基于 Agent 的分布式空间数据挖掘模型及实现*)

甄 彤[1,2] 范艳峰[1]

(河南工业大学计算机科学系 郑州450052)[1] (华中科技大学控制科学与工程系 武汉430074)[2]

摘 要 在分析数据挖掘、空间数据挖掘、Agent 的概念和技术特点的基础上,提出了一种基于代理的分布式空间数据挖掘系统,描述了其实现方法,因为本系统只传送数据挖掘的中间结果,所以大大减少了网络的数据传输量,并加强了数据的安全性和保密性。

关键词 空间数据挖掘,Agent,分布式

The Model and Implementation of Agent-Based Distributed Spatial Data Mining

ZHEN Tong[1,2] FAN Yan-Feng[1]

(Department of Computer Science, Henan University of Technology, Zhengzhou 450052)[1]

(Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074)[2]

Abstract Through analyzing the concepts and technology characters of data mining, spatial data mining, and agent, a agent-based distributed spatial data mining system and its implementation are introduced. As this system only transmited intermedia results, not only data transportation is reduced in the network, but also data security and confidentiality are improved.

Keywords Spatial data mining, Agent, Distribute

随着超大规模数据库的出现,先进的计算机技术,对海量数据的快速访问,以及对这些数据应用精深的统计方法计算的能力的提高和发展,数据挖掘应运而生。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,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。而遥感技术(RS)、地理信息系统(GIS)和全球定位系统(GPS)以及生物学的蛋自质分子结构等的发展则产生了对空间数据挖掘(Spatial Data Mining,简称 SDM),也即空间知识发现(Knowledge Discovery in Spatial Databases,简称 KDSD)的需求,它指从空间数据库中提取隐含的、用户感兴趣的空间和非空间的模式和普遍特征的过程。

Intemet 是巨大的分布式并行信息空间和极具价值的信息源。网络技术的飞速发展和广泛使用,使得各个领域之间的数据交流与共享成为可能,交换信息也更加电子化和海量化。但是因网络所固有的开放性、动态性与异构性,故而从网上得到的数据便是没有经过组织的、多型的,而且分布于世界各地的服务器网站上。更重要的是,可得到的信息服务的类型及可信度正在不断地变化和更新,用来解决问题的信息越来越难收集、筛选、评估和使用。因而,定位信息资源、访问、筛选、集成用来进行数据挖掘的信息以及协调信息检索成了一个关键的任务。面对基于 Internet 上的"信息海洋",我们需要提取有用的、可以指导决策的知识。在这种分布式环境下的空间数据挖掘,与传统的基于单一数据表相比,具有很重要的现实意义。可以充分利用已有的资源,也可以实现并行空间数据挖掘,解决数据挖掘的空间和时间"瓶颈"。基于这种地理分布式、形式异构的信息资源上的空间数据挖掘,需要一套复杂的方法来访问、清理原数据以及对挖掘过程的协调。

1 移动 Agent

软件 Agent 是能自动执行一定功能的软件实体。它代表用户,是用户实现其意图的软件助手。它因用户向它指派工作而起作用。从系统的角度看,软件 Agent 是生存于一个执行环境中的软件对象,由程序代码、持久化的内部状态数据和一系列属性组成。它具有反应能力、自主性、目标驱动、连续性等特点。

移动 Agent 可以通过网络从一台机器移动到另一台机器。移动 Agent 具有自主性的学习能力。移动 Agent 在目标机器上执行一系列的操作后,可以返回主机执行结果或者转到另外一台目标机器上继续执行。移动 Agent 最为明显的两个特征就是自适应性和移动性。

为了支持 Mobile Agent 的运行,系统还应该有支持 Mobile Agent 的服务设施,提供 Agent 运行的环境和服务接口,通过 AcL 互相通信和访问服务设施提供的服务。

采用移动 Agent 技术处理分布式计算,具有以下明显的优点:

(1)对网络的依赖性减少。移动 Agent 可以根据需要动态地迁移到数据源处执行,而不是将数据移往计算,避免了大量原始数据在网络中的流动,节约网络带宽,减少了对网络的依赖性。

(2)具有异步自主执行功能。移动 Agent 被创建之后,被派遣到网络上,独立于其创建进程,异步自主地完成所肩负的任务。移动 Agent 到达目的地工作时可以与网络断开,完成任务之后再与网络相连,用户取回结果。

(3)能与环境进行交互。移动 Agent 能自动地监测、感知环境,环境发生变化时能自动地作出反应。

(4)对异构平台适应。分布式计算环境、硬件和软件都可能是异质的。移动 Agent 独立特定的主机和传输协议,只依赖于它们的执行环境,不受平台的异质所影响。

(5)并行处理能力。对并发执行的任务,可以利用多移动 Agent 技术,创建多个移动 Agent,分别派遣到不同的主机

*)基金项目:河南省自然科学基金项目(0324220027);河南省科技攻关项目(0424220189)。

上，实现并行处理。

(6)状态的持续性。必须保持在不同的地址空间中连续运行，即保持运行的连续性，也就是移动 Agent 在转移到另一节点上运行时的状态必须是在上一节点挂起的那一时刻的状态。

由基于移动 Agent 的运行方式和特点可知，上文所述的分布式数据挖掘的技术难点均可以得到一一解决。

2 空间数据挖掘技术

由于 GIS 数据库是空间数据库的主要类型，并且从 GIS 数据库中发现的知识类型及知识发现方法可以涵盖其它类型的空间数据库，借鉴 DM 和 KDD 技术的成果，针对空间数据的特点，从 GIS 数据库可以发现的主要知识类型有：

(1)普遍的几何知识。普遍的几何知识是指某类目标的数量、大小、形态特征等的普遍的几何特征。计算和统计出空间目标几何特征量的最小值、最大值、均值、方差、众数等，还可统计出特征量的直方图。在足够样本的情况下，直方图数据可转换为先验概率使用。在此基础上，可根据背景知识归纳出高水平的普遍几何知识。

(2)空间分布规律。空间分布规律是指目标在地理空间的分布规律，分成在垂直向、水平向以及垂直向和水平向的联合分布规律。垂直向分布即地物沿高程带的分布，如植被沿高程带分布规律、植被沿坡度坡向分布规律等；水平向分布指地物在平面区域的分布规律，如不同区域农作物的差异、公用设施的城乡差异等；垂直向和水平向的联合分布即不同的区域中地物沿高程分布规律。

(3)空间关联规律。它是指空间目标间相邻、项链、共生、包含等空间关联规则，例如，村落与道路相连，道路与河流的交叉处是桥梁等。

(4)空间聚类规则。空间聚类规则，或空间分类规则，是指特征相近的空间目标聚类成上一级类的规则，可用于 Gls 的空间概括和综合，例如，将距离很近的散布的居民点聚类成居民区。

(5)空间特征规则。它是指某类或几类空间目标的几何和属性的普遍特征，即共性的描述。普遍的几何知识属于空间特征的一类，作用十分重要。由于它在遥感影像解译中，因此分离出来的单独作为一类知识。

(6)空间区分规则。它指两类或多类目标间几何的或属性的不同特征，即可以区分不同目标的特征。

(7)空间演变规则。若 GIS 数据库指时空数据库或 GIS 数据库中存有同一地区多个时间数据的快照(snapshot)，则可以发现空间演变规则。空间演变规则指空间目标依时间的变化规则，即哪些地区易变，哪些地区不易变，哪些目标易变及怎么变，哪些目标固定不变。

3 基于 Agent 的分布式空间数据挖掘模型

基于 Agent 的分布式空间数据挖掘系统，能从不同的数据站点中进行分布式空间数据挖掘。主要由用户接口代理 UIA (user interface agent)、用户信息库、适配器(faciliator)、数据挖掘代理 DMA(data mine agent)和元数据管理 MDM (meta data management)5个部分组成。DMA 在有挖掘请求时由系统动态创建，虽然 DMA 可能和某个特定的操作系统绑定，但 UIA、适配器、系统的定义和通信部分可用 Java 编码，独立于操作系统平台。图1是本系统的结构图。

(1)*用户接口代理* UIA(user interface agent)　UIA 用来实现用户和计算机之间的交互，它用良好的界面来收集用户的需求和输出数据挖掘的结果，使得使用本系统的用户不需要太多的专业知识。UIA 在和用户交互的过程中，会根据用户信息库中的信息对用户的身份和需求进行分析，并把分析结果提交给用户信息库。

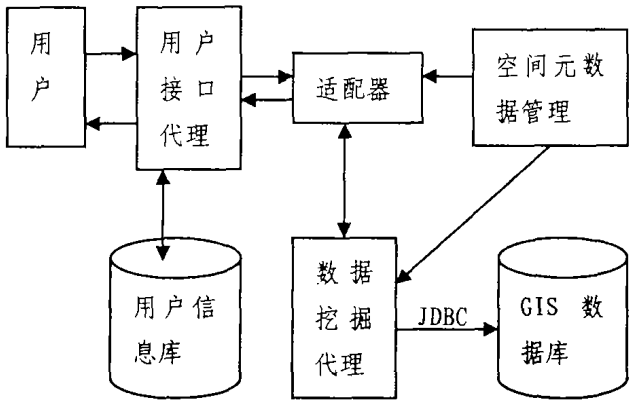

图1　基于 Agent 的分布式空间数据挖掘模型

用户界面全部采用 Web 方式，这样做的主要好处在于用户存取方便、独立于平台、低廉的系统建立和管理耗费。

(2)*用户信息库*　用户信息库中存放着两类信息：一类是用户管理信息；另一类是用户的兴趣和爱好等信息。用户管理信息用来注册、维护和管理用户信息。在系统中每个用户被分配一个帐号，相同属性的用户被分配为一组，还可以进行用户优先级设置、磁盘使用空间限制等。用户的兴趣和爱好等信息用作为 UIA 和用户交互的推理规则。

(3)*适配器*　适配器用来初始化 Agent，实现 Agent 之间的通信，以及和远程数据挖掘系统之间的通信。它保存着各个 Agen 的运行状态、运行位置、当前利用系统资源的情况。每个 Agent 在适配器中都有一个别名，这样 Agent 通信时只需要知道别名就可以了，而无需知道运行位置。

适配器的另一个功能是分解 UIA 传送过来的挖掘请求然后传送给相应的 DMA，在 DMA 完成挖掘过程后，综合 DMA 的挖掘结果再传送给 UIA。

(4)*数据挖掘代理* DMA　DMA 是系统的核心，主要由两部分组成：

① Agent 定义模块：它实现 DMA 的一般定义和状态等，还包括和适配器之间的消息传递；

②数据挖掘模块：它实现本地空间数据库的数据挖掘分析，本模块可用任何语言编码，这样做主要是考虑到可以重用已有的代码和本地代码的计算效率要比 Java 高，但数据挖掘模块和空间数据库的接口全部采用 JDBC。

在文中选择基于分类的空间方法中的归一化(generalized)的方法，原因是当涉及到层次信息时，它能相对比较容易地表示相关信息。例如：叶子层包含不同的数值范围；中间层的值可能是“严重”，“非常高”，“高”，“中等”和“低”。然而最高层可以有两个值，“可接受”或“严重”。

挖掘代理的主要作用是进行空间数据挖掘处理。各个挖掘代理收到请求后，发送一个 DQML 消息给协作代理，然后进行相应的空间数据挖掘。完成挖掘处理之后，结果显示给用户或存储在数据库里。

(5)*空间元数据管理*　元数据(Metadata)，是关于数据的数据(Data About Data)，是关于数据和信息资源的描述性信息。

(下转第110页)

细(如表达整形的数据类型就包括6种),而且还定义了特有的复杂类型,如序列、值类型、对象引用等。由于 XML 本身具有良好的可扩展性,因此通过扩展 XMLSchema 和 SOAP 数据类型,使得每个 IDL 类型唯一地映射为一种 XML 数据类型。产生 WSDL 文档时,分布对象接口中的 IDL 类型声明映射为 WSDL 文档中相应的 XML 类型 Schema,使得 SOAP 客户能够以 XML 数据类型构造 SOAP 请求。目前,OMG 已经制定了这一映射的相关标准[2]。

4 StarWebService 系统实现

StarWebService 是我们实现的基于适配模式的 Web 服务运行支撑环境,它基于 J2EE 体系结构,采用可插拔的灵活设计支持多种后端服务实现。目前,StarWebService 提供了 SOAP 消息的 HTTP 绑定,支持多线程并发地处理 HTTP 请求。其服务容器运行在符合 Servlet2.3规范的 Web 容器之中,主要包括服务管理模块、数据类型转换模块和实现适配模块。StarWebService 全部采用动态机制设计实现适配模块,从而支持在运行时刻动态地部署、修改、去部署各种类型的服务。StarWebService 对当前主流的分布对象技术的 Web 服务化提供了全面的支持,允许使用 CORBA 对象、EJB 等多种后端系统提供 Web 服务的功能。特别地,对于后端系统为 CORBA 的服务而言,StarWebService 遵循了 OMG 的最新标准[2],支持 IDL 语言的全部简单类型及主要复杂类型的编码转换,与国外同类产品功能相当。StarWebService 还包括友好的图形化部署工具及应用开发 API。

总结 本文针对 Web 服务运行支撑环境的构造问题,提出"总线+容器+服务"的体系结构,并着重讨论了基于分布对象的实现适配的关键问题和自主研制的支持分布对象适配的 Web 服务运行支撑环境 StarWebService。下一步工作将重点围绕可管理的 Web 服务容器展开,并在此基础上展开支持网格计算[3]的 Web 服务运行支撑环境的研究。

参 考 文 献

1 W3C Working Draft. Web Services Architecture. http://www.w3.org/,May 2003
2 OMG. CORBA to WSDL/SOAP Interworking Specification. http://www.omg.org, 2003,1
3 Foster I, Kesselman C, Nick J M, Tuecke S. Grid Services for Distributed Systems Integration. IEEE Computer, 2002,35(6)
4 Sun Microsystems Inc. Jav™2 Platform Enterprise JavaBeans™ Specification, v2.1, Final Draft, 2002
5 OMG, CORBA Component Model Specification, 2002,6
6 Gamma E, Helm R, Johnson R, Vlissides J. Design Patterns: Elements of Resusable Object-Oriented software, 1994,10

(上接第97页)

空间元数据(GeoMetadata),是关于地理相关数据和信息资源的描述性信息。它通过对地理空间数据的内容、质量、条件、位置和其他特征进行描述与说明,帮助和促进人们有效地定位、评价、比较、获取和使用地理相关数据。对空间数据某一特征的描述,称为一个空间元数据项。空间元数据是一个由若干复杂或简单的元数据项组成的集合。如果说地理空间数据是对地理空间实体的一个抽象映射,那么,可以认为,空间元数据是对地理空间数据的一个抽象映射。从这个意义上来说,空间元数据和地理空间数据是对地理空间实体不同抽象层次的描述,是对地理信息的不同深度的表达,它们统一于它们所反映的客观内容,同时说明数据集的主要内容、属性的统计特性和属性在数据挖掘算法中的角色,是空间数据挖掘的基础。空间元数据管理包括元数据的产生和维护等。

4 系统实现

Agent 之间的通信实际上是通过 Faciliator 完成的,Faciliator 概念来自对中间件技术的推广与发展。Agent 之间的通信实际上首先是与同一结点上的 Faciliator 通信。Facihator 负责将消息分解,然后发往不同结点上的 Faciliator,再由本地的 Faciliator 传送给相应的 Agent。

在本系统中,数据挖掘代理 DMA 负责存取空间数据和从空间数据中挖掘高级别的用户信息。DMA 以并行方式工作,DMA 之间通过适配器来通信和共享信息。适配器协同代理,提供信息给用户,并反馈用户信息给代理。系统的基本工作原理如下:

(1)用户(或远程数据挖掘系统)发出挖掘请求;

(2)UIA 接受挖掘请求,并把挖掘请求按照预定格式打包后转发给 Faciliator;

(3)Faciliator 对挖掘请求进行分析,确定需要涉及到的 DMA;

(4)Faciliator 检查 DMA 状态,如果 DMA 没有运行,就创建 DMA;

(5)Faciliator 把挖掘请求广播给 DMA;

(6)DMA 根据挖掘请求自动挖掘出相应的信息;

(7)Faciliator 从各个 DMA 中收集相应的信息,然后进行综合分析,得出最终的结果信息;

(8)把结果信息提交给用户(或远程数据挖掘系统)。

在空间数据挖掘的过程中,用户可以通过消息通信机制对 Agent 进行监控或终止。如果用户的挖掘请求不能在本地完成,则可以根据远程空间数据库的信息,由 Faciliator 把挖掘请求转给远程数据挖掘系统的用户接口代理,直到远程数据挖掘系统挖掘出结果信息后再把结果信息传回。

结束语 网络的开放性、动态性与异构性以及信息不断变化更新的特点,是空间数据挖掘所面临的一个挑战。本文就此问题提出了基于 Agent 的解决方案并进行了实现。进一步的工作中需要实现更多的挖掘算法并集成到该系统以及选择更多的数据进行实验。

参 考 文 献

1 Kotz D, Gray R S. Mobile Agents and the Future of the Internet. ACM Operating Systems Review, 1999, 33(3): 7~13
2 Lange D B, Oshima M. Seven Good Reasons for Mobile Agents. Communications of the ACM, 1999, 42(3): 88~89
3 周海燕,王家耀,等. 空间数据挖掘技术及其应用[J]. 测绘通报, 2002, 37(2):11~13
4 邸凯昌,李德仁,等. 空间数据挖掘和知识发现的框架[J]. 武汉测绘科技大学学报,1997,22(4):328~333
5 Agrawal R, Imeielinski T, Swami A. Mining association rules between sets of items in large databases[C]. Processing of ACM SIGMOD, May 1993. 207~216
6 王大玲, 于戈, 鲍玉斌, 等. 基于概念层次树的数据挖掘算法的研究与实现[J]. 计算机科学, 2001, 28(6):88~91
7 Han Jiawei, Micheline K. Data Mining: Concepts and Techniques [M]. Morgan Kaufmann Publishers, 2000
8 Han J, Koperski K, Stefanovic N. GeoMiner: A system prototype for spatial data mining[M]. In: Proc. ACM SIGMOD Int. Conf. on Management of Data, Tucson, Arizona, 1997. 560~563
9 Koperski K, Adhihary J, Han J. Spatial Data Mining: Progress and Challenges. SIGMOD'96 Workshop on Research Lssues on data Mining and knowledge Discovery (DMKD'96). Canada: Montreal, 1996