文章编号:1000 - 582X(2003)12 - 0092 - 03

Web Agent 在空间数据挖掘框架中的应用*

文俊浩[1,2],耿玉营[3],吴中福[2],徐 玲[1]
(1. 重庆大学 软件学院,重庆 400044; 2. 重庆大学 计算机学院,重庆 400044;
3. 华融资产管理公司郑州办事处,郑州 450000)

摘 要:空间数据是人们认识自然和改造自然的重要数据,SDMKD 是利用数据挖掘方法从空间数据库抽去知识以及步骤的人机交互工程。网络作为巨大的分布式并行信息空间,在其中进行知识发现更具有意义。本文在分析数据挖掘、空间数据挖掘、Web Agent 的概念和技术特点的基础上,笔者给出了一个利用 Web Agent 在 Internet 上实现空间数据挖掘的基本框架,并进行了实现验证。

关键词:数据挖掘;空间数据挖掘;智能代理
中图分类号:TP391 **文献标识码**:A

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]以及生物学的蛋白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用户感兴趣的空间模式与特征、空间与非空间数据的普遍关系及其他一些隐含在数据库中的普遍的数据特征。[1]

Agent 是指驻留于环境中的实体,它可以解释从环境获得的反映环境中所发生事件的数据,并执行对环境产生影响的行为。[2]

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,逐渐成为巨大的分布式并行信息空间和极具价值的信息源,交换信息也更加电子化和海量化。但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。分布式对象技术(如 CORBA 或 DCOM 技术)使分布且异构的应用程序之间能以一种共同的方式提供和获得服务,实现其在分布状态下的"软集成"。另一方面,从 Internet 上可得到的信息服务的类型及可信度正在不变地变化和更新着。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,利用 Agent 的智能性进行数据挖掘便发挥了其潜在的优势。

1 Agent 技术

Agent 是具有智能的,它应对环境有响应性、自主性和自动性等。它不仅能够作用于自身,而且能够施动作于环境并能够接受环境的信息,重新评估自己的行为;同时,它能够与其它 Agent 协同工作。[3] 图 1 表示了 Agent 的属性。

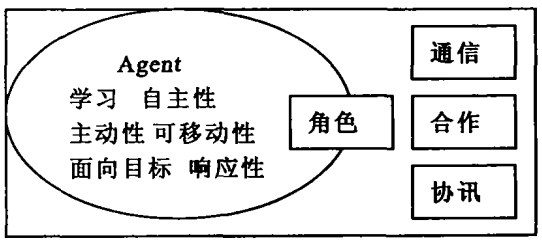

图 1 Agent 的属性

自主性:一个 Agent 在没有与环境的交互作用或来自环境的命令的情况下自主执行任务。响应性:A-

* 收稿日期:2003 - 04 - 18
基金项目:国家十五重大科技计划项目(2002BA107B);建设部科技攻关项目建科(001 - 4 - 70);重庆大学基础及应用基础研究支持项目。
作者简介:文俊浩(1969 -),男,河南临颍人,重庆大学博士研究生,主要研究方向:数据挖掘、信息安全。

gent 必须对环境的影响和信息做出适当的响应。主动性/面向目标：Agent 不仅对环境变化做出适当反应，而且在特定情况下采取主动行为，这种自身采取主动的能力需要 Agent 有严格定义的目标。推理/学习/自适应能力：Agent 的智能主要由 3 个部件来完成，即内部知识库、学习或自适应能力以及基于知识库内容的推理能力。可移动性：一个 Agent 在计算机网络中的漫游能力。角色：Agent 在社会活动中对安全性、风险、信任、诚实等因素的考虑。通信/合作/协调：这是 Agent 在群体中应具有的社会属性。

在本文中，Agent 的基本工作是给各种数字和服务提供一个访问中介，通常用来检索、转化和处理信息。对于大多数的系统，专门地设计了 Agent 通信语言，以便在两个或更多的 Agent 之间通信。KQML（知识查询和操纵语言）便是其中之一。它是交换信息和知识的一种语言和协议。KQML 可被看作由 3 层组成：内容，消息和通信层。内容层详尽描述了信息的实际内容。消息层提供了表示格式并规定包含内容的消息的传送协议。通信层为低层通信参数编码。[4]

2　基于 Web Agent 的空间数据挖掘框架

利用三种类型的 Agent：界面 Agent，协作 Agent 和信息 Agent。界面 Agent 和用户交互，接收用户的具体要求并传递结果。协作 Agent 通过规划问题解决方案，通过查询并和其他软件 Agent 交换信息来执行任务。信息 Agent 提供对异构信息资源收集的智能访问。这些 Agent 被设计来支持查询和空间数据挖掘[5]。

挖掘框架分为两层。前层接受用户查询要求和显示详细的用于挖掘的数据的信息报告。结果可以是挖掘的结果或特殊查询。后层是收集相关数据，为前层转化和准备数据。在系统中，不同 Agent 之间的交互作用如下：

1）预定事件

* 收集 Agent 访问网络资源，定期地收集相关数据。收集 Agent 把原始数据或 HTML 文件送给协作 Agent。

* 协作 Agent 转化从收集 Agent 来的数据，并以 DQML 的格式保存。

* 协作 Agent 把结构化的数据信息传给挖掘 Agent。

* 挖掘 Agent 对数据进行空间数据挖掘。

* 挖掘 Agent 以 DQML 文件的形式把结果传递给 GIS 数据库。

2）请求事件

* 用户输入请求到查询 Agent。

* 查询 Agent 把查询格式化成 DQML 消息，并从数据库中得到相应的结果。

* 如果查询 Agent 不能从 DQML 数据库中得到想要的结果，它把这个要求传输给挖掘 Agent。那么挖掘 Agent 和协作 Agent 合作提取、转化和分析数据。最后，挖掘 Agent 的结果便返回给查询 Agent 用来显示。

收集 Agent 用来进行挖掘的数据可以从各种相关网站上得到，并且这些数据在不同的时间随着不同的数据源不断地变化更新着，更为复杂的是，数据的存储格式各不相同。收集 Agent 用来从相关网站收集各种相关数据，这些数据用来规划数据检索、实际的提取和在 GIS 数据库中的归档；协作 Agent 的主要功能是把没有加工的原始数据转化成可以用来查询和挖掘 Agent 的数据格式。在许多情况下，协作 Agent 需要清洗 HTML 里不需要的信息。更重要的是，未加工数据以不同的频率出现。因此，Agent 也需要重新配置和协调数据以便能得到一致的数据格式，使用 DQML（数据挖掘和构造语言）即可完成；空间挖掘 Agent 的主要作用是进行空间数据挖掘处理。各个挖掘 Agent 收到请求后，发送一个 DQML 消息给协作 Agent，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；查询 Agent 允许用户输入 ad - hoc 请求或检索数据挖掘结果。查询 Agent 以 DQML 的格式传送查询消息给服务器来搜索答案。如果不能在 GIS 数据库里找到答案，查询 Agent 便会激发协作 Agent 去得到数据。然后，协作 Agent 返回数据之后，查询 Agent 执行查询并在网页浏览器上显示结果。查询 Agent 也可能需要一个挖掘结果。类似于上述过程，如果得不到结果的话，查询 Agent 便促使协作 Agent 工作，结果返回给查询 Agent。

3　实现

JAVA 被用来实现挖掘框架。Servlet 被翻译为服务器小程序，主要用于处理服务器和客户机之间的消息传递。Servlet 是一个基于 Java 的 Web 服务方构件，它是由一个可以生成动态 Web 内容的 Servlet 引擎管理的。正如其他基于 Java 的构件一样，Servlet 也是与协议无关、跨平台的，它被翻译为平台无关的字节码，这些字节码可被加载到一个支持 Java 的 Web 服务器上。Servlet 引擎是具有能够提供 Servlet 功能扩展的 Web 服务器。Servlet 通过请求/响应方式与用户端进行交互，Servlet 可以通过动态构造一个发回客户机的

响应来客户机请求。它可以与运行于客户端的 Applet 进行交互,也可以直接与客户端的 HTML 页交互。

Servlet 运行于服务器端,它接受来自客户端的请求,并把处理结果返回给客户端。它作为现有 Internet 技术和 Java 技术的重要中间桥梁,可以用于构成当前网络大型应用中普遍使用的三层服务结构中的 Web 服务器层。三层服务结构现在有着广泛的应用,其中第 1 层是用户服务,第 2 层是业务服务,第 3 层是数据服务。图 2 显示了该三层服务结构。

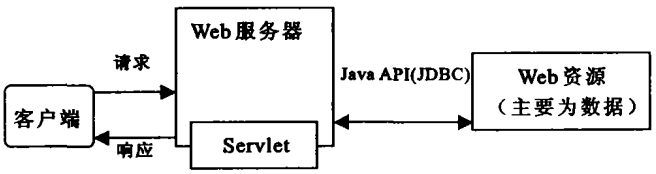

图 2 Web 的三层服务结构模型

与传统的 CGI 和其他类似的技术相比,Servlet 有如下优点:

在传统的 CGI 中,每个请求都要启动一个新的进程;而在 Servlet 中,每一个请求由一个轻量级的 Java 线程处理。Servlet 在服务器上运行一个 Java 虚拟机,在多次调用同一个 Servlet 时,它只需加载一次就可以了,而且加载一个修改后的 Servlet 时不需要重新启动服务器,因此它具有高效性。

Sun 公司为 Servlet 提供了一套标准的 Servlet API,它为请求和响应消息定义了一个标准的接口。许多 Web 服务器支持 Servlet API.

Servlet 是用 Java 编程语言编写的,它具有 Java 语言的所有优点,例如:方便易用和良好的可移植性[6]。

4 结 论

笔者就空间数据挖掘中,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,提出了基于 Web Agent 的解决方案并进行了实现,引入了 Agent 的概念和技术,进一步的工作中还需要将更多的挖掘算法实现并集成到本系统,也还需要选择更多的数据进行实验。随着 Agent 技术的进一步发展,在空间数据挖掘中,Agent 无疑将会发挥更强有力的作用。

参考文献:

[1] 李德仁,王树良,史文中,等. 论空间数据挖掘和知识发现[J]. 武汉大学学报(信息科学版),2001,26(6):491-499.

[2] 张云勇. 移动 Agent 及其应用[M]. 北京:清华大学出版社,2002.

[3] 罗英伟. Agent 及基于空间信息的辅助决策[J]. 计算机辅助设计及图形学学报,2001,13(7):667-671.

[4] CHORAFAS D N. Agent technology handbook[M]. New-York: N. Y. :McGraw-Hill,c1998.

[5] NG V, CHAN S, AU S. Web Agents for Spatial Mining on Air Pollution Meteorology[A]. The Sixth International Conference on Computer Supported Cooperative Work in Design Advance Program[C]. London:Ontario CANADA,2001.

[6] 赵京胜,顾训穰. 基于 Agent 技术的应用框架分析[J]. 计算机工程与应用,2003,22:94-97.

Application of Web Agent in Spatial Mining Framework

WEN Jun-hao[1,2], *GENG Yu-ying*[3], *WU Zhong-fu*[2], *XU Ling*[1]

(1. Faculty of Software Engineering, Chongqing University, Chongqing 400044, China;
2. College of Computer Science, Chongqing University, Chongqing 400044, China;
3. Huarong Capital Administration Co., Zhengzhou 450000, China)

Abstract: Information available from the Web is unorganized, multi-model and constantly changing and being updated. Therefore, information becomes more difficult to collect, filter and evaluate. An agent-based framework to support spatial data mining is described. It makes use of three types of agents: Collect agents, Coordinate agents and Query/Mining agents. And then the implementation model is given.

Key words: data mining; spatial data mining; web agent

(编辑 吕赛英)