基于 Internet 的分布式文档管理技术

赵文耘 勉玉静
(复旦大学计算机系软件工程实验室 上海 200433)

摘 要 网上合作研究中心的出现,对文档动态管理及发布提出新的要求。本文提出了一种基于 Internet 的分布式文档管理的体系结构,通过对多种分布式数据复制技术的分析,给出了基于邮件系统的数据复制技术,并给出了软件科学与技术网上合作研究中心的应用实例。

关键词 Internet 分布式文档管理 数据复制

INTERNET BASED DISTRIBUTED DOCUMENTS MANAGEMENT

Zhao Wenyun Mian Yujing
(Department of Computer, Fudan University, Shanghai 200433)

Abstract With the development of web based cooperation research, dynamic documents management and publish become important. We describe an architecture of Internet based distributed document management. After analyzing some kinds of data replication technology, this paper describes the mail based data replication technology and then gives an instance about SSCC.

Keywords Internet Distributed documents Replication

1 背 景

1.1 网上合作研究的出现

全球高速网络的建立,为信息资源的共享和基于网络的联合研究提供了实现前提。及时充分有效地利用 Internet 不断增长的研究信息资源,已成为科学技术研究与开发的重要手段。科研也逐步由以地域为基础向以因特网为基础的方向发展。为此,国家教育部提出了网上合作研究的思路,现已成立并验收了十多个诸如生物、软件科学等的网上合作研究中心。顾名思义,网上合作研究就是采用当前的因特网技术,若干个对某领域有相当基础的大学、研究所或企业,进行紧密研究与项目实施,实现资源共享与整合,这种新颖的研究方式正成为一种新的科研模式。

1.2 建立网上合作研究中心的主要工作及可能存在的问题

网上合作研究,除了应对外进行信息发布及宣传外,还应做到将多个异地的网上合作中心联成一个整体,以一致的形势及手段对外。此外,应可采用多种手段实现网上的合作研究及讨论。由于当前 Internet 联结的速率安全等不尽如人意,各网上合作中心均采用成员单位建立各自网站的形式。因此在当前条件下实现分布式网站的整合就显得尤为重要。

文档是科研最重要的资源之一。各单位积累的学术文档各有侧重,为便于对文档信息浏览、检索,达到资源共享,通常将它们集中在一台文件服务器上进行存储管理。要求成员单位将本地文档及其索引信息通过 Internet 上传至远程服务器。这些文档大小从几百 K 到数十兆不等,在目前国内网络条件下,远程上传容易丢失数据,很不可靠;而多用户同时下载文档时对网络和服务器的要求较高,会出现速度过慢甚至连接中断的现象。给网上合作研究带来不便。

网上讨论是网上合作主要的交流方式。通常的网上讨论仅限于文字形式,而在合作研究中,如能借助适当文档加以说明,将更有助于参与者对讨论内容的理解,提高讨论的深入程度,使讨论更为有效。因此希望能够方便的将文档作为附件上传,供参与讨论的人们浏览下载。如果这些文档都存储在一台机器上,与文档资源共享相同的问题:大量文档在 Internet 上传输的可靠性,将成为实现的最大障碍。

1.3 分布式文档管理的提出

可见集中式的存储管理模式有很大弊端,特别是在国内,跨地区访问还经常受到网络环境的制约。因此变集中为分布的管理模式被自然的提出。如果建立分布式数据库存储文档索引信息,将文档资源分布到各成员单位,各节点用户可就近访问,上传下载,这样既能减轻中心服务器压力,提高效率,又能提高网络传输的可靠性。因此,基于 Internet 的分布式文档管理模式是网上合作研究的一种更为有效的形式。

2 基于 Internet 的分布式文档管理的实现技术

2.1 体系结构

本文给出一种简单可行的基于 Internet 的分布式文档管理模式,其要点为:

1) 将文档或动态的网上讨论片断分别存放在各相关成员单位;

2) 为各类动态信息简历一致的文档索引;

收稿日期:2003-04-06。赵文耘,教授,主研领域:软件工程,CASE。

3) 动态更新索引, 据索引进行动态访问。如图 1 所示。

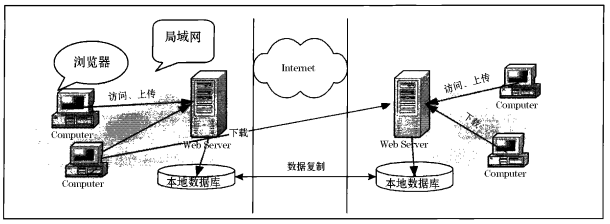

图 1 基于 Internet 的分布式文档管理的体系结构

由图 1 可知: 各节点(成员单位) 开发自己本地的 web 应用程序, 提供文档上传、下载功能, 存储文档体; 建立本地数据库, 存储上传至本地的文档索引信息。另一方面, 通过一定的复制手段, 定时将本地文档数据的变化量复制到其他节点的数据库中, 从而使各数据库中都存有所有节点的文档信息, 供本地网站访问者浏览和检索。这样每个节点只维护数据库的一个横向切片, 可以保证在做数据复制时不会出现更新冲突。当网站访问者请求下载文档全文时, 再根据该文档的索引信息, 连接到文档存储站点, 读取文件, 传递给访问者。这样既实现了文档的分布式存储, 缓解了集中式存储带给服务器的压力, 也减少了文档在网络上的传输量。显然, 实现分布式文档管理的关键是如何实现各站点文档索引数据的同步, 即分布式数据库的数据复制。下文对各种实现技术进行分析。

2.2 分布式数据复制实现技术分析

2.2.1 DBMS 提供的数据复制技术

很多大型数据库管理系统, 如: Microsoft SQL Server, Oracle 等自身都提供了数据复制技术。一般分为快照复制模式、被复制方主动的"订阅时复制"模式以及被复制方被动的"发布式复制"等三种模式。但它们又都有其自身的局限性。

1) 不能有效支持异构数据库环境。如 Oracle 的数据复制技术不能向 SQL Server 中同步数据;

2) 这些复制技术本身都不考虑网络的连接问题。比如, SQL Server 的数据复制要求在一个按照 windowsNT/2000 的标准组建的局域网内部进行, 才能保证复制过程中的可靠性与安全性。而 ORACLE 的复制技术也要求参与复制的各节点之间的网络连接是可靠稳定的。

因此, 对于我们讨论的基于 Internet 的远程数据复制, 使用 DBMS 自带的数据复制技术是不够理想的。

2.2.2 C/S 结构的应用程序解决方法

另一种方法是开发 C/S 结构的复制应用程序, 由客户端程序访问数据库服务器。类似订阅与发布模式, C/S 结构的复制应用程序也分为两种复制模式, 一种是被复制方 push(推) 模式, 一种是复制方 pull(拉) 模式。与 DBMS 的数据复制技术相比, 这两种方式在两个数据库之间加一层应用程序, 对数据格式进行转换, 可以解决异构数据库环境问题, 同时能够做到基于 Internet 的远程数据库访问和操作, 实现分布式的数据复制, 但仍有一定的缺陷:

1) push 模式中, 数据库安全性受到损害。分节点的数据库要为远端的应用程序开放访问权限, 甚至要允许远端程序对本地数据库进行更新操作, 因此, 数据库用户名及口令需要在 Internet 上的传送, 从安全角度考虑很不可靠。

2) pull 模式中, 传送数据的安全性不够。应用程序只有读取远程数据库的权限, 即使用户名口令被截获也不会对数据库本身产生影响, 但数据在网络传输中没有任何的加密措施, 因此从保密性角度看仍然不够安全。

3) 两种方案均可能影响事务完整性。因为网络连接建立在传输层协议 TCP/IP 之上, 网络传输的可靠性并不强, 跨越 Internet 的网络传输因各种外部原因很不稳定, 时常会发生客户端应用程序与远程数据库服务器之间的连接突然中断的现象。为减少网络传输量, 尽量不采用快照复制的方法, 只能每次传输数据的增量部分, 为此, 必须在本地数据库中记载已经完成的更新, 因此不论是 push 还是 pull, 事实上都需要分布式事务。不可靠的网络会严重影响事务的一致性和完整性。

4) 以上分析的两种技术还都有一个无法克服的缺点: 无法穿越防火墙的阻挡。因为 TCP/IP 协议要求进行点对点的连接。而事实上, 因为安全与保密的缘故, 数据库不允许直接暴露在 Internet 中。如校园网内部的数据库服务器不拥有实地址, 校园网也不开放外部对数据库访问的端口, 数据库完全隐藏在局域网之中。在这样的网络条件下, 通过数据复制实现数据同步就必须设法绕过防火墙, 基于更高层的网络协议进行数据传输。

2.2.3 基于电子邮件传输的数据同步技术

通过以上分析看到: 基于 Internet 的分布式数据库的复制技术必须解决以下几个问题:

1) 保证数据传输的稳定和可靠性。

2) 既能穿越防火墙的阻挡, 又能保证数据库的安全性与数据本身的保密性。

3) 最好能够支持异构数据库之间的数据复制。

4) 可采用压缩及加密技术对传输数据进行处理。

我们提出了基于电子邮件传输协议的数据复制模式, 能够较好的解决以下问题。如图 2 所示。

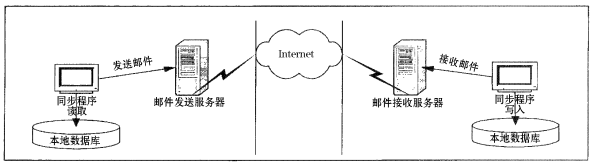

图 2 基于邮件传输协议的数据复制

在这种模式中, C/S 模式的推拉方式被结合在一起: 各节点的同步程序均要完成两个功能。一方面, 自动收集本节点数据库中的变化数据, 压缩或加密后借助邮件服务器将数据发送至其他站点, 另一方面, 应用程序主动收取邮件, 经解压或解密后根据邮件内容获得其它站点的变化数据, 修改本地站点数据库中的索引数据。

与传统技术相比, 这种解决方案有以下优势:

1) 网络传输的可靠性方面:

® 基于应用层的网络协议, 不再基于点到点的连接, 从而绕过了防火墙的阻碍。真正实现了基于 Internet 的分布式管理。

® 与 TCP/IP 协议相比, SMTP(邮件接收协议), POP3(邮件发送协议) 协议可以被认为是更为可靠的。而且在目前的网络条件下, 高质量且可靠的邮件服务器并不难得到。

2) 数据复制过程的安全性方面:

® 各站点的应用程序只能够访问本地数据库, 对本地数据做操作, 从而避免了数据库用户名口令被截获的可能。提高了数据库本身的安全性。

® POP3, SMTP 协议都要求身份认证过程, 进一步提高了安全性。

(下转第 114 页)

间的关系,用类图来描述 DIM。还可运用 UML 中包(package)的机制,将许多类集合成一个高内聚、低耦合的类的集合,以表示模块或库,用包图来显示类的包以及这些包之间的依赖关系、继承关系和组成关系。

从领域信息模型所反映出来的 OOA 活动有:

® 划分主题,建立主题图

将具有较强联系的类组织在一起;

® 建立详细说明

为每个类建立一个类描述模板。

3 相关问题

领域模型是对领域知识的一种描述形式,它来源于对领域十分熟悉的领域工程师及领域专家,同时领域模型也是对领域问题空间解的一种图形化表示形式。通过领域模型可以看到领域内许多相关系统的一些共性,因而有利于在这些共性基础上开发其它的特定应用系统。通过本文在领域模型如何进行系统分析的讨论,从中可以看出,从领域模型所反映出来的分析活动与一般的 OOA 过程是相符合的,并且这些分析活动以及分析结果是比较直接的,同时使用 UML 对领域模型进行描述,有利于对领域对象及其关系的形式化表示,同时对领域结构和领域行为有了一个较规范的描述。此外,有关领域分析、领域工程、领域设计、领域实现等方面都有待于作进一步深入研究。

参 考 文 献

[1] Ronald J. Norman, Object-Oriented Systems Analysis and Design, 清华大学出版社, 1996, James.

[2] Rumbaugh, Ivar Jacobson, Grady Booch, UML 参考手册, 机械工业出版社, 2001. 1.

[3] 邵维忠、杨芙清, 面向对象的系统分析, 清华大学出版社, 1998. 12.

[4] 邹咸林,“领域分析过程框架及 UML 描述”[J],《计算机应用与软件》, 2002. 19(12): 14~16.

[5] 邹咸林等,“面向对象的软件重用成熟度模型”[J],《计算机科学》, 2001. 28(2): 38~39.

(上接第 22 页)

® 复制的数据如需保密,既可使用邮件服务商提供的加密协议,也可使用自行定义的加密协议。

3) 功能与性能方面:

® 传输的数据可采用与 DBMS 无关的格式,如符合 XML 规范的文档。只要各站点应用程序能够解析出所需数据,对本地数据库做相应操作即可。从而解决了异构数据库之间的数据复制问题。

® 每次只传送增量数据,传输量小。网络传输可靠,分布式事务的一致性和完整性也能得到保证。

® 费用低无需额外的网络和软硬件投入,数据复制同步完全自动化,无需人为干预

当然,基于邮件服务的数据复制方式也有自身的不足:对邮件的处理,由于多种原因会存在延时现象,从数据更新请求到完成更新需要一定的时间。但是对于实时性要求并不高的应用,如网上合作研究中心的文档资源共享来说,则是完全可行的。

3 实例分析: SSCC

作为国家教育部网上合作研究建设项目之一, 2001 年,以北大、南大、复旦、北航、大连理工等五校在软件科学与技术领域已有的合作为基础,发挥中心成员单位在软件以及网络技术中各自的特长,建立软件科学与技术网上合作研究中心(SSCC),开展网上合作研究。

软件合作中心采用建立中心网站与各院校分网站的形式,由中心网站发布实时信息,提供交流平台;分网站实现内部的资源共享,各网站间通过 Internet 互连以达到合作研究的目的。各单位上传近百兆的文档,而从各分站点均能浏览到这些文档的索引信息。而另一方面,两所学校的数据库隐藏在学校局域网中,其他院校无法访问;复旦和北京之间的网络连接很不可靠,时常发生从北京无法访问复旦站点的情况。

在这样比较恶劣的网络条件下,我们采用了基于 Internet 的分布式文档管理,借助基于邮件服务的数据复制技术,取得了很好的效果:

1) 因为就近上传,大量文档的上传在局域网内完成,与集中式的文档管理相比, Internet 上的数据传输量大大减少;下载的总数据量虽然没有变化,但原本一台服务器的压力被分在五台服务器上,减小了网络的压力,节省了带宽,提高了下载的速度和可靠性。

2) 每篇文档的描述信息不过几百个字节,向其他四个站点复制的总数据量也不过几 k,通过采用压缩技术,使得传输量更小;使用本校的邮件服务器,安全可靠,没有发生过邮件传输失败的问题。确保只需就近访问任一网站就能浏览到所有站点文档的索引信息。如:标题,作者,摘要、文件大小,上传日期等。

3) 在安全角度:各站点只需其他站点的 pop3 服务器和邮件地址,没有任何敏感信息在网络上传送。

实践证明,我们的基于邮件服务的数据复制技术和基于 Internet 的分布式文档管理方案是有效可靠的。

4 结 语

随着网上合作研究的不断深入,将有更多的文档资源需要得到共享,基于 Internet 的文档管理将越来越重要。我们提出了基于 Internet 的分布式文档管理的方案,并借助邮件服务实现了关键技术:分布式数据库数据复制技术。在保证数据传输的稳定和可靠性,数据库的安全性与数据本身的保密性的基础上,穿越了防火墙的阻挡,并且支持异构数据库之间数据复制。

在今后的工作中,还需要对基于 Internet 的文档管理模式做更多的探讨。设法提高数据复制的时效性,以适用于对实时性要求较高的应用;简化同步过程,尝试基于应用层网络协议 HTTP, HTTPS, SOAP 的数据复制技术,以 Web Service 的形式进行分布式文档管理,减少对第三方系统如邮件系统的依赖等等。

参 考 文 献

[1] 周龙骧等著, 分布式数据库管理系统实现技术, 科学出版社, 1998 7.

[2] [美] Jeffrey R. Shapiro 著, SQL Server 2000 参考大全, 清华大学出版社, 2002. 6. 1.

[3] [美] Kevin Loney/Marlene Theriault 著, Oracle8i 数据库管理员手册, 机械工业出版社, 2000 年 7 月.

[4] [美] Andrew S. Tanenbaum 著, 熊贵喜等译, 计算机网络(第 3 版), 清华大学出版社, 1998 年 7 月.