一种基于星型拓扑的分布式挖掘系统的设计与实现

Design and Implementation of a Distributed Mining System Based on Star Topology

陆培伦　胡书卫　（上海大学 自动化系 上海 200072）

摘　要：数据挖掘技术与分布式对象技术、Web 技术的融合是数据挖掘技术未来发展的方向。本文提出了一种基于星型拓扑结构分布式挖掘系统，它采用 ORB 技术进行分布式处理，利用多线程技术实现多用户挖掘。该系统具有跨平台，可移植及可扩展的优点，极大地提高了分布式数据挖掘的效率。

关键词：数据挖掘 分布式计算 ORB 中间件 星型拓扑

1　引言

数据挖掘技术与分布式对象技术、Web 技术的融合是数据挖掘技术未来发展的方向[1]。随着网络技术的发展，数据库的存放越来越趋向于分布式，计算模式也逐渐由传统的 B/S 结构向分布式的方向发展。不管是从网络的传输速度，还是从数据的安全性和保密性考虑，对这些分布式数据库的数据挖掘最好都采用分布式数据挖掘[3]。本文设计的一种基于星型通信模式的分布式数据挖掘系统，该系统采用 ORB 技术，能实现分布式网络环境下与异地异构平台上的分布式站点通信，完成分布式数据挖掘的任务。该系统具有很强的平台适应能力及可移植能力，可以实现对多个异构的并行分布式数据源协同挖掘，极大地提高了分布式挖掘的效率。

2　相关技术介绍

2.1 数据挖掘技术

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中抽取隐含在其中的，人们事先不知的，但又是潜在有用的信息和知识的过程，并且它是一个反复的过程[5]。抽取的知识可以表示为概念（concept）、规则（rules）、模式（patterns）等形式。挖掘的原始数据可以是结构化、半结构化（文本、图形、图像数据），甚至是分布在网络上的异构型数据。

2.2 CORBA 技术

CORBA 由对象管理组（OMG）定义，是分布式异构环境下实现可重用、可移植、可交互的面向对象软件系统的规范[3]。CORBA 由四部分组成：对象请求代理（ORB），CORBA 服务（CORBA service），CORBA 设备（CORBA facilities），以及应用对象（Application object）。ORB 是 CORBA 的核心，是建立在各个对象之间客户/服务器关系上的中间件。通过 CORBA，客户可以透明地调用本地或远程服务器对象。在此情况下，ORB 负责呼叫对象，通过它的参数、方式和返回结果等手段实现请求。在对象确定中，客户不必知道所处的硬软件环境，不必知道对象在网络中的位置。另外，客户/服务器模式仅利用了两个对象之间的交换，而 ORB 上的对象还能作用于另外的客户和服务器。

图 1 给出了 ORB 及其接口结构。接口定义语言（IDL）是 CORBA 规范中定义的一种中性语言。它可以完整地描述客户所需接口的全部信息。IDL 定义的接口经过 IDL 编译器编译后产生客户的桩（Stub）及执行对象的构架（skeleton）这类能与 ORB 通信的接口。

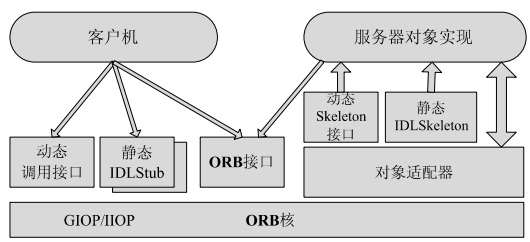

图 1　CORBA2.0 ORB 结构图

上述结构中，ORB 就是连接各个对象的软总线，它使得位于同台计算机或分布在网络各结点的对象之间

实现互操作,而屏蔽了这些对象所在库的结构和平台的差异。

3　系统星型拓扑结构的设计

数据挖掘有很多的研究方向,关联规则挖掘是其中最活跃的研究方向之一,而频繁项目集的求解是关联规则求解中最重要的步骤,随着分布式数据库的发展,如何求解分布式数据库中的关联规则已成为了人们研究的热点之一。

传统的分布式数据挖掘系统大都采用“从站 – 从站”的环形通信模式,为了统计某个候选项目集的全局支持和计数,每个站点都要向其它所有站点发送轮询请求,虽然每个站点的通信负荷比较均衡,但每个站点的通信量都较大,从而容易导致整体网络开销较大。为了进一步减少站点间的通信量,节约网络通信开销,我们将提出另一种方案,即“从站 – 服务器总站”的星型通信模式(如图 2 所示)。

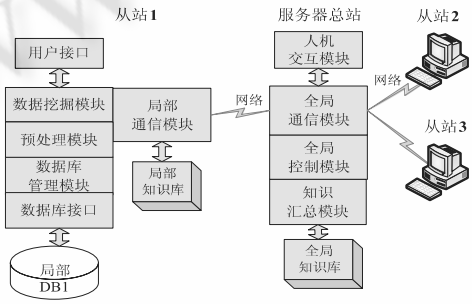

图 2　挖掘系统星型网络模型

该系统主要由多个分布式站点和应用服务器总站组成。从站主要由用户接口,数据挖掘模块,数据预处理模块,数据库管理模块,通信模块及局部知识库组成;服务器总站主要由人机交互模块,全局通信模块,全局控制模块,知识汇总模块及全局知识库组成。

在该方案中引入了局部知识库及全局知识库的存储结构,由于采用了“从站 – 总站”的星型通信模式,各个分布式站点基于各个局部数据库运用本地挖掘模块生成局部频繁项目集及支持数,此过程各从站可异步进行。应用服务器负责对所有分布站点上求出的局部频繁项目集及支持数进行汇总,进而求出各局部频繁项目集的超集,统计其中每个项目集的全局支持合计数,并最终确定全局频繁项目集。最后根据求出的全局频繁项目集和用户给定的最小可信度得出基于全局数据库的关联规则,并把它存储到全局知识库中,从而完成整个系统挖掘的过程。

由此可见,环形的通信模式虽然可以使各个站点间的通信负载达到均衡,但是各站点间的网络通信开销较大,并且各站点间要求同步运行,系统运行效率较差;而星型模式可以显著的减少站点间的通信负载,而且各站点可实现完全异步运行,从而大大提高了系统挖掘的效率,也降低了网络的通信开销。

4　系统实现与挖掘步骤

该分布式挖掘系统采用 C/S 模式,从站及服务器端通过运行 Java 语言编写的功能组件来完成挖掘任务。同时,它基于 CORBA 和 Java 体系结构,按功能要求主要划分为四层结构:数据层、事务层、通信代理层、应用服务器层等(如图 3 所示)。

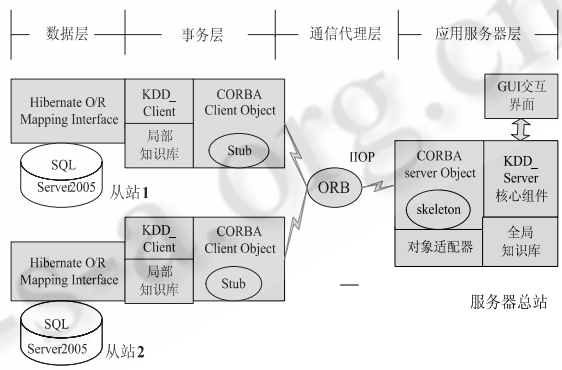

图 3　系统的四层结构图

(1) 数据层

数据层主要包括关系型数据库及中间件 Hibernate。数据库负责存放各个从站挖掘所需的数据源,中间件 Hibernate 是一个面相 Java 环境的对象/关系型数据库映射机制,通过它可以方便地对数据库进行访问控制,省去了很多以往繁杂的 SQL 语句的困扰,一切对数据库的操作如同普通的面向对象编程一样,真正对数据库的操作都由 Hibernate 引擎隐式完成,处理过程被封装在 Hibernate 组件中,我们可以不用关心内部的处理过程,所要了解的仅仅是构成 Hibernate 的主要模块及其使用接口,以便我们把主要精力集中在事

务层的业务处理上。

(2)事务层

事务层为分布在不同位置的从站上处理的所有业务逻辑。最核心的部分就是 KDD_Client 挖掘组件,它采用 JAVA 语言编写,负责生成局部频繁项目集及支持数;CORBA 组件通过远程代理 ORB 负责向服务器端发送和接受请求及数据,完成局部频繁项目集信息的传递及汇总工作,此过程各站点可异步进行。

(3)通信代理层

该层主要实现各分布式站点与应用服务器端的无缝连接。通信代理中间件由对象请求代理 ORB 实现,提供从站对应用服务器对象请求的透明通信,对象请求代理中间件 ORB 使得不同的对象间可轻松地实现交互,而不必考虑两者通讯机制,地理位置,语言实现,操作系统的差异。通信协议采用 IIOP 技术,IIOP 是一种实现互操作性的协议,它使得由不同语言编写的分布式程序可在 Internet 中实现彼此的交流沟通。

(4)应用服务器层应用服务器层包括 CORBA 的服务器对象,对象适配器,KDD_Server 数据挖掘核心组件,GUI 交互界面及全局知识库组成。a)服务器对象由遵循 CORBA 规范的服务器请求对象组成,为各从站的数据请求提供代理服务;b)对象适配器是对象实现访问 ORB 提供服务的主要方法,这些服务包括对象引用的生成和解释,方法的调用,交互的安全,对象和实现的激活与释放等;c)KDD_Server 为服务器端运行的核心数据挖掘组件,采用 Java 语言编写,通过继承 Thread 类而实现了多线程技术,通过创建多线程,可以同时满足多个客户的挖掘请求,相当于每个客户的服务请求分别对应一个独立的服务器线程,从而可以并行地为客户进行处理,也可以把一个复杂的任务分成不同部分运行在多个线程上,实现了系统的并行化;d)GUI 交互界面是人机交互的接口,它主要负责接收客户的请求并完成挖掘结果的显示。

该系统的挖掘步骤如下:

1) 各局部站点通过 IDL STUB 及 ORB 核心向远程服务器端的对象适配器 OA 发出连接请求,对象适配器接收到从站请求后,查找请求对象在服务器上 IDL SKELETON,完成两地的连接绑定。

2) 当用户在应用服务器端通过 GUI 界面向系统发出挖掘请求后,挖掘组件提取出挖掘参数(最小值支持度及其它参数),并向各分布式站点传递挖掘请求及参数。

3) 各站点接收到请求后调用挖掘组件,通过运行当地的数据挖掘算法生成局部频繁项目集及支持数,并将挖掘结果存储在局部知识库中,并通过 CORBA 向服务器总站传送,此过程各站点间可异步进行。

4) 应用服务器总站将收到的所有局部频繁项目集合并汇总,生成全局候选频繁项目集,并再次通过 CORBA 向各局部站点广播。

5) 局部站点根据服务器总站广播的数据确定本地新增项目集(在全局候选项目集中但不在本地频繁项目集)及支持数,修改局部知识库,并将结果返回服务器端。

6) 服务器端接收各分布式站点传送的新增项目集及支持度,并进行累加,求出全局候选频繁项目集的全局支持合计数。

7) 应用服务器根据用户传递的最小支持度筛选出最终的全局频繁项目集。

根据在全局站点求得的全局频繁项目集及最小可信度,通过调用核心挖掘组件 KDD_Server 生成全局关联规则,并把它保存在全局知识库中,从而完成关联规则的整个求解过程。

5 结束语

本文设计的这种基于 ORB 技术的分布式挖掘系统采用星型通信模式,能实现分布式网络环境下与异地异构平台上的分布式站点通信,完成分布式数据挖掘的任务。该系统与以往常见的挖掘系统相比有以下特点:

1)主站与从站间采用星型拓扑结构

这种"从站 – 总站"的星型通信模式减少了各局部站点间的通信负载,实现了各站点间的完全异步,减少了网络开销和提高了系统挖掘的效率。

2)采用中间件技术简化了系统的设计

对象请求代理中间件 ORB 使得不同的对象间可轻松地实现交互,而不必考虑两者通讯机制,地理位置,语言实现及操作系统的差异;数据访问中间件 Hibernate 简化了数据库的访问机制,也使得系统具有了良好的可扩充性,易管理性,和高可用性等优点。

3)采用 Java 语言编写的程序实(下转第 49 页)

(上接第 8 页)

现了跨平台特性

Java 技术的跨平台特性使得用 Java 编写的服务程序具有“一处编写,到处运行”的特性,正好可以满足不同分布站点运行环境存在差异的要求,同时 Java 编写的程序也能满足挖掘结果实时显示的要求。

4)采用多线程技术实现了系统的并行性

应用服务器可以创建多个线程来同时处理多用户的挖掘请求,也可以把一个复杂的任务分成不同部分运行在多个线程上,实现了系统的并行化。

5)采用 CORBA 技术使系统具有良好的可扩展性

由于采用了对象请求代理体系结构,当有新的分布站点要加入时,无需改动和重新编译应用服务器程序 KDD_Server,只需让新的对象在启动时向 KDD_Server 注册即可,使系统便于扩展。

参考文献

1 Kosala R, Blockeel H. Web mining research: A survey. ACM SIGKDD Explorations, 2000, 2(1): 1-15.

2 David Hand, Heikki Mannila. Principles of Data Mining. Massachusetts Institute of Technology, 2001.

3 汪芸. CORBA 技术及其应用. 东南大学出版社, 1999.

4 李雄文等. 三种三层 Web 体系结构的特点与比较. 计算机应用研究, 2001(8).

5 朱玉全, 杨鹤标, 孙蕾. 数据挖掘技术. 南京: 东南大学出版社, 2006.