网格环境下多数据源的协同查询

全卫新 赵正德 刘宜宁 张 伟
(上海大学计算机科学与工程学院,上海 200072)

摘 要 伴随网格技术的日益发展,作为对数据应用需求的快速回应,在数据网格之后又出现了网格数据库的概念,其中查询又是数据库应用中一个使用频繁的核心功能,由于每个节点上的数据库可能只包含所需信息的一部分,而且查询操作还涉及到数据库异构性、通信开销等问题,这些都给网格环境下的数据库查询性能带来了新的困难。为了提高网格环境下数据库查询的性能,提出了一种运用协同(CSCW)机制来协调网格用户的查询操作的方法,这不仅体现了系统的协同性,而且降低了数据传输开销,还提高了查询效率。

关键词 计算机支持的协同工作 协同查询 网格技术 数据库
中图法分类号: TP39 **文献标识码**: A **文章编号**: 1006-8961(2006)11-1539-04

Collaborative Query of Muti-datasource on Grid

QUAN Wei-xin, ZHAO Zheng-de, LIU Yi-ning, ZHANG Wei
(*School of Computer Engineering and Science, Shanghai University, Shanghai* 200072)

Abstract With the development of the grid technology, for quick response to the requirement of data application, the concept of data grid has appeared after computing grid. The query function is used very frequently in the database application, the same to the grid environment. It appears frequently that composing the database which distribute in various computer nodes to make a virtual database dynamically to finish a query because every database only includes a part of the information. This paper detailed research the question that how to finish the query and shorten the response time under the grid of heterogeneity of the nodes and the widely various communication speed between the nodes. It has important significance for the further research and development of grid database.

Keywords computer supported cooperative work(CSCW), collaborative query, grid, database

1 引 言

目前对网格环境下数据资源的研究和开发工作主要侧重于文件系统,另外,网格环境下的数据源都是在原有数据库的基础上进行研究发展的,然而现有的这些数据库系统的许多应用则侧重于数据存储、访问、组织、授权、重组等方面[1]。由于网格环境的异构性,致使数据库查询技术具有一定局限性,若完全使用原来基于传统计算环境的数据库查询技术,则不能高效地执行数据查询,为此本文提出了一种基于协同技术的查询方法。

2 相关技术

2.1 协同技术

计算机支持的协同工作(computer supported cooperative work, CSCW)是 Greif 和 Cashman 于 1984 年提出的概念,作为一门新兴的学科,它具有多学科交叉的特征,计算机协同工作[2]将计算机技术、网络通信技术、多媒体技术以及各种社会科学紧密地结合起来向人们提供了一种全新的工作环境和交流方式。CSCW 的关键是协调,而信息共享则是协同工作的基本任务,CSCW 系统的目的是支持多

收稿日期:2006-05-25;**改回日期**:2006-08-05

第一作者简介:全卫新(1983 ~),男。2005 年获上海大学计算机工程与科学学院学士学位,现为上海大学硕士研究生。主要研究方向为 CSCW、分布式、网格技术等。E-mail:quanweixin@163.com

个用户参与同一工作，共同协调与协作来完成一项任务[3]。总的来说，CSCW 可以定义为地域分散的一个群体借助计算机及其网络技术，其可通过共同协调与协作来完成一项任务。

另外，网格技术还支持和促进了协同工作模式及协同技术的发展。

2.2 网格技术

网格技术主要研究在分布、异构、自治的网格资源环境来动态建构虚拟组织，并实现跨组织的资源共享与协同工作，共享与协作是网格的基本理念[4]。由于 Internet 实现了计算机系统与网络设施的互联，且 Web 实现了网页的联通，从而使得信息的共享与获取不受时空限制，而网格则试图实现在全球 Internet 范围按需共享与整合各种 IT 资源。

3 协同查询的分析

3.1 网格环境下数据源的特点

在网格的分布的、异构的环境下，网格数据库具有以下特点：

（1）能够接受和容纳多个异构数据库的系统，其对外呈现出一种集成结构，而对内又允许各个异构数据库的“自治性”；

（2）网格环境下的数据源是对已有数据源的网格化，并不存在一个统一的数据库管理系统软件；

（3）数据访问具有并发性，且数据关系复杂；

（4）由于异构情况在前而集成要求在后，因此宜采用自下而上的数据集成方法；

（5）具有很强的动态性，致使各个网格节点上的数据源可以方便地加入或退出网格环境。

3.2 协同查询的目标

网格的主要目标是支持在共享资源上的协同工作，由于网格环境的异构性及动态性，不仅给不同节点间的数据传输造成了很大的网络延迟，并严重影响着系统的查询效率，因此如何减少数据传输量，加快查询响应速度就显得尤为重要了。为此本文研究了利用协同技术来减少数据传输量的查询方法。协同查询的目标主要是：

（1）有效利用前期用户的查询结果，以减少数据源的交互操作和降低数据传输成本；

（2）协调网格用户间的访问操作，使用户有效利用数据资源，以提高查询效率。

4 协同查询设计

系统为了实现用户对系统的透明访问，采用了 Web 服务技术向用户屏蔽了系统的具体实现细节，因此协同查询模块没有向用户显式地呈现各节点间的具体协作查询过程，而是在用 Web 技术将数据源提供的具体网格服务 GDS（grid database service）进行集成实例化时隐式地嵌入进来了。

用户间的查询协调主要是通过对表 recordTable 的控制来实现的，且隐式地嵌入到了系统中。recordTable 表中的内容包括：用户提交的查询访问服务名（Web 服务）、命令参数和查询后的结果在系统服务器中的存放位置，该位置是指系统服务器存放结果数据的表名。

4.1 用户相同查询的协同处理

相同查询是指查询内容（select 内容）相同，查询条件（from 子句和 where 子句）也相同的情况。网格环境下，由于用户数量巨大，因此对同一系统的很多查询是完全相同。具体过程如图 1 所示，由于使用与用户相同的协同处理方法最大程度上利用了用户间的协作关系，并利用了历史用户的查询结果，因此能很大程度大提高查询效率，减少通信开销。

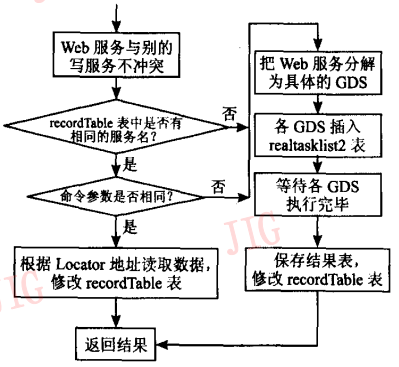

图 1 用户相同查询的协同处理

Fig. 1 Collaborative query process

4.2 用户相似查询的协同处理

在网格环境下，网格用户对同一系统的查询访问有很多是相似的。这里的相似查询是指用户间的查询条件（from 子句和 where 子句）相同，而查询内容则有包含或交集的情况。若能利用前期用户的访问结果，就可以很好地提高查询效率。查询的具体

过程如图 2 所示。

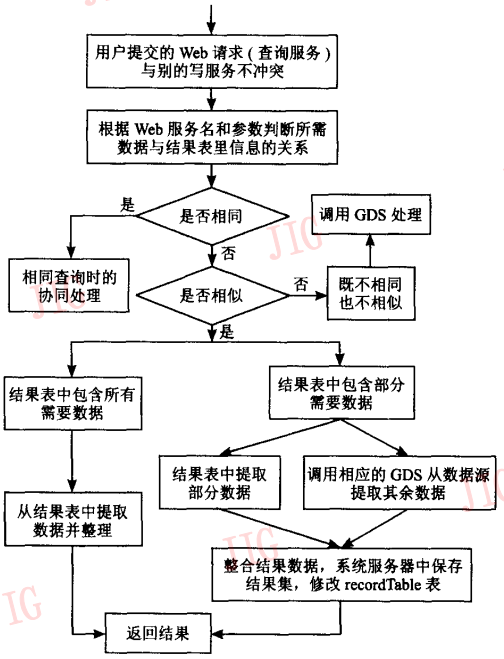

图 2　用户相似查询的协同处理

Fig. 2　Similar collaborative query process

4.3　查询结果的保存与管理

4.3.1　查询结果的保存

如果某次查询调用了数据源提供的网格服务，那么为了方便后来用户就需要将本次数据用 DBMS (database manager system) 来处理验证数据的有效性,以便于数据的完整性一致性管理及多用户访问时的并发控制。若采用文件形式保存,则对数据的管理工作就需要很多管理员手工完成,这将是一项很大的工程。作者为了实现方便把结果数据保存在了系统服务器中数据库的结果表里。

4.3.2　结果信息管理

随着用户查询数据的增加,系统服务器中结果的数量也相应增加,当增加到一定程度时就给服务器带来了负担,影响系统的性能,因此必须对结果信息进行管理,即可以对结果表的数量设定了一个上限值,当超过上限值时就根据 recordTable 表的 count 字段值和 time 字段值,选取最近很少访问的结果表将其删除,再删除 recordTable 表中的相关信息。

4.4　写操作对协同查询的影响

由于网格环境的动态性增加了操作实现的复杂性,因此上面所说的系统服务器中存放的查询结果不可能是一直有效的,本文实现了一个超市连锁信息系统,比如节点 A 处的用户在提交了查询北京超市存放的货物信息后,系统服务器中就保存了相关的查询结果。若节点 B 处的用户要查询北京超市存放的货物信息,就可以直接利用保存在系统服务器中的结果了。然而当 A 用户执行完查询后,节点 C 处的用户(比如销售人员)卖掉了部分货物,这时北京超市实际存放的货物就比保存在系统服务器中的结果要少,若用户 B 直接利用系统服务器中的结果,则显然是错误的。如果发生这种情况,就意味着系统服务器中的相应信息无效。为了不让用户读取错误数据,有以下两种解决办法:

(1)当修改(写操作)相关数据时,同时修改系统服务器中的信息。这种方法可保证系统服务器中的内容一直是有效的,用户查询数据时只需在系统服务器中读取前期用户的结果即可;但是当数据频繁修改时,系统服务器中的内容也必须不断改动,而且随着用户读取的数据的增加,系统服务器中保存的信息也越多,这就不可避免对其存储与维护造成一定的影响。

(2) 当修改(写操作)相关数据时,即通知系统服务器,删除相关的结果信息和 recordTable 表中的相关记录。这种方法虽可以减少系统服务器结果表的数量,但是由于当修改很少时,下一个查询必须到数据源读取数据,因此查询费用比上一种方案高。

本文中综合比较了上述两种方案,实际应用中可根据问题应用域的数据是否具有易变性等特点来选择一种或组合使用。

5　系统实例

本文设计了一个超市连锁店信息系统,该系统首先利用 GT4(GlobusToolkit 4)在 Eclipse 平台开发网格基础服务,便于数据源的扩展;然后采用 Web Service 技术和 Tomcat 服务器来为终端用户开发可访问的服务,该系统体现了网格的集成性、共享性和协同性等特征,完全实现了网格环境下的相同和相似的协同查询。

6　结　论

本文首先分析了网格环境下数据源的特点,然

后确定了协同查询的目标,提出用户查询的协同处理方式,讨论了结果数据的保存和管理技术,通过利用用户的访问结果来实现用户之间的协同访问,并讨论了写操作对协同查询的影响,提出了两种可选的解决方案来保证结果数据的有效性。

参考文献(References)

1 Zhou Long-xiang. Multi-levels distributed dbms technology [M]. Beijing: Science Press, 1998-07. [周龙骧. 多层分布式数据库管理系统实现技术[M]. 北京: 科学出版社,1998-07.]

2 Shi Mei-lin, Yang Guang-xin. Computer supported cooperative work: Past, present and future[A]. In: Proceedings of the First Country-wide CSCW Conference[C], Beijing: 1998: 1 ~ 8. [史美林,杨光信. 计算机支持的协同工作:过去,现在和未来[A]. 见:第一次全国 CSCW 会议论文集[C], 北京,1998: 1 ~ 8.]

3 Douglis F, Foster I. The grid grows up [J]. IEEE Internet Computing, 2003, **7**(4): 24 ~ 26.

4 Yan Jun-yong. The Essence and Framework software evolution in grids[J]. Computer Applications and Software, 2005, **22**(10): 73 ~ 75. [严隽永. 网格的实质及其架构软件的发展[J]. 计算机软件与应用, 2005,**22**(10):73 ~ 75.]