

采用移动代理技术建立数据仓库的研究

钟国祥¹ 张为群²

(重庆教育学院学生处 重庆400067)¹ (西南师范大学计算机与信息科学学院 重庆400715)²

Research on Building Data Warehouse by the Means of Mobile Agent Technology

ZHONG Guo-Xiang¹ ZHANG Wei-Qun²

(Student Office, Chongqing Education College, Chongqing 400067)¹

(College of Computer and Information Science, Southwest Normal University, Chongqing 400715)²

Abstract The thesis sets forth for the policy maker a way of achieving the aim of building model data warehouse based on relation model distributed database by using Mobile Agent technology. According to the design principal of data warehouse and the standard technology, author puts forward forming metadatas by tree pattern or forests structure, giving the definition of metadata. Model and the general principal of combining metadatas to form data warehouse and showing a picture tool of data warehouse designing system. It creates a Mobile Agent orignal form that can realize the constrection of data warehouse based on relation model distributed database.

Keywords Mobile agent, Data warehouse

1 问题的提出

随着信息技术的飞速发展,企业更需要一种能支持决策的全新信息系统。如何在分布式数据库上建立数据仓库是决策支持系统开发人员面临的一个普遍性难题。由于移动代理可以在异构网络中自主地从一台主机迁移到另一台主机并与其他代理或资源进行交互,并具有高效处理大量数据的能力,因此本文针对上述问题为决策支持系统开发者在关系型分布式数据库上构造数据仓库提出了一种采用移动代理技术实现的办法。

2 以元数据为核心的数据仓库的构建

2.1 元数据模型

2.1.1 元数据的定义 元数据是关于数据的数据,它是用来描述数据的。元数据:

$MD = (K, N, W, G, T)$

其中,K:键码,是数据所属“事实”的关键标识码;N:名称,是数据元素的名称;W:类型,是数据的数据类型;G:域,是数据的有效值范围;T:时间,是数据的时间属性。

2.1.2 元数据结点的模型 元数据结点表示数据仓库的主题或属性。

主题型元数据结点定义为一个三元组:

$SD := (K, N, T)$

其中,a,K:键码,是结点所表示主题的所在关系表的关键字段名;b,N:名称,也就是结点所表示主题的所在关系表的名称;c,T:时间,是结点所表示的主题的时间属性。

属性型元数据结点定义为一个五元组:

$AD := (K, N, W, G, T)$

其中,a,K:键码,是结点所表示属性的所在关系表的关键字段名;b,N:名称,也就是结点所表示属性在关系表中的字段名称;c,W:域,也就是结点所表示属性的值的宽度;d,G:类型,也就是结点所表示属性的数据类型;e,T:时间,是结点所表示的属性的时间属性。

2.1.3 对于单主题的数据仓库的元数据采用树型结构定义,对于多主题的数据仓库的元数据采用森林结构定义。

2.2 元数据的建立和维护

数据仓库的建立是一个在实践中不断完善的过程,且决策支持系统开发者对分布式数据库也有一个逐渐熟悉的过程,因此决策者根据操作型数据库一分布式数据库建立数据仓库的元数据也应该是逐步推进的。在建立和维护元数据的过程中,我们定义了插入、删除和替换三种操作。在此不再赘述。

2.3 以元数据为核心构造数据仓库的规则

数据仓库作为支持管理决策过程的、面向主题的、集成的、随时间而变的、持久的数据集合,主要存贮大量的历史数据和汇总数据,数据既来源于源数据库,也来源于外部数据。定义了数据的数据一元数据之后,我们根据数据仓库的设计原理和技术标准提出基于元数据构造数据仓库的如下九条规则:

1. 选择规则(SEL)

1)选择主题(SEL subject)。在元数据树或森林中,若指定的元数据结点是主题型元数据结点,则取出该结点元组中的N作为数据仓库一待建关系表的主题。

2)选择属性(SEL attribute)。在元数据树或森林中,若指定的元数据结点都是属性型元数据结点时,则取出这些结点元组中的N作为数据仓库一待建关系表的属性。

2. 分割规则(partition) 把满足条件元组中的N构成的属性值记录从分布式数据库中取出来组建成数据仓库中一新的关系表。

3. 样本规则(sample) 从分布式数据库中取出指定结点元组中的N构成的属性值记录组建成数据仓库中一新的关系表。

4. 清理规则(remove) 当满足条件时,则先清除指定数据仓库关系表的所有记录,然后追加从分布式数据库中取出指定的结点元组中的N构成的属性值记录。

5. 快照规则(snapshot) 当满足条件时,则对指定的数

据仓库关系表追加从分布式数据库取出的指定结点元组中的 N 构成的属性值记录。

6. 计数规则(count) 对满足条件的分布式数据库实例计数,结果作为数据仓库关系表。

7. 求和规则(sum) 对指定的分布式数据库结点元组实例求和,结果作为数据仓库关系表的一属性值。

8. 求极值规则(MM) 对指定的分布式数据库结点元组实例求极值,结果作为数据仓库关系表的一属性值。

9. 求平均规则(average) 对指定的分布式数据库结点元组实例求平均,结果作为数据仓库关系表的一属性值。

3 建立数据仓库的移动代理系统原型的构建

3.1 建立数据仓库移动代理的定义

建立数据仓库移动代理是指帮助决策支持系统开发者从分布式数据库出发建立数据仓库并与决策支持系统开发者交互的移动代理。它根据决策支持系统开发者提供的建立数据仓库的规则在指定的分布式数据库的有关站点中进行信息检索和加工,并把结果追加到数据仓库的关系表中。

3.2 建立数据仓库移动代理系统原型

建立数据仓库移动代理系统原型如图1所示。

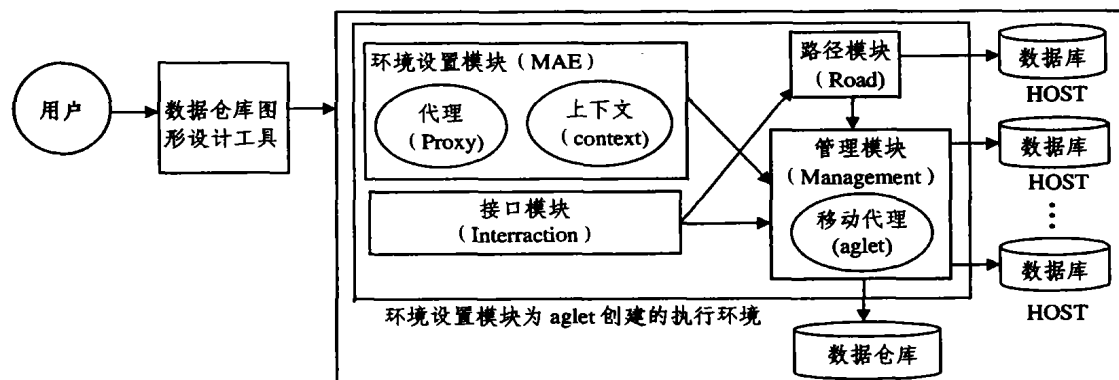


图1

3.2.1 数据仓库设计系统图形工具模型 其模型如图2所示。

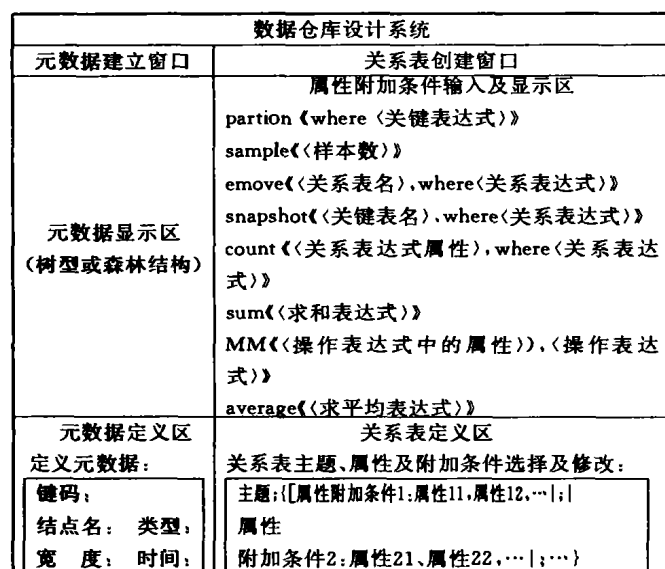


图2

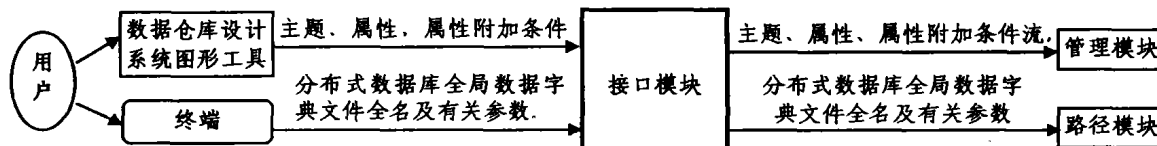


图3

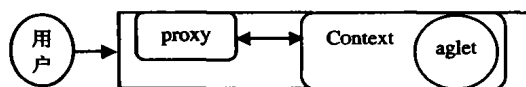


图4

3.2.2 接口模块 如图3所示。其中:主题是数据库一个关系表的名称;属性是数据库关系表或分布式数据库关系表的字段名;属性附加条件是对分布式数据库关系表有关字段进行信息检索和加工的约束,用构造数据库的规则来表示;文件全名由文件所在站点的 Url、文件路径和文件全名构成。

3.2.3 环境设置模块 如图4所示。环境设置模块为 agent 在分布式数据库各站点的工作提供执行环境。该环境实现了移动代理定义中的大部分模型,并提供了与自身相关的支持服务,与自身的建造环境相关的支持服务,访问其他移动代理的支持服务及访问非代理软件环境的开放性支持。

1) Proxy. 是 aglet 的代表。它作为一个屏障,阻止外界对 aglet 公有方法的直接访问。Proxy 还提供 aglet 的位置透明,即它隐藏了 aglet 的真实位置。这意味着一个 aglet 和它的 Proxy 可以分离,使得本地的 Proxy 能够隐藏 aglet 的远程性。

2)Context。是 aglet 的工作场所。它是一个静态对象,提供了方法来维护和管理运行的 aglet,在一致的执行环境中使主机免受恶意 aglet 的攻击。

3.2.4 路径模块 如图5所示。路径模块根据接口模块

提供的主题、属性及属性附加条件流,通过对创建数据库关系表所需求的分布式数据库中的属性在分布式数据库全局数据字典的查询,得到这些属性所在分布式数据库关系表的文件全名,并提供给管理模块作为 `aglet` 移动的路径。我们称通过查询得到的分布式数据库中所有关系表文件全名及相应的

属性为文件全名流。

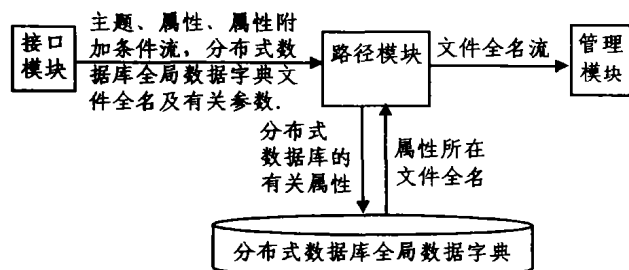


图5

3.2.5 管理模块 如图6所示。管理模块根据接口模块提供的主题、属性及属性附加条件流和路径模块提供的文件全名流,负责创建和发送 aglet 到远程主机,接收局部主机的 aglet 和 aglet 的检索和加工结果,并将检索和加工结果信息装入数据仓库的关系表中。管理模块连续传输 aglet,通过连续传输在目标主机形成它的副本。一收到 aglet,管理模块就重构 aglet 并以它作为参考对象,产生它的执行文本。

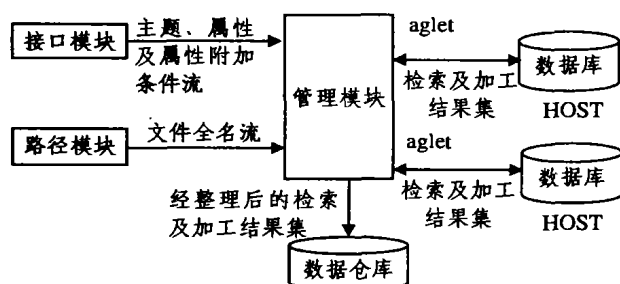


图6

1)aglet,是一个可移动的 Java 对象,它能访问计算机网络中具有 aglet 使能的主机。它根据用户的要求在分布式数据库有关站点完成信息检索和加工,并把结果反馈给管理模块。

它有自主性,因为它抵达某一主机后运行它自身的执行线程;它有反应性,能对到来的消息作出反应。

2)identifier。是 aglet 的标识符,每个 aglet 与一个 identifier 相关联,aglet 的标识符是全球唯一,终身不变的。

4 实验系统设计和结果分析

我们利用 ORACLE8i 采用重构法建立了一个分布式考生信息数据库实验原型。按照定义的格式,通过数据仓库设计系统图形工具定义主题、属性及属性附加条件流,我们一一验证了构造数据仓库的九条规则,并进行了组合测试,建立数据仓库移动代理系统都能正确运行,并能产生准确结果。

结束语 利用数据仓库进行决策支持已成为决策支持系统开发者关注的热点,关键是如何直接为决策支持系统开发者提供一个操作简单、效率较高的基于分布式数据库建立数据仓库的系统工具。本文在研究数据仓库设计概念和理论的基础上,提出了以层次模型构建元数据,提出了以元数据为核心的建立数据仓库的九条规则,并构建了数据仓库设计系统图形工具原型和建立数据仓库移动代理系统原型。通过这个系统创建数据仓库,不仅可以免去开发者去学习一些繁杂的概念、理论,而且还能为开发者节约大量的时间和精力。甚至,有兴趣的决策者自身也可以通过这个系统构建数据仓库。

参考文献

- 1 Lange D B, Oshima M. Programming and Deploying Java Mobile Agents with Aglets. Addison-Wesley, 1998
- 2 <http://www.cs.dartmouth.edu/~agent>
- 3 Aglets Software Development Kit(askd). <http://aglets.trl.ibm.co.jp/>
- 4 邵佩英. 分布式数据库系统及其应用. 科学出版社, 2000
- 5 [美]Giovino W A, 著. 潇湘工作室译. 面向对象数据库设计. 人民邮电出版社, 2000
- 6 袁鹏飞. Oracle 8i 数据库高级应用开发技术. 人民邮电出版社, 2000
- 7 (上接第93页)
- 8 Lempel R, Moran S. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. In: Proc. 9th Intl. World Wide WEB Conf. 2000
- 9 Chakrabarti S, et al. Mining the WEB's link structure. IEEE Computer, Aug. 1999
- 10 Chakrabarti S, et al. Automatic resource compilation by analyzing hyperlink structure and associated text. In: Proc. 7th Intl. WWW Conf. 1998
- 11 Borodin A, et al. Finding Authorities and Hubs From Link Structures on the World Wide WEB. In: Proc. 10th Intl. WWW Conf. 2001
- 12 Cohn D, Chang H. Learning to probabilistically identify authoritative documents. In: Proc 17th Intl. Conf. on Machine Learning, 2000
- 13 Cohn D, Hofmann T. The Missing Link—A Probabilistic Model of Document Content and Hypertext Connectivity. Advances in Neural Information Processing Systems (NIPS), 2000, 13
- 14 Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Addison Wesley, New York, NY, USA, 1999
- 15 Amento B, Terveen L, Hill W. Does "Authority" Mean Quality? Predicting Expert Quality Ratings of WEB Documents. In: 23rd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2000
- 16 Mendelzon A, Rafiei D. What do the Neighbours Think? Computing WEB Page Reputations IEEE Data Engineering Bulletin, 2000, 23(3): 9~16
- 17 韩家炜, 孟小峰, 王静, 李盛思. WEB 挖掘研究. 计算机研究与发展, 2001, 38(4)
- 18 Google Inc. Google search engine. <http://www.Google.com>
- 19 Topic system. <http://www.cs.toronto.edu/db/topic>
- 20 Dempster A, Laird N, Rubin D. Maximun likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1977, 39: 1~38
- 21 IBM Almaden Research Center Clever Project. <http://www.almaden.ibm.com/cs/k53/clever.html>
- 22 Arasu A, et al. PageRank Computation and the Structure of the WEB: Experiments and Algorithms. In: 11th Intl. World Wide WEB Conf. 2002