文章编号:1005-8451(2003)06-0007-03

基于 Web 的数据仓库数据集成问题的探讨

冯茉莉,张 喜

摘 要:基于 Web 的数据仓库数据集成就是将 Web 技术与数据仓库技术有机结合。它将涉及到数据仓库技术,Internet 技术,数据挖掘技术和搜索引擎技术等。从丰富数据仓库数据源的技术角度出发,分析了基于 Web 的数据仓库体系结构,并对数据仓库中基于 Web 的数据集成方法、实施数据集成中存在的问题以及目前可利用的解决方案进行了探讨。

关键词:数据仓库;互联网;网络集成
中图分类号:TP392 文献标识码:A

Study on data integration of Web-based data warehouse

FENG Moli, ZHANG Xi
(School of Traffic and Transportation of Northern Jiaotong University, Beijing 100044)

Abstract: Data integration of Web-based data warehouse was to combine the Web and the data warehouse .It involved data warehouse, Internet, data mining and search engine technologies. In terms of riching data source of data warehouse,it was analysed the architecture of Web-based data warehouse, made a deep study on Web-based data integration approach, questions in data integration implementation and available solution at present.

Keywords: data warehouse; Internet; Web integration

1 引言

数据仓库技术是20世纪90年代以来兴起的一门数据库技术,经过10多年的不断发展,在高效地存储和处理大型数据集合方面已经相当成熟。然而,随着计算机网络通讯技术的迅猛发展和Internet的日益普及,为了充分利用 Web 数据,更好地支持企业的决策,人们开始把关注的重点由提高数据仓库数据处理能力转移到如何丰富数据仓库数据源的问题上来。因此,基于 Web 的数据仓库的数据集成问题就成了该领域亟待研究和解决的问题。文章从丰富数据仓库数据源的技术角度出发,分析了基于 Web 的数据仓库体系结构,并对数据仓库中基于 Web 的数据集成方法、实施数据集成中存在的问题以及目前可利用的解决方案进行了较深入的探讨。

2 数据仓库及数据集成

数据仓库就是面向主题的、集成的、不可更新的(稳定性)以及随时间不断变化(不同时间)的数据集合。它主要是从大量的事务型数据库中提取出有用的数据,经过数据集成,形成统一的决策支持数据库的数据存储格式,存储到数据仓库中,作为公用数据为决策者提供更好的访问支持,用以辅助经营管理的决策制定过程。

由于数据仓库的数据来源可以是企业内部的或是外来的数据,它们通常是由不同的数据系统,不同的操作系统以及应用系统生成,所以,我们必须对各个数据源进行数据集成。所谓数据集成,就是从这些数据源中将有用的数据提取出来,进行净化、整理、综合和概括,去掉没用的数据项,转换成统一的格式加载到数据仓库中。

3 数据仓库的体系结构

数据仓库是一个复杂的工作过程,它从各个独立的系统中搜集所有相关的数据,按照一定的机制对这些数据进行清理,转换,组织,集成为一种统一的数据格式加载到数据仓库中去。

3.1 传统模式下的数据仓库

在传统的模式下,数据仓库使用的数据主要来源于文件和各种信息系统,例如采购系统,生产管理系统,工资系统等。传统的数据仓库系统有2种类型的数据源:操作型系统和外部环境。操作型系统是主要

收稿日期:2002-12-17
作者简介:冯茉莉,在读硕士研究生;张 喜,教授。

的信息来源，而外部数据主要包括新闻，各种统计数据以及政府的公文等等，是可选择的。目前，数据仓库的数据来源主要局限于企业的内部环境。

3.2　基于Web的数据仓库的体系结构

随着网络技术的发展，Internet也在以指数级的速度迅猛发展，尤其是WWW（万维网）已经发展成巨大的分布式信息空间。Web技术更是目前Internet上发展最快的也是最重要的信息检索手段，它提供了一种全球范围内的信息共享方式。只有数据仓库技术和Internet技术相结合才可以为更有效管理企业数据提供潜在解决方案。基于Web的数据仓库体系结构如图1所示，该体系结构分3层，即客户端，Web服务器和应用服务器。

基于Web的数据仓库的优点：（1）方便存取。基于Web的数据仓库提供一种以Web为中心的扩展方案，使得在任何地方只要能与Internet连接的计算机就可以方便地访问公司的数据仓库；（2）与平台无关。Web技术是建立在某种计算机标准上的，它包括用于通讯的TCP/IP，HTTP应用导航，用于显示的HTML。使得用户不管使用哪一种计算机平台，都可以对数据仓库中的商业信息进行访问；（3）采用瘦客户机机制，降低建设和管理成本。

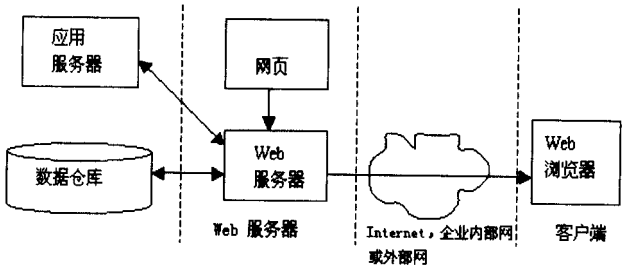

图1　基于Web的数据仓库的体系结构

4　基于Web的数据集成

近年来，随着从Web上可获取信息数量的激增，如何从Web上搜集与企业决策相关的商业信息，如何有效地分析处理这些数据，并且把Web中的数据集成到数据仓库中等问题成为企业亟待解决的问题，下面，把数据仓库技术和网络技术结合起来，从系统分析的观点介绍基于Web的数据集成—网络集成的概念。

4.1 基于Web的数据集成方法—网络集成

网络集成（WI）是一种对Web数据进行筛选，分析，转换，并加载到数据仓库系统中的系统化方法，同时也是管理和充分利用Web数据一种强有力的工具。它主要解决的是如何在快捷的Web上畅游，发现和获取具有商业价值的有用信息，重新组织这些信息到数据仓库中去，以备将来使用。图2给出了网络集成的体系结构。其中，Web　工具包具备自动搜索机制，它根据WI目录中定义的筛选网站的标准和规则，自动地在网上浏览，并且把筛选出来的这些网站上的信息加载到Web　服务器中。WI　目录中定义了筛选网站的标准和规则，它直接影响到一个企业对于Web数据的商业利益和价值倾向。它还继承了网络域名组织的等级结构，同时还包含了Web　工具包每次都需要定期访问的网站列表。它必须适应企业对于商业要求的变化而不断地进行更新。Web　服务器用来存放从目标站点提取来的Web数据，并进行清洗，转换，最后集成到数据仓库系统中去。

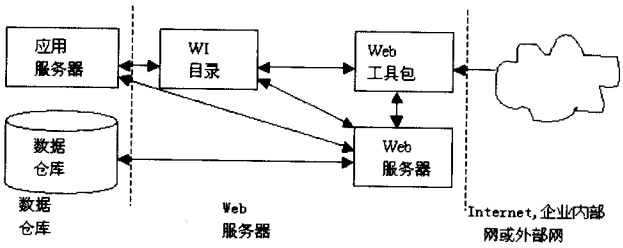

图2　网络集成的体系结构

4.2　网络集成技术

网络集成涉及到数据仓库技术，Internet，数据挖掘技术（特别是多媒体的数据挖掘）和搜索引擎技术。对数据仓库技术而言，确定的数据仓库已经不能满足企业决策支持的需要，人们引入了非确定的数据仓库来处理非结构的多媒体信息。数据挖掘技术（Data　Mining）在近些年来得到不断的发展，多媒体数据库，Internet信息库等数据类型的挖掘也有一定的成就。基于索引的Web搜索引擎，以及Web挖掘，这些都对网络集成的实施提供了技术支持。基于索引的Web搜索引擎可以完成对Web的搜索，对Web页面作索引，建立和存储大量的基于关键字的索引，用于定位包含某些关键字的Web页面。Web挖掘实现对Web存取模式，Web结构和规则，以及动态的Web内容的查找。它分为Web内容挖掘（Web content mining），Web结构挖掘（Web structure mining）和Web使用记录的挖掘（Web usage mining）。基于统计的特征提取技术，Web日志挖掘技术等都为网络集成搜索WWW上文本信息，以及优化网络集成访问站点结

构提供了可利用的解决方案。

5　网络集成需要解决的主要问题及解决方案

5.1　需解决的问题

网络集成的主要任务就是从 Web 中提取有价值的数据。但是，由于 Internet 本身是一个开放的、动态的和异构性的全球分布式网络，并且对这些新技术的集成应用仍然是一个新的挑战，使得人们在实施网络集成、以及在使用 Web 信息的过程当中有许多潜在的问题需要解决。

（1）数据格式多种多样：信息载体的异构性，内容的非结构化），跟踪和检索 Web 信息困难。Internet 是一个多媒体环境，混合了文本，图形，表格，图片，动画，视频和音频。将文本，图像，声音等多媒体数据类型转化成一种统一的数据格式加载到关系数据库表中的同一个数据域中需要先进的技术来完成。

（2）Web 数据不稳定，生存期短，访问的不一致性：Internet 的动态性）。数据仓库一个突出的特点就是它是稳定的数据集合，Web 数据则是易变的并且生存期短。Web 数据另一大特点是访问的不一致性。我们经常遇到"failure to response"情况,而这种情况的原因大部分是由于域名解析出现问题或是没有找到网页。站点的变动导致域名解析出现问题，网页的变动，删除或重新编辑导致找不到该网页在 Internet 上是经常遇到而且是普遍的事情。

（3）Web 数据的质量不高。过多的信息，特别是不相关信息，常常会影响决策支持效果，降低企业整体性能。因此，通过向 Internet 的扩展，网络集成的任务是增强数据仓库决策支持的能力，而不是把数据仓库转化成另一个 Internet。

5.2　解决方案

就目前的技术而言，可以从以下几个方面来解决上面遇到的问题。

（1）采用信息挖掘系统作为对搜索引擎的补充，提高搜索的 Web 数据的精度。Web 搜索引擎的快速定位包含某些关键字的 Web 页面，对网络集成搜索相关信息有一定的帮助。但是由于现有的搜索引擎各自不同的搜索规则，返回的结果也不尽相同。它们只能通过盲目的匹配来处理以关键词形式表示的简单目标，这有 2 个缺陷：对任一范围的话题检索出来的数目过于庞大，相关性也不大；有很多与话题相关的文档因为不包含相应的关键字而检索不到。由于这些缺陷它们无法处理用户给出的样本形式的复杂模糊目标，而信息挖掘系统则能够从样本中提取出目标信息的特征，然后根据目标特征在网络中进行有目的的搜寻；

（2）采用 Web 挖掘技术，不仅对 Web 上非结构化和半结构的文本内容进行挖掘，还对 Web 的链接结构进行挖掘，识别出权威 Web 页面，来优化网络集成的访问结构，帮助网络集成发现高质量的 Web 结构和资源。通过对 Web 日志挖掘，网络集成还可以对其目录中网络数据的变更进行动态的追踪；

（3）采用面向对象数据仓库技术，用对象来代替原来传统的数据类型的数据项，把 Web 中多种数据格式：文本，图形，表格，图片，动画，视频和音频）转换为统一的对象类型加载到数据仓库中；

（4）制定规则，建立质量模型，有效地对搜索到的 Web 数据进行商业价值判断。

6　结束语

文章试图从丰富数据仓库数据源的技术角度出发，通过分析基于 Web 的数据仓库的体系结构，对数据仓库中基于 Web 的数据集成方法—网络集成进行了的探讨，并提出了实施数据集成所面临的主要问题及解决方案。随着 Internet 的飞速发展，Web 将会为企业提供越来越多的有价值的数据资源，所以，对基于 Web 的数据仓库的网络集成方法的研究和对实施数据集成所面临问题的提出与解决，不但对现代数据仓库技术的研究具有理论价值，而且对提高企业科学决策的整体性能也具有现实意义。基于 Web 的数据仓库的网络集成就是数据仓库技术，Internet，Web 挖掘技术与原有数据集成的融合，在自动地数据采集，多媒体的数据挖掘，数据质量的评价模型，以及多种技术的协作等方面都有待我们继续探索。

[参　考　文　献]

[1] Zhenyu Huang,Leida Chen,and Mark N.Frolick. Integrating Web-based Data Into A Data Warehouse[J].Information Systems Management, Winter 2002, 19(1).

[2] Leida Chen, Mark N.Frolick. Web-based Data Warehousing: fundamentals,cha-llenges,and solutions[J].Information Systems Management, 2000, 17(2).