基于联机分析处理的数据仓库分析
Analysis of Data Warehouse Based on Online Analytical Processing

尹 松[1],周永权[2]
Yin Song[1],Zhou Yongquan[2]
(1. 广西大学计算机与电子信息学院,广西南宁 530004;
2. 广西民族学院计算机与信息科学学院,广西南宁 530006)
(1. Coll. of Comp. & Elec. Info., Guangxi Univ., Nanning, Guangxi, 530004, China; 2. Coll. of Comp. & Info. Sci., Guangxi Univ. for Nationalities, Nanning, Guangxi, 530006, China)

摘要:介绍数据仓库的特性、联机分析处理技术、数据挖掘技术和决策支持系统,以及这些技术与数据仓库的关系。

关键词:数据仓库 联机分析处理技术 数据挖掘 决策支持系统 数据集市
中图法分类号:TP311.13

Abstract: The concept, structure of data warehouse and its relationship to database and data mart are introduced ROLAP and MOLAP are analyzed. The theory and model of data mining which are based on data warehouse are expounded.

Key words: data warehouse, online analytical processing, data mining, DSS, data mart

随着企业事务处理系统的运行和建立,信息量越来越大,企业数据源越来越多,传统的面向数据操作的数据库已经不能满足形势发展的需要,数据仓库应运而生,它是体系结构化环境的核心,是决策支持系统(DSS)处理的基础。

数据仓库是一种概念,不是一种产品,它包括电子邮件文档、语音邮件文档、CD-ROM、多媒体信息以及还未考虑到的数据[1],数据仓库建立在一个较全面、较完善的信息应用的基础之上,用于支持中高层决策分析。完整的数据仓库应包括数据仓库技术、联机分析处理技术(Online Analytical Processing,简称 OLAP)和数据挖掘技术(Data Mining,简称 DM)等方面的内容[2]。

数据仓库中的数据只有通过数据挖掘技术才能变为决策支持系统(DSS)的有用信息,DSS 主要是为企业的中高层领导进行决策服务,OLAP 是数据仓库中的一种数据分析技术,它的数据来源于数据仓库。OLAP 能提供数据的多维概念视图,使用户能从多角度、多侧面、多层次考察数据库的数据,以便更好地为决策支持系统服务。本文将分析数据仓库的特性、联机分析处理技术、数据挖掘技术和决策支持系统,以及这些技术与数据仓库的关系。

1 数据仓库的特性

1.1 数据仓库

数据仓库历来没有一个标准化的定义,20 世纪 80 年代中期,“数据仓库”首次出现在 William. H. Inmon 的《建立数据仓库》一书。随着人们对大型数据系统的深入研究,人们对数据仓库的定义达成了以下共识:数据仓库中的数据是面向主题的、集成的、不可更新的并随时间不断变化的数据集合,建立数据仓库的目的就是为了更好地支持决策分析[3]。

1.2 数据仓库结构

数据仓库是从多个信息源中获取原始数据,经过加工整理后,存储到数据仓库的内部数据库中,通过数据仓库访问工具,向数据仓库的用户提供统一、协调和集成的信息环境,支持企业全局的决策过程和企业经营管理综合分析。数据仓库的系统结构如图 1 所示。

2004-07-19 收稿。

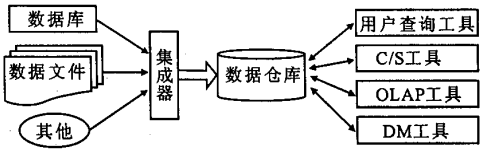


图1 数据仓库结构

1.3 数据库、数据集市与数据仓库的关系

随着数据库数据量的增大和用户查询要求趋向复杂，传统的数据库出现许多问题，如数据分散，缺乏组织性；数据难以转化为有用的信息；不能满足复杂查询的要求；不能为企业决策提供有效的支持等。为了克服数据库的这些弊端，人们开始尝试对数据库中的原始数据进行再加工，形成一个综合的、面向分析的环境，以支持决策的产生。为此，形成数据仓库的思想，该思想主要是要在数据库技术的基础上快速、高效、准确地将决策支持所需的数据信息从日常运行的数据中分离出来，对于一些分散的数据，要对它进行归纳、综合、分析，转化为集中统一、随时可以方便调用的管理决策信息。

数据库与数据仓库的主要区别如表1所示。

表1 数据库与数据仓库的主要区别

数据方式	生存周期	存取结果	存取方式
数据库	生存期短，数据经常变化	满足记录层的需求	反复的事务存取模式
数据仓库	生存期长，数据是相对静态的	数据以聚集方式的集合存取	做定期报告式的查询

数据方式	修改方式	驱动方式	操作方式	规模大小
数据库	对数据做实时修改	事件驱动	可读可写	几个GB
数据仓库	周期性地大批量修改数据	数据驱动	只读	可达数百GB

数据集市是数据仓库中的一个术语，分为独立的数据集市(Independent Data Mart)和从属的数据集市(Dependent Data Mart)两种[4]，它与数据仓库的区别不仅是数据量的大小，数据仓库是企业级的，它为企业各个部门的运行提供决策支持手段，而数据集市是部门级的，一般只能为某个局部范围内的管理人员服务，所以数据集市也称为部门级数据仓库(Departmental Data Warehouse)，它是一种更小的、更集中的数据仓库，是企业分析商业数据的一条廉价路径，它也可以升级到完整的数据仓库。

数据库、数据仓库、数据集市三者的关系如图2所示。

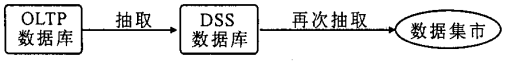


图2 数据集市的形成

2 数据仓库中的OLAP

2.1 OLAP[5]

OLAP是一种数据分析技术，它能够完成基于某种数据存储的数据分析功能，具有快速性、可分析性、多维性、信息性等特性。联机事务处理OLTP是以数据库为基础，面对操作人员和底层管理人员，完成基本数据的查询、删除、修改等处理工作。它是操作型应用，而OLAP是在OLTP的基础上发展起来的，它是分析型应用。OLAP是以数据仓库为基础的多维分析处理，它具有灵活的分析功能、直观的数据操作、分析结果可视化表示等优点，能满足数据仓库对大量的数据进行对比、综合、归纳和预测的需要。OLAP通过面向对象方式组织数据，以多维数据结构存储数据，采用多维分析方法，使用户对大量复杂数据的分析变得清楚高效。

2.2 OLAP的2种基本模式[6]

(1)多维OLAP(Multi-OLAP)，基于多维数据库(MDB)的OLAP技术，尤其适用于预算编制、财务建模、盈利分析和预测等领域，MOLAP以多维结构形式存储数据，便于用户查看，使得MOLAP工具快捷灵活。

MOLAP的数据存储一般有2种方式：

1)超立方结构(Hypercube)：指用三维或更多的维数来描述一个对象，每个维彼此垂直，数据的测量值发生在维的交叉点上，数据空间的各个部分都有相同的维属性。

2)多立方结构(Multicube)：指将大的数据结构分成多个多维结构，这些多维结构是大数据维数的子集，面向某一特定应用对维进行分割，将超立方结构变为子立方结构，因此多立方结构具有很强的灵活性，提高了数据(尤其是稀疏数据)的分析效率。

(2)关系型OLAP(Relational-OLAP)，其功能类似于MOLAP，但其底层数据库是关系型数据库，而不是多维数据库。ROLAP对传统RDBMS进行扩充以实现对数据仓库的联机分析处理，与MOLAP使用数据立方直接实现多维数据视图不同，ROLAP将对视图的相应操作映射到关系表与SQL查询上。

实际应用中，通常将ROLAP和MOLAP结合使用，就形成一个新的OLAP结构，即混合型OLAP，它将两者的优势结合起来，利用关系数据库存储历史数据、细节数据，发挥关系数据库技术成熟的优势，而在多维数据库中存储当前数据和常用数

据，以提高操作性能。

2.3 OLAP 的分析方法

多维数据分析是 OLAP 的一个重要属性，对以多维形式组织起来的数据采取切片和切块、旋转、上卷和下钻等各种分析动作，以求破析数据，使得用户能从各个角度和各个侧面观察数据仓库里的数据[7]。

(1)切片与切块。多维数据结构按二维进行切片，按三维进行切块，得到分析所需要的数据；

(2)旋转。它是一种视图的操作，通过旋转可以从不同得视角得到所需数据；

(3)上卷和下钻。上卷是在某一维上将低层次的细节数据概括到高层次的汇总数据。下钻则相反，它从汇总数据深入到细节数据进行观察，这两者都是改变维的层次和改变分析的粒度的操作。

2.4 ROLAP 和 MOLAP 特点分析

ROLAP 的主要特点是灵活性强，用户可以动态定义统计或计算方式，而且现有的关系数据库的技术可以沿用，缺点是它对用户的分析请求处理时间要比 MOLAP 长，SQL 无法完成多行计算和维之间的计算。

MOLAP 结构的主要特点是支持高性能的决策支持计算，能迅速地响应决策分析人员的分析请求并快速地将分析结果返回给用户，这主要得益于它的独特的多维数据结构以及存储在其中的预处理程度很高的数据，缺点是需要进行预计算，增加系统复杂度，且不支持维的动态变化。

2.5 数据联机分析挖掘

数据联机分析挖掘(Online Analytical Mining，简称 OLAM)是将数据 OLAP 和 DM 有机地结合在一起的。OLAP 和 OLAM 都是基于 C/S 的模式，它们的特点是与用户的交互性。OLAM 的挖掘分析处理建立在超级立方体的基础上，虽然 OLAM 的多维计算需要更多的维数和更强大的访问工具，但是用于 OLAP 和 OLAM 的立方体没有本质的区别，OLAM 服务器通过用户图形接口接收用户的分析指令，在元数据的指导下，对超级立方体做一定操作，然后将挖掘分析结果返回给用户，这个过程是动态的[6]。

从 OLTP 到 OLAP 再到 OLAM，是数据库到数据仓库的发展，是联机事务型处理到联机挖掘型处理的发展，是知识管理向知识发现的发展，是浅层利用数据向深层利用数据的发展，也是人工手动型向人工智能型的发展。

3 数据挖掘技术[8]

数据库知识发现(Knowledge Discovery in Database，简称 KDD)是指从数据库的大量数据中提取正确、新颖、潜在有用和最终可理解模式的过程。KDD 过程一般由数据准备、数据挖掘和解释评估三部分组成。

数据挖掘是 KDD 中的一个特定步骤，是在可接受的计算效率的限制下，应用数据分析和发现算法，从数据库的大量数据中提取隐含的、目前未知的、潜在有用的和最终可理解模式的过程。数据挖掘是建立在数据仓库上的决策支持技术，将数据挖掘应用到数据仓库系统环境中，可以增强用户的决策支持能力。数据挖掘从大量的数据中抽取潜在的有用信息的过程可以分为数据选择、数据转换、数据挖掘、结果分析。数据挖掘与传统数据分析方法的区别是在没有明确假设的前提下挖掘信息并发现知识。

数据挖掘的过程如图 3 所示。

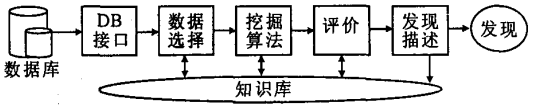


图 3 数据挖掘逻辑模型

4 决策支持系统[9]

决策支持系统(DSS)的目标是支持企业中高层管理者的决策过程，它使中高层管理者可以方便地根据已经获得的数据，运用定量分析和定性分析技术，从多个侧面对决策问题进行分析、评价和判断。决策支持系统中各种模型、规则的建立，都是以历史的和现在的数据为基础，所以数据是 DSS 的分析基础。而数据仓库用于大量数据存储和组织，数据挖掘建立在数据仓库的基础上，数据挖掘和数据仓库技术的结合为企业 DSS 的建立提供新的、更有效的解决方案。

数据仓库和 OLAP 是现代 DSS 的重要组成部分，与传统的 OLTP 不同，它们是对现有的数据进行归纳、分析和推理，从而为决策服务的。基于数据仓库的 DSS 应用是一种“客户－多维分析服务器－服务器”三层体系结构，数据仓库和 OLAP 技术将不同的数据集成到一起，克服传统数据库的多数据源、历史数据利用不充分、分析效率低等问题，为决策者提供有力的决策支持，可以说数据仓库和 OLAP 技术为现代 DSS 的开发和应用开辟一条新的快捷路径。

5 结束语

数据仓库作为数据库领域发展起来的一种新型的技术，它在数据的管理和使用上与传统数据库有着本质的不同，数据仓库是大量集成化数据的集合，是多种技术的综合体。OLAP提供给数据仓库系统一种高灵活性、高性能地存取、浏览和分析数据的手段，它的分析结果可以为数据挖掘提供挖掘依据，而数据挖掘又可以拓展OLAP分析的深度，可以发现OLAP不能发现的更为复杂的信息，因此，将OLAP与数据挖掘相结合将会在数据仓库中发挥更好的效用，这也是OLAP发展的一个新的方向。

参考文献：

1 Inmon W H. What is a data warehouse? http://www.billinmon.com,2000.

2 马宏鹏，赵 新，李 明，等. 数据仓库原型系统设计. 计算机工程与应用，2000，11：109～111.

3 彭木根. 数据仓库技术与实现. 北京：电子工业出版社，2002.6.

4 柳 莺，赵艳红，钱 旭，等. 数据仓库技术研究和应用探讨. 计算机应用，2001，2(2)：46～48.

5 Inmon W H. A Data Warehouse/OLAP Framework for Web Usage Mining and Business Intelligence Reporting. http://www.billinmon.com,2002.

6 张 旭，董有田. OLAP多维数据分析与应用研究. 黑龙江科技学院学报，2002，9(3)：16～19.

7 Inmon W H. OLAP and Data Warehouse. http://www.billinmon.com,2002.

8 Inmon W H[美]著. 构建数据仓库. 王志海，林有芳，等译. 北京：机械工业出版社，2003.3.

9 戴超凡，邓 苏，黄宏斌，等. DSS中数据管理新技术研究. 计算机工程与应用，2000，12：21～24.

（责任编辑：黎贞崇）

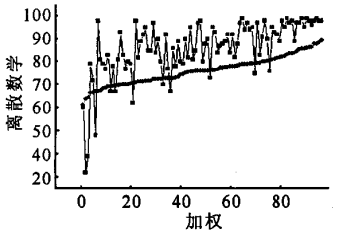
（上接第242页）

以及高级程序员的通过率可以看出，其中高级程序员的通过率占全院的40%左右。从这个比率看，2000级的软件开发能力和编程能力非常突出。

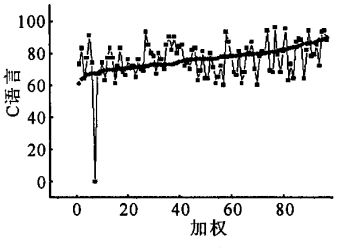
2000级和1999级的一些重要专业基础课程未能出现在前几个主成分中，为此，针对离散数学和C语言两门专业基础课进行分析，以加权主成分排序作为自变量，专业课成绩作为因变量作散点图（图3）.

从图3可以看出，两课程的成绩不能很好地区分按加权主成分排序的学生的学习能力，这也说明，两课程的成绩不能有效地区分学生的学习能力。

通过表3的分析，可认为我系新办专业的课程设计和建设还需要进行整合，课程的开设的合理性



(a)



(b)

图3 散点图

(a)离散数学；(b)C语言

还有待论证，课程成绩中反映出的问题需要认真分析、解决。

参考文献：

1 范金城，梅长林. 数据分析. 北京：科学出版社，2002.

2 樊 欣，邵谦谦. SAS8.X经济统计. 北京：北京希望电子出版社，2003.

3 朱 宁，符名培，方进一. 教学研究中的主成分模型. 桂林：桂林电子工业学院学报，2004，(2)：97～99.

（责任编辑：黎贞崇）