数据仓库与数据发掘的应用

许向东 （北 方 交 通 大 学 100044）
张全寿 （铁道部电子计算中心 100045）

摘要：本文介绍了数据仓库的基本概念、特征以及创建数据仓库的方法及关键问题；论述了数据发掘的目的、一般过程及与此相关的一些技术。最后分析了当前市场上数据仓库和数据发掘工具产品的发展现状。

关键词：数据仓库 数据发掘 数据库 数据

一、数据仓库

Prism Solutions 公司的创始人之一，W. H. Inmon 是世界公认的数据仓库概念的创始人，在他的《建立数据仓库》一书中指出，“数据仓库就是面向主题的、综合的、不同时间的、稳定的数据的集合，用以支持经营管理中的决策制定过程”[2]。根据 Inmon 的解释，数据仓库是一个处理过程，该过程从历史的角度组织和存储数据，并能集成地进行数据分析。简而言之，数据仓库就是一个大的数据库，其中存储了从公司所有业务数据库，如：联机事务处理(OLTP)系统中获取的综合的数据，这些数据可能驻留在许多不同的数据源中。这些数据源可能是文件、层次型数据库(如 IMS)、网络结构化的数据库(如 IDMS)、反向列表数据库(如 DATACOM/DB)，关系型数据库(如 Oracle)，或者由上述系统组成的混合系统。

1. 数据仓库的特征

数据仓库除了具有传统 DBMS 的共享性、完整性、数据独立性外，还具有下列特征：(1)面向主题而集成。传统的数据库是面向应用设计的，它的数据只是为处理具体应用而组织在一起的。应用是客观世界既定的，它对于数据内容的划分未必适用于分析所需。而主题是一个在较高层次将数据归类的标准，每一个主题基本对应一个宏观的分析领域，基于主题组织的数据被划分为各自独立的领域，每个领域有自己的逻辑内涵互不交叉[3]。因此，在数据进入数据仓库之前，必然要经过加工与集成，将原始数据结构做一个从面向应用到面向主题的大转变。

(2)历史性。数据仓库通常存有关于各个主题的不同时间的综合信息，一般为 5～10 年。与此不同的是，业务数据库通常只保存有用事务数据 30～90 天。数据经集成进入数据仓库后是极少或根本不更新的。因此，常用操作是追加操作和历史性查询。

(3)时间属性。数据仓库数据的码键都包含时间项，以标明该数据的历史时期。由时间维和各个主题域一起可以构成多维数据。

2. 建立数据仓库的方法及几个关键问题

数据仓库的建立离不开四个重要的方面：对数据来源的分析；对数据的转化与综合过程的定义；构造数据仓库本身；提供用户赖以从数据仓库中获取所需信息的工具。下面是建立数据仓库的具体步骤。

(1)确定终端用户的需要，为数据仓库中存储的数据建立模型。通过数据模型，可以得到企业完整而清晰的描述信息。数据模型是面向主题建立的，同时又为多个面向应用的数据源的集成提供了统一的标准。数据仓库的数据模型一般包括：企业的各个主题域、主题域之间的联系、描述主题的码和属性组。

(2)深入地分析企业的数据源，记录数据源系统的功能与处理过程。Prism Solutions 公司的顾问 J. D. Welch 指出，设计数据仓库最重要的一步便是要理解商业动作的规律，只有了解数据是如何被处理的，才能分解商业处理过程，从中获取数据元素。

(3)利用现有系统的信息，确定从源数据到数据仓库的数据模型所必须的转化/综合逻辑。这涉及到应该合并转化多少数据；是综合所有的数据文件还是综合发生变化的操作文件；转化/综合过程应该多长时间执行一次等问题。决定数据转化与更新频率是重要的商业事件。无论数据仓库的更新是采用事件驱动还是时间驱动，当某种事件发生时就需要更新数据。

(4)生成元数据。元数据是关于数据的数据，类似于传统数据库的数据字典。描述了数据的转化与综合逻

辑,定义了数据仓库的数据模型。目前,国外有几家第三方厂商的软件产品,如 Carleton 公司的 Passport、Prism Solutions 公司的 Warehouse Manager 可帮助用户完成数据仓库过程中这最重要的一环。

(5)生成物理的数据仓库数据库,并从各种源系统中获取数据装入数据仓库之中。

(6)生成必须的终端用户应用软件,或通过其他方法为终端用户提供查询工具,以便终端用户从数据仓库中获取所需的信息。

在建立数据仓库的过程中,特别要考虑的几个关键问题是:

①数据的粒度(Granularity of Data)。粒度是指数据仓库中数据单元的详细程度和级别。数据越详细,粒度越小,级别就越低;反之,数据综合度越高,粒度越大,级别就越高[3]。粒度的划分将直接影响到数据仓库中的数据量和所适合的查询类型,对数据仓库中其他的设计工作有很大影响。通常需要将数据划分为:详细数据、轻度综合、高度综合三级或更多级粒度。不同粒度级别的数据用于不同类型的分析处理。

划分粒度的方法是:首先,估算数据仓库中数据的行数和所需的 DASD(Direct Access Storage Device)数;其次,由估算出的数据量和 DASD 数,决定如何划分粒度,但需要注意的是,划分粒度的决定性因素并非总的数据量,而是总的行数。这是因为对数据的存取通常是通过存取索引来实现的。

②数据的分割(Partitioning of Data)。数据的分割是指把逻辑上统一的数据分割成较小的、可以独立管理的物理单元进行存储,以便于重构、重组和恢复,以提高索引创建和顺序扫描的效率[3]。数据的分割使数据仓库开发人员和用户具有更大的灵活性。

数据分割分为两种:系统级和应用级。系统级的分割是由 DBMS 和 OS(操作系统)实现的;应用级的分割由开发人员通过代码来直接控制。因此,应用级的分割更为灵活。在数据仓库的设计中,使用得较为普遍的是应用级的分割。数据分割完毕后,可用如下方法检验其正确性:能否将索引加到一个物理单元而不影响其他的操作。如果增加一个索引的操作非常复杂,就有必要对数据的分割进行调整或重新分割。

③元数据的设计和管理。元数据(Metadata)是数据仓库结构的一个重要部分,利用元数据能最有效地管理数据仓库。元数据的定义包括:

·面向程序员的数据结构;
·面向 DSS 分析员的数据结构;
·数据仓库的数据来源;
·数据模型;
·数据模型与数据仓库的联系;
·数据抽取的历程

二、数据发掘

数据发掘是从大型数据库或数据仓库中发现并提取隐藏在其中的信息或知识的过程。目的是帮助分析人员寻找数据间潜在的关联,发现被忽略的要素,而这些信息对预测趋势和决策行为是十分有用的。数据发掘的一般过程可用图 1 表示。

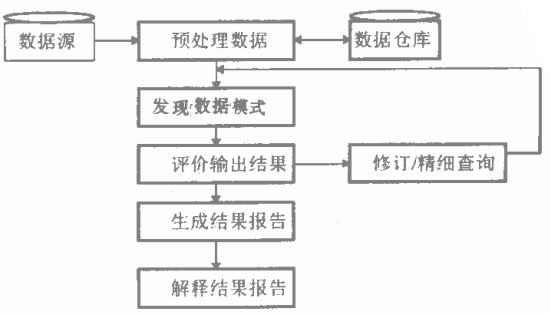

图 1 数据发掘的一般过程

(1)预处理数据:收集和净化来自数据源的信息,并加以存储,一般是将其存放在数据仓库中。

(2)模型搜索:利用数据发掘工具在数据中查找模型,这个搜寻过程可以由系统自动执行,自底向上搜寻原始事实以发现它们之间的某种联系,也可以加入用户交互过程,由分析人员主动发问,从上到下地找寻以验证假定的正确性。对于一个问题的搜寻过程可能用到许多工具。例如,神经网络、基于规则的系统、基于实例的推理、机器学习、统计方法等。

(3)评价输出结果。数据发掘的搜寻过程一般需要反复多次,因为当分析人员评价输出结果后,他们可能会形成一些新的问题或要求对某一方面作更精细的查询。

(4)生成最后的结果报告。

(5)解释结果报告。对结果进行解释,依据此结果采取相应的商业措施,这是一个人工过程。

①数据发掘的相关技术。为了简化和加快数据发掘

过程,使数据发掘真正方便、实用,还需其他的技术支持。如数据净化、数据仓库技术、强大的并行处理技术和存储技术。

(1)数据净化(Data Scrubbing)。为了使数据发掘能够产生合理的结果,数据在进入数据仓库以前必须清除错误,形成统一的格式,如用"1"和"0"代表性别,而不是用"male"、"female"、"man"、"woman"表示。这个过程可能用的很慢。此外,尽管有现成的软件工具可以辅助开发人员净化数据,将数据搬迁到数据仓库中,但开发人员还是要考虑数据如何表示、采用哪种格式等问题。

(2)数据仓库技术。一个企业在没有建立自己的数据仓库之前,有许多分散的、未集成的、不精练的信息,采掘这样的数据,效率是很低的。数据仓库为数据发掘提供了有效的结构,有利于数据发掘。

(3)并行处理技术。毫无疑问,强大的并行处理计算机可以提高数据发掘的应用,因为并行处理技术可以将一个复杂查询分解成多个子查询,每个子查询交给不同的处理器处理,这一处理过程是并行执行的,不象串行处理机,任务只能顺序执行。因此,并行处理技术可以大大加速数据发掘的过程。反过来,人们对数据发掘的兴趣也有助于并行系统的销售。

(4)存储技术。现在的数据仓库存储的数据量是 GB 到 TB 级别。随着时间的推移,在未来五年,可能会达到几百个 TB 级。因此,廉价可行的存储技术对于数据发掘来说变得非常重要。目前,普遍采用的是二级存储技术,即磁盘(磁光盘)-主存两级存储。由于缺乏快速地访问和存储磁盘的技术,随着存储容量的增长,数据发掘查询越来越复杂,并行处理器速度的加快,存储技术可能会成为数据发掘的新瓶颈。

三、数据仓库和数据发掘工具产品

数据仓库和数据发掘技术能够帮助用户从历史性数据中挖掘知识,进而支持决策,极大地吸引了用户,而用户造就的数十亿美元的市场又极大地吸引了数据库厂商。各大公司纷纷开始了自己的数据仓库开发与研制计划,以及数据发掘工具产品的研制工作。

Oracle 公司率先推出了企业级数据仓库解决方案。其 Designer/2000 是一个 CASE 产品工具,可以实现数据仓库的设计。它通过使用共享分析库,记录设计过程中对数据仓库的需求分析和说明信息,辅助设计人员完成数据仓库的数据建模工作。Sybase 公司一方面收购第三方软件厂商的数据仓库产品,一方面积极开发自己的联机企业级的智能仓库。Informix 公司在新产品 Online8.0 版本中融入了数据仓库功能。

在微机 Windows 平台上的数据仓库产品也进入研究开发高潮。开放式数据连接(ODBC)是数据仓库必须遵循的基本标准。目前 InterData 公司正在开发的 Smart Data 工具软件通过 ODBC 连接各个数据产品,如 Excel、Lotus 1-2-3 以及 FoxBase,从而建立数据仓库。

此外,第三方软件厂商也纷纷推出数据仓库工具产品,如 Prism Solutions 公司的 Warehonse Manager 产品,可以自动生成数据转化、综合与映射的代码,并可将源数据库中的数据映射到支持多种 RDBMS 的大型机与服务器中。Carleton 公司的 Passport、Evolutionary Technologies 公司的 Extract ToolSuite、Vality Technology 公司的 Integrity Programming Environment 等软件工具都可用于管理数据仓库繁重的维护工作,包括管理数据的获取与源数据的维护。

目前为数据发掘所提供的主要工具有:联机分析处理(OLAP:On-Line Analytical Processing)工具及包含一些 AI 技术的工具,如 IDIS(Information DIscovery System)。

OLAP 描述的是一种多维数据服务(这里的维是指人们观察客观世界的角度,如时间、地域、业务等),这种服务的设计目的是保证分析员、经理和决策者针对特定问题,通过快速、一致、交互式的实时数据访问和分析,获得有创意的发现。目前,典型的产品有 Pilot 公司推出的 Lightship 产品;Oracle 公司新近推出的 Oracle Express 系列产品;美国 Business Objects 公司于 1996 年底推出的 Business Miner 产品。Business Miner 产品是一个桌面式数据发掘工具,可在所有 Windows 平台(包括 Windows 95, Windows NT 和 Windows 3.X)上使用。

随着商业竞争愈来愈激烈,数据仓库、数据发掘技术的应用会越来越普遍,其产品会更加成熟。

参考文献

[1] Ari Silberschatz, Mike Stonebraker, Jeff Ullman. Database Research: Achievements and Opportunities Into the 21st Century.

[2] W.H. Inmon. Building the Data Warehouse. Boston: QED Technical Publishing Group, 1992.

[3] 王珊、刘方. 创建数据仓库的方法、模型与步骤. 计算机世界报. 1996.7.15.

(来稿时间:1997 年 10 月)