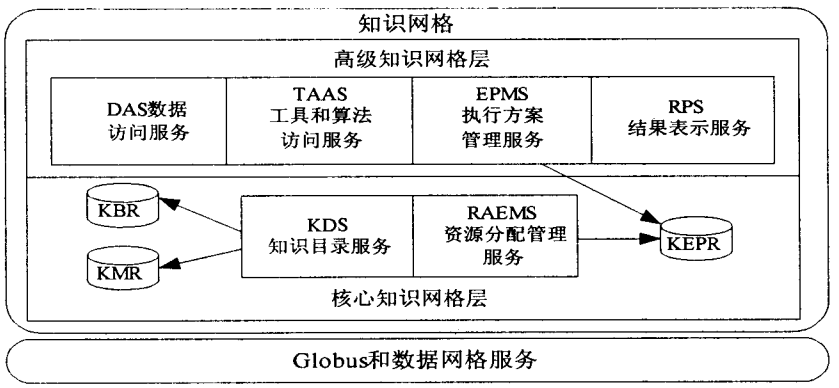


图 1 知识网格体系结构

2.1 核心知识网格层

核心知识网格层提供了网格上分布式知识发现计算的基本服务的定义、组装和执行功能,主要管理描述数据源特性、第三方数据挖掘工具、数据管理和数据可视化的元数据,通过实现应用需求和可用网格资源协调应用的执行。该层主要由两种服务构成:

(1)知识目录服务(KDS)。

该服务扩展了基本的 Globus 元数据目录服务(MDS),负责维护知识网格中数据和工具的描述。由 KDS 管理的元数据有:

①数据源,如数据库,XML 文档和其它的结构化和非结构化数据。

②数据抽取、过滤和操纵的工具和算法。

③分析被挖掘数据的工具和算法。

④数据可视化工具,即用于可视化、存储和操纵挖掘结果的算法和工具。

⑤分布式知识发现执行方案。执行方案是知识网格应用的抽象描述,是描述数据源、数据挖掘工具、可视化工具和结果存储间的数据流和交互的图示。

⑥挖掘过程结果获取的知识,如学习模型和发现模式等。

由 XML 文档表示的元数据信息存放在知识元数据库(KMR)中。描述了能用于挖掘的不同数据源的特征,如位置、格式、可用性等。

要维护从一个特定数据仓库中挖掘出来的数据是不切实际的,但是维护一个已发现知识的数据库是非常有用的[5]。这些信息被存放在知识仓库(KBR)中,但是描述它们的元数据仍由 KDS 管理。KDS 不仅可用于搜索和访问原始数据,也可以发现原先已发现的知识,以便在数据改变时比较给定挖掘计算的输出,或者以递增的方式应用数据挖掘工具。

对于知识网格来说,数据管理、分析和可视化工具通常都是事先存在的。然而,为了使它们对知识发现计算是可用的,相关的元数据就必须存放在 KMR 中。同时,元数据对于数据源的使用来说是非常有用的。另一个重要的库是知识执行方案仓库(KEPR),用于存放数据挖掘过程的执行方案。

(2)资源分配和执行管理服务(RAEMS)。

该服务用于在执行方案和可用资源间查找最佳映射,以满足应用需求(如计算能力、存储能力、主存、数据库、网络带宽和延迟)和网格约束。在执行方案激活之前,该层管理和协调应用的执行。该层并不是使用 KDS 和 Globus MDS 服务,而是直接基于 Globus GRAM 服务的。每个数据挖掘程序的资源请求都是用资源规格描述语言(RSL)描述的。执行方案的分析和处理将产生全局资源请求,并为本地 GRAM 将全局资源请求转换成本地 RSL 请求。

2.2 高级知识网格层

高级知识网格层包括组建、实现和执行并行分布式知识发现计算的服务。另外,该层提供了存储和分析已发现知识的服务。主要的服务包括:

(1)数据访问服务(DAS)。

数据访问服务负责搜索、选择、抽取、转换和交付被挖掘的数据。搜索和选择服务是基于核心知识目录服务的。在用户需求和约束的基础上,数据访问服务自动进行查询和查找由数据挖掘工具分析的数据源。

被挖掘的数据的抽取、转换和交付是基于 Globus 的全局访问二级存贮器服务(GASS)和知识目录服务(KDS)的。当有用数据被发现后,数据挖掘工具可以进行转换操作,也可按照用户请求或安全约束在数据抽取之前进行过滤。这些操作通常都是在选择了数据挖掘工具之后进行的。抽取功能可嵌入到数据挖掘程序中,或者编码存放到 KDS 能够访问的数据库中。

(2)工具和算法访问服务(TAAS)。

该服务负责数据挖掘工具和算法的搜索、选择和

下载。描述其可用性、位置和配置的元数据存放在KMR中,并由KDS管理,而算法和工具则存放在每个知识网格结点的本地存储系统中。需要向其他用户导出数据挖掘工具的结点首先必须使用KDS服务来发布该工具。还有其他的相关元数据,如参数、数据输入/输出格式、实现的数据挖掘算法、资源请求和约束等。

(3)执行方案管理服务(EPMS)。

执行方案是描述数据源、抽取工具、数据挖掘工具、可视化工具和KBR中的知识结果之间的数据流和交互的图形化表示。最简单的情况是,用户可使用可视化构造工具直接描述一个执行方案。然而,由于DAS和TAAS产生结果的多样性、数据和工具的位置、中间结果表示方法等的差异能产生多种不同的执行方案。因此,EPMS是由用户自行选择数据和程序的半自动化的工具,产生一系列满足用户、数据和算法需求及约束的多种可执行方案。

执行方案存放在KEPR中,它允许反复地进行知识发现。如:随着时间变化的相同数据源的阶段性分析。同样,相同的执行方案也可用于分析不同的数据集。另外,不同的执行方案也可用于并行地分析相同的数据集,从不同角度来比较结果(性能、精确性等)。

(4)知识表示服务(RPS)。

知识可视化是数据挖掘过程中的重要步骤,它可以帮助用户解释发现的模式。该服务指出了如何产生、表示和可视化抽取的知识模型(关联规则、聚类模型、分类等)。结果元数据存放在由KDS管理的KMR中。KDS不仅用于搜索和访问原始数据,还可查找已经发现的知识。

3　基于知识网格的数据挖掘

3.1　知识网格的数据挖掘过程

知识网格的数据挖掘[6]过程如图2所示。

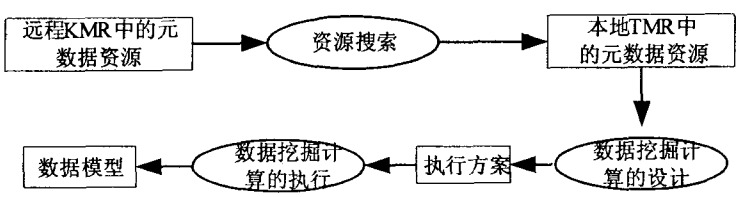


图2　知识网格的数据挖掘过程的设计和执行步骤

(1)资源搜索。由DAS(数据访问服务)和TAAS(工具和算法访问服务)工具搜索网格中各结点的KMR中的XML元数据文档,将满足搜索条件的元数据存放到本地任务元数据库(TMR)中。

(2)数据挖掘计算的设计。由执行方案管理服务(EPMS)工具执行,为数据挖掘计算生成由XML文档表示的可执行方案,并存入知识执行方案仓库(KEPR)。

(3)数据挖掘计算的执行。由资源分配和执行管理(RAEMS)服务执行数据挖掘计算,并将执行后产生的结果存放到知识仓库(KBR)中。用户可以利用结果表示(RPS)工具使执行结果可视化并对其进行分析。

3.2　网格应用可视化环境VEGA

VEGA[7](Visual Environment for Grid Applications)是用于在基于Globus的网格中实现分布式计算组合和执行的知识网格工具。该工具集允许用户从可用的远程资源集合(如计算结点、数据源)开始建立计算。通过网格的信息服务来定位和选取这些资源以对象集的形式呈现给用户,用户使用可视化机制来组合这些对象以生成一个数据挖掘计算的图形化表示。VEGA就将该图形化表示转换成一个可执行方案,通过Globus资源管理工具在网格上处理和执行。它使知识网格用户可以简单有效地开发和执行分布式数据挖掘。如图3所示。

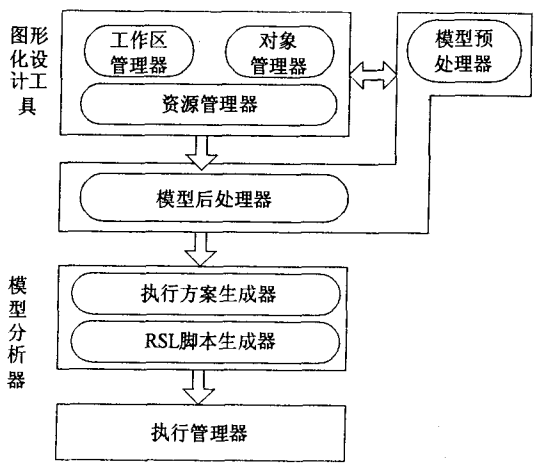


图3　VEGA软件模型

VEGA包括一系列工具,允许执行以下操作:

(1)任务组建:如包含在计算中实体的定义和它们直接关系的规则说明。

用户可以利用一系列图形对象表示各种资源,如数据集、数据挖掘工具和网格结点。由工作站管理器、资源管理器和对象管理器等软件组件构成。工作站负责组织设计环境、管理图形表示的内部模型。资源管理器允许用户浏览任务元数据仓库(TMR),搜索选择用于计算的资源。对象管理器在可视化组件实现阶段处理图形对象,包括数据、软件和主机。

(2)任务一致性检查。

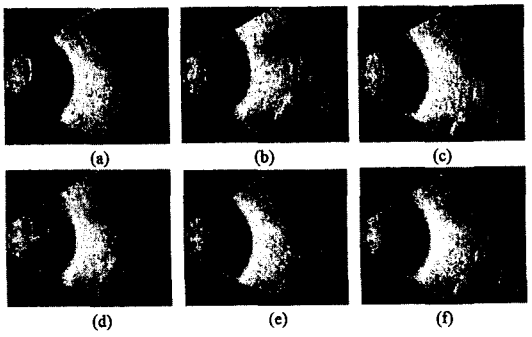
一致性检查是为了获得正确一致的计算模型,由两个组件实现:模型预处理器和模型后处理器。

(3)为任务产生执行方案。

(下转第104页)

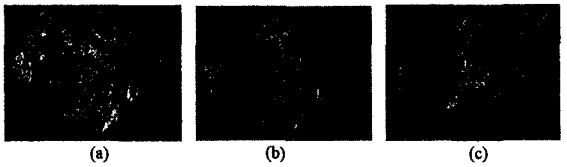
前两者小。

图 2 是三种算法 1/2 处的插值图像与实际图像的误差图像。可见,改进的对应点匹配的灰度插值算法的实验结果要优于线性插值和传统对应点匹配的实验结果。



(a) 原始图片 1 (b) 原始图片 3
(c) 实际位于原始图片 1 和 3 之间的图片 2 (d) 1/2 处直接线性插值图像
(e) 1/2 处传统对应点匹配插值 (f) 1/2 处改进的对应点匹配插值

图 1 眼球图像不同灰度插值结果示意图



(a) 1/2 处直接线性插值误差图像 (b) 1/2 处传统对应点匹配插值误差图像
(c) 1/2 处改进的对应点匹配插值图像

图 2 误差图像

3 结束语

改进的基于对应点匹配的图像插值算法在原有对应点匹配思想的基础上,有效地减少了由于不同组织区域匹配点插值所导致的边界模糊重叠和形态失真,保证了图像插值的质量,插值结果可较好地运用于眼球的三维重建。并且,本算法在 PC 机上易于实现,是一种实用、有效的图像插值算法。

参考文献:

[1] Grevera G J, Udupa J K. An objective comparison of 3 - D image - interpolation methods[J]. IEEE Trans Med Imaging, 1998,17(4):642 - 652.

[2] Chuang K S, Chen C Y, Yuan L J. Shaped - based gray - level image interpolation[J]. Phus Med Bio, 1999,44(6): 1565 - 1577.

[3] 邹鹏程,尹学松. 基于断层图像分割的三维匹配插值[J]. 计算机工程与应用,2004(24):80 - 82.

[4] 屈景怡,史浩山. 一种基于轮廓形状的胸部医学图像插值方法[J]. 计算机工程与应用,2005(11):218 - 220.

[5] 陈灵娜,陈增科. CT 断层图像匹配插值算法的研究与实现[J]. 南华大学学报,2005,19(1):73 - 75.

[6] Penney G P, Schnabel J A, Rueckert D. Registration - Based Interpolation[J]. IEEE Transactions on Medical Imaging, 2004,23(7):922 - 926.

(上接第 101 页)

该阶段由两个软件模型完成。XML 生成器将计算模型转换成基本的由 XML 文档表示的执行方案,然后由 RSL 生成器将 XML 文档转换成 RSL 脚本。

(4)通过资源分配管理执行产生的执行方案。

执行管理器通过 Globus GSI 服务为网格用户认证授权,并将 RSL 脚本递交给 Globus 网格资源分配管理服务(GRAM)执行。执行管理器在整个数据挖掘计算的生命周期中还起到作业监视的功能。

4 结 论

网格架构发展迅速,所支持的程序的种类日益多样化,可使用的工具也日趋完善和复杂。网格服务的发展方向已从原先的基本的面向计算的服务转到高级信息管理和知识发现服务上来。知识网格系统为分布式数据挖掘和基于网格服务的知识发现定义了一个集成的体系结构。该体系结构推动了地理位置分布的大规模数据集的数据挖掘。利用 VEGA 图形化辅助工具,可以方便地实现分布在网格各结点中的数据挖掘和知识抽取,为科学和工商业提供潜在而非常有价值的信息,促进经济和科技的迅猛发展。

参考文献:

[1] 陈 平,王 柏,徐六通,等. 数据挖掘网格的关键技术与挑战研究[J]. 电信科学,2006(22):52 - 56.

[2] 都志辉,陈 渝,刘 鹏. 网格计算[M]. 北京:清华大学出版社,2002.

[3] Cannataro M, Talia D, Trunfio P. KNOWLEDGE GRID: High Performance Knowledge Discovery on the Grid[C]// Proceedings of the Second International Workshop on Grid Computing. [s.l.]:[s.n.],2001:38 - 50.

[4] Cannataro M, Talia D, Trunfio P. Distributed data mining on the grid[J]. Future Generation Computer Systems, 2002, 18 (8):1101 - 1112.

[5] 魏定国,彭 宏. 基于知识网格的数据挖掘[J]. 计算机科学,2006,33(6):210 - 213.

[6] Cannataro M, Talia D, Trunfio P. Design of distributed data mining applications on the Knowledge grid[C]//In: Proceedings National Science Foundation Workshop on Next Generation Data Mining. Baltimore:[s.n.],2002.

[7] Cannataro M, Congiusta A, Talia D, et al. A Data Mining Toolset for Distributed High - performance Platforms[C]// Proc. Conf. Data Mining 2002. Bologna, Italy: Wessex Inst. Press,2002.