应用数据仓库技术实现决策支持系统

曹重英　陈洛资　肖锋　单莹　（长沙铁道学院信息工程学院　410012）

摘要：本文简述了数据仓库、联机分析处理、数据挖掘的概念和技术。提出和实现了一种利用数据仓库技术及其工具，结合传统 DSS 的四库结构，设计和实现决策支持系统的新方法。

关键词：决策支持系统　数据仓库　联机分析处理　数据挖掘

近年来，数据仓库技术的兴起给信息社会带来新的契机。逐渐成为 Internet 之后的又一技术热点。目前，在美国 30% 到 40% 的公司已经或正在建造数据仓库。以数据仓库技术为基础，以联机分析处理和数据挖掘工具为手段的决策支持系统日渐成熟。我们针对某市自来水公司的 CMIS 工程，利用数据仓库技术构建了一个企业管理决策支持系统。

一、早期决策支持系统的问题

由数据库、模型库和知识库为核心的旧的决策支持系统越来越不适新的要求。其规模受到限制，不能访问或以快速方式访问大型数据存储器或有高度标准结构的数据。传统的数据库作为数据管理手段，主要用于事务处理。其数据缺乏组织性，大多数以原始数据的形式存储，难以转化为有用的信息，效率低下，对分析处理的支持不能令人满意。以往的多数 DSS 只能停留在演示阶段，灵活性和可用性差，不实用，未能进入大规模工业工程实践。三库很难形成有机结合，容易形成数据孤岛。

二、主要技术介绍

1. 数据仓库技术（DW）

数据仓库技术是“面向主题的集成的、稳定的和随时间变化的数据集合，主要用于决策制定”。它从多个同构或异构的传统数据库中获取原始数据，先按辅助决策的主题要求形成当前基本数据层，再按综合决策的要求形成综合数据层。其数据组织形式可分为虚拟存储方式、关系数据库存储方式和多维数据库存储方式。以多维数据库形式存储比较适合。

数据仓库中的数据大体分为四级：远期基本数据、近期基本数据、轻度综合数据和高度综合数据。还有一部分重要数据是元数据。无数据是“关于数据的数据”，如关系数据库中的数据字典就是一种元数据。在数据仓库中用来与终端用户的多维商业模型/前端工具之间建立映射的元数据称为 DSS 元数据。

如果数据在应用数据库中未发生变化，则不需向数据仓库中追加，数据追加的内容仅限于上次数据仓库输入后在应用数据库中变化了的数据。

目前，高性能数据库服务器可处理海量数据，进行复杂和要求优化的查询；并行数据库技术可并行存储管理超大规模数据库，提供高速复杂查询的能力；网络技术为大量数据通过网络传输和转化提供方便。这些使数据仓库的发展有了坚实的基础。

2. 联机分析处理（OLAP）

联机分析处理是针对特定问题的联机数据访问和分析。通过对转换后信息的很多种可能的观察形式进行快速、一致和交互的存取，得到对数据更进一步的观察。

OLAP 与以往的 OLTP（联机事务处理）是不同的，主要表现如下：

OLAP	OLTP
面向决策人员，支持管理需要	面向操作人员，支持日常操作
导出的，综合的，历史的数据	原始的，细节的，当前的数据
分析驱动	事务驱动
不可更新，但周期性刷新	可更新
数据处理量大	数据处理量小

目前，有基于多维数据库的 OLAP 和基于关系数据库的 OLAP。多维数据库以多维方式来组织数据，以多维方式显示数据。其多维概念清晰，占用存储少，统计速度远远超过 RDBMS。现有关系数据库大量存在，基于关系数据库的 OLAP 也是可行的。

3. 数据挖掘（DM）

数据挖掘是一个决策支持过程，主要基于 AI、统计学、机器学习等技术，高度自动化地分析企业原来的数

据,作出归纳性的推理,从中挖掘出潜在的模式,预测客户的行为,帮助企业的决策者调整市场策略,减少风险,作出正确判断和决策。

DM 利用了人工智能中的一些成熟的算法和技术,例如,人工神经网络、遗传算法、决策树、邻近搜索方法、规则推理、模糊逻辑等。DM 系统利用的技术越多,得出的结果精确性越高。

从功能上讲,DM 可分为以下四种:关联分析、序列模式分析、分类分析和聚类分析。DM 的数据分析过程为四个步骤:数据准备、挖掘、表叙及评价。

三、应用数据仓库技术的决策支持系统

1. CZ_DSS 系统开发环境

系统采用 Client/Server 模式,以 WindowsNT 4.0、Windows95 为运行环境,选择 ORACLE7 Server 为数据库服务器环境,以 VB、Express Object 等为前端开发工具。我们把数据库服务器和数据仓库服务器捆绑在一起。

2. 系统架构

该系统分为六个组成部分:基础系统网络、数据采集、数据仓库、多维数据库、以及(知识库、模型库、方法库)三库组合和应用系统。它们之间相互作用,构成一个层次分明的信息分析环境。(见图 1)

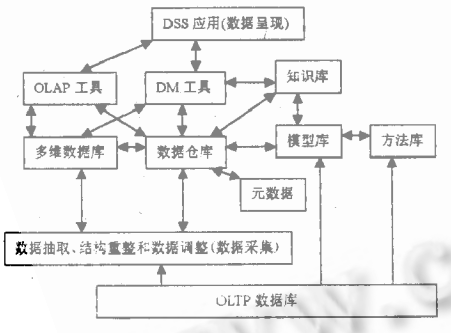

图 1　CZ_DSS 系统结构

3. CZ_DSS 的设计与实现

CZ_DSS 系统利用数据仓库技术及其工具,结合传统 DSS 的四库结构,设计和实现决策支持系统,以实现这种新方法,克服传统数据库技术的缺点。

数据仓库的设计不同于传统 OLTP 数据库的设计。其设计是数据驱动的,开发是一个不断循环、反馈、完善的过程。在整个开发过程中,决策者和开发人员应密切合作,周密筹划,减少无效或重复劳动。其包括概念模型设计、逻辑模型设计、物理模型设计、实施等。

经过调研,我们完成如下整体设计。CZ_DSS 系统中的数据流程为:数据采集系统采集 OLTP 数据库中的各类数据,重整结构和调整数据后归类存放在数据仓库,然后由多维数据库多层次分类成有效信息,知识库、模型库和方法库三库有机结合,给数据挖掘以强大的支持,最后通过 OLAP 工具和 DM 工具将数据呈现给用户。

(1)基础系统网络。系统支持网络分两层。第一层是全公司的网络数据库服务中心;第二层为各分散的二级单位的局域网。这为数据仓库的建立提供完整的技术支持手段,包括网络连接、数据库互连和访问等。

(2)数据采集系统。由于数据源的多样性和异构性,我们建立一个数据采集系统按数据仓库的设计要求从 OLTP 应用数据库中提取数据,重整后存放在数据仓库,根据多维数据库的特性调整部分数据。各业务数据库的数据类型一般是不一样的,因此,必须确保数据的一致性和可用性,进行必要的数据转换。例如,各个数据库对时间的表达并不是一样的,我们把它们转化为同一数据类型。要根据业务的需要,抽取必要的数据,进行综合。例如,我们对用户记录表上的电话号码之类无关的数据就不抽取。选取合适的数据采集算法是十分必要的,否则数据仓库就会成为数据垃圾收集站。

(3)数据仓库。数据仓库的一个重要特征是要求数据按照其自然属性来组织,即面向主题。所谓主题,是构成企业经营运作的主要框架或方向,是在较高层次将数据归类的标准。通过调研,确定"销售"、"财务"等主题。

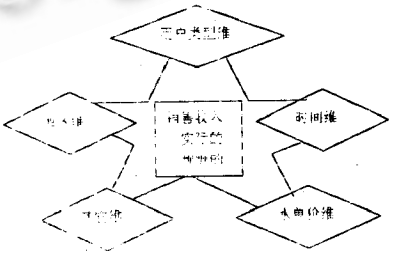

图 2　"销售"主题星型图

数据建模对数据仓库是至关重要的。与传统的数据库建模技术相比,星型图最适合查询为基础的情况。其最大优点是:在用户查询和收集时,可对大量指标实体进行筛选,以减少其最终容量。图 2 是 CZ_DSS 中的一个例子。

数据仓库中的数据结构是在现有业务系统数据结构基础上,针对管理信息的特征:时间特性和汇总特性,对数据的名称、类型、描述及关联进行重新定义,主要包括:统一数据类型、调整数据长度和增加时间属性。例如,通过汇总等,我们可在"销售"这个主题中得到欠费额,其在业务数据库中是没有的。

CZ_DSS 中元数据描述了整个数据仓库系统环境,包括数据字典和数据处理规则,对数据仓库、多维数据库和数据采集系统的开发设计起主导控制作用。

因在数据仓库中导库的数据量巨大,我们采用增量视图维护的方法,应用了 ECA 算法(Eager Compensating Algorithm)。这种算法基于 FIFO 模型,针对原始数据变化进行补偿请求查询,并将变化反映到实视图上。这种视图维护方法对减少通过网络访问各业务数据库的数量是十分有效的。同时,CZ_DSS 对实时性的要求不是太高,我们利用夜间从各业务库导入数据,进行脱机维护。目前,如何在联机状态下为用户提供服务并同时对系统进行热维护还处于研究状态下。

建立数据仓库的目的是在不调整现有各类业务系统的情况下,提供一个适应 OLAP 和 DM 应用的统一、全面、详尽的数据源。这些数据包括公司当前和历史的详尽数据。

(4)多维数据库。本多维数据库按该公司管理人员分析决策的自然方式构建数据模型,从而形成信息分析的多维视图,它是 OLAP 的数据引擎。确定多维数据库中维的数目和内容,即多维数据库的结构,是设计多维数据库的关键。我们是通过 ORACLE EXPRESS ADMINISTATER 设计多维数据库的。

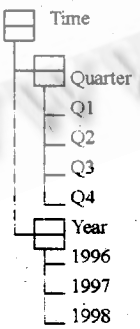

图 3　时间维结构图

维数据模型涉及到维、变量、公式、关系等概念。例如,我们在营销决策子系统中建立的维有地区、片、时间、用户名、用户类型、交费类型等,变量有水单价、水成本价、实际用水量、实际交纳水费等,关系有地区·片、用户名·用户类型等。

多维数据库中各维的层次划分,基本上确定了每个垂直的汇总路径。数据按照这些汇总路径构造之后,当沿其中任何一条路径自上而下分析时,可实现下钻分析。图 3 是时间维结构图,从中可看出层次结构。

(5)数据仓库工具应用和三库组合。在 CZ_DSS 中,数据挖掘算法作为数据挖掘模型存储于模型库,模型库利用方法库提供的方法工作,通过数据挖掘模型发现的模式和知识输入知识库。DM 工具利用数据仓库、方法库、模型库和知识库共同完成数据挖掘过程。DM 过程的启动通过时间机制来完成。我们采用贝叶斯网络和关联规则,通过数据挖掘进行故障诊断、生产过程优化、经营风险评估等。

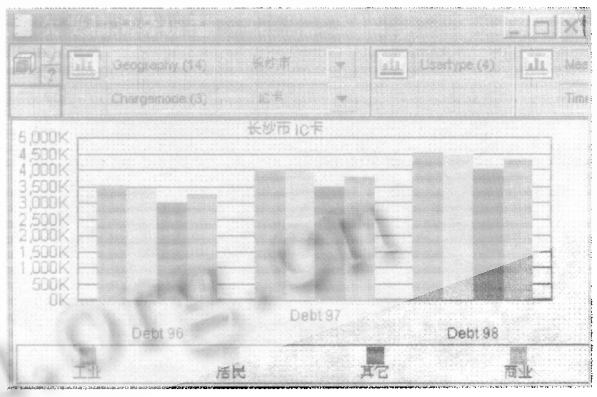

图 4　欠费情况分析图

OLAP 工具使用 DW 中的集成数据构建面向分析的多维数据模型。其快速、一致地访问各种信息视图,帮助公司管理人员发现问题和规律。这些视图从原始数据转化而来,以用户易于理解的方式反映公司的真实情况。OLAP 支持用户进行多维分析,如跨维、跨层次的计算和建模:趋势分析,预测分析;切片和切块;旋转分析等。系统提供了多视角查询、分析、预测和制作动态图表的功能。我们使用 ORACLE EXPRESS OBJECTS 作为联机分析工具。通过数据呈现和形象化,我们可以发现诸如哪些类型的用户老是欠费,用水的趋势,哪类用户是重点

的利润来源等。图 4 是 CZ_DSS 中的一个欠费情况分析表。此图反映了各个时间跨度，该市各区各片、居民、商业、工业、其他四种用户类型，IC 卡、银行托收、直收三种收费方式的欠费情况。从此图上，管理人员可发现诸如什么时间欠费情况较严重、那种方式收费情况较好等。通过分析，营销管理人员可抓住重点，及时进行调整，指导商务活动，从而达到整个商业目的。在以前，只能等到季度报表才能知道情况，而且分析起来也不是很方便。

四、结论

本文利用我们所作的一个课题介绍数据仓库技术设计和实现决策支持系统，提出一种利用数据仓库技术及其工具，结合传统 DSS 的四库结构，设计和实现决策支持系统的新方法。CZ_DSS 统一了公司管理信息的采集，大大减少了人为干预，为用户提供了多角度、多层次查询、分析、预测的功能，使得管理人员能及时发现问题和进行预测，大大地提高了效率。

参考文献

[1] W. H. Inmon. Building the Data Warehouese: QED Technical Pulishing Group, 1992

[2] R. Agawal et al. Modeling Mulitidimimensional Databasees: Proc. of the ICED, 1996

[3] GILL H S, Rao P C. 数据仓库——客户/服务器计算指南。(王仲谋，刘书舟译)，北京：清华大学出版社，1997

[4] Tom Hammergren. 数据仓库技术。(曹增强，王备战，岳晓奎 译) 北京：中国水利水电出版社，1998

(来稿时间：1999 年 8 月)