基于医院信息系统的数据仓库研究及实现

陈　玲　　广州市红十字会医院 计算机中心

摘要：本文主要研究在现有的医院信息系统数据库的基础上，应用数据仓库技术、联机分析技术和数据挖掘技术，结合统计分析模型，建立了初具规模的医院信息系统数据仓库，探讨了医院信息系统中数据查询不准确问题和历史数据查询问题；利用 OLAP 技术和统计模型，对引入数据仓库的数据进行分析和应用，根据决策者的需求，设计不同的主题，并在此基础上建立了基于医院信息系统的数据仓库系统，使医院信息系统成为了医院决策者可信赖的决策依据平台。

1 相关理论基础

1.1 数据仓库

1 数据仓库的定义和特点

著名的数据库专家W.H.Inmon在其著作《Building the Data Warehouse》中给予如下描述：数据仓库（Data Warehouse　DW）是一个面向主题的（Subject Oriented）、集成的（Integrate）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant）的数据集合，用于支持管理决策[1]。对于数据仓库的概念可以从两个层次予以理解：首先，数据仓库用于支持决策，面向分析型数据处理，它不同于企业现有的操作型数据库；其次，数据仓库是对多个异构的数据源有效集成，集成后按照主题进行了重组，并包含历史数据，而且存放在数据仓库中的数据一般不再修改。

针对以上定义，数据仓库有以下特点[2]：

(1) 面向主题：操作型数据库的数据组织面向事务处理任务，各个业务系统之间各自分离，而数据仓库中的数据是按照一定的主题域进行组织；

(2) 集成的：面向事务处理的操作型数据库通常与某些特定的应用相关，数据库之间相互独立，并且往往是异构的，而数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库的信息是关于整个企业的一直的全局信息；

(3) 时变的：操作型数据库主要关心当前某一个时间段的数据，而数据仓库中的数据通常包含历史信息，系统记录了从过去某一时点（开始应用数据仓库的时点）到目前的各个阶段的信息，通

过这些信息，可以对发展历程和未来趋势做出定量分析和预测。

(4) 相对稳定的：操作型数据库中的数据通常实时更新，数据根据需要即使发生变化。数据仓库中的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期的加载、刷新。

数据仓库最根本的特点是物理地存放数据，而这些数据并非是最新的、专有的，而是来源于其它的数据库，数据仓库的建立并不是要取代原有的数据库，而是建立在一个较全面、完善的信息应用的基础上，用于支持高层决策分析，数据仓库是数据库技术的一种新的应用，它是用数据库管理系统来管理其中的数据。

2 数据仓库与关系数据库的区别

传统的数据库技术是面向事务处理的，它是对现有数据的归纳、分析和推理，其主要功能是为联机分析处理提供支持。传统的数据库技术和数据仓库技术的区别见表1[3]。

表1　数据库技术和数据仓库技术的区别

数据库	数据仓库
面向应用	面向主题
详细的或全面的	综合的或是提炼的
在存取瞬间是准确的	代表过去的数据
为日常工作服务	为管理者或分析服务
可更新	不更新
重复运行	启发式运行
处理需求事先可知	处理需求事先不知道
对性能要求高	对性能要求宽松
事务处理驱动	分析处理驱动
更新控制主要涉及所有权	无更新控制问题
高可用性	松弛的可用性
整体管理	以子集管理
非冗余性	时常有冗余
静态结构：可变的内容	结构灵活
一处处理数据量小	一处处理数据量大
支持日常操作	支持管理需求
访问的高可能性	访问的低可能或适度可能性

1.2 联机分析处理（OLAP）[4]

OLAP是使分析人员、管理人员或执行人员能够从多种角度对从原始数据中转化出来的、能够真正为用户所理解的、并真实反映企业维特性的信息进行快速、一致、交互地存取，从而获得对

数据的更深入了解的一类软件技术(OLAP委员会的定义)

OLAP的技术核心是“维”。“维”是人们观察客观世界的角度，是一种高层次的类型划分，一般包含着层次关系,这种层次关系有时会相当复杂。OLAP通过把一个实体的多项重要的属性定义为多个“维”，可使用户能对不同“维”上的数据进行比较，因此可认为它是多维数据分析工具的集合。OLAP的基本多维分析的操作有上卷(rollup)、下探(drilldown)、切片(slice)、切块(dice)、旋转(pivot)等。

1.3 数据挖掘

Joseph P.Bigus 在他的书《数据挖掘和神经网络》写道，数据挖掘（Data Mining DM）是从一个大的数据聚合中有效地发现不明显却有价值的信息。数据挖掘以自动发现新事实和数据的关系为中心。用传统的查询工具，你只能查询已知的信息。假定更多的有用信息是隐藏着的，而数据挖掘工具能够使你揭开这些隐藏的信息[5]。

2 建立医院数据仓库系统是医院信息系统发展的必然

目前的医院信息系统着重于医院业务流程,医院各个业务部门都有一个清晰、功能完善的子系统来完成相应的业务工作，各个子系统都是相互联系的，这种数据库是操作型事务处理数据库，随着计算机应用和网络技术的发展，医院管理决策者希望医院信息系统能够更多地参与数据分析和辅助决策，而不再是对数据简单的收集、整理、查询和统计。传统的数据库系统只能较好完成数据统计分析原始动态数据和日常统计报表的任务。不可否认，医院信息系统的报表查询功能在辅助决策、统计服务等方面也发挥了较大的作用，能够完成日常统计报表服务和在此之上的医疗质量指标的数据分析。

建立数据仓库，可以将原始的操作数据进行多方位的分析，用户利用数据仓库和联机分析处理技术（OLAP）可以实现对数据的多维分析、向下探查分析和变化趋势分析、掌握各个层次的数据并和前期或同期的数据作对比、分析数据的变化趋势等等。同时，把操作型数据与分析型数据分开，大大减轻了操作型数据库的负担，同时大大提高了数据统计分析的速度，这种分析、统计比传统数据库系统全面、灵活、快速。

3 医院数据仓库系统的总体设计

在构建医院的综合数据仓库系统时，根据医院信息系统的现状，综合数据仓库系统主要分三个

层次：第一层，数据仓库层，实现数据仓库和简单的查询和基本的统计分析，并使用报表和图的方式进行结果显示，构造决策支持系统的基础；第二层，联机分析处理（OLAP）。包括切片与切块、上卷、下钻和旋转等操作，它与简单的信息处理相比的主要优势是支持多维数据分析功能；第三层，实现数据分析功能（数据挖掘）。构造分析模型，进行分类和预测，并用可视化工具展示分析结果，完成决策支持系统。

本系统采用客户/服务器体系结构，后台以部门数据集市和全局数据仓库为支持，用户使用前端的OLAP工具和统计分析模型进行数据分析，得出有统计学意义的分析结果，提供决策支持。

根据医院的医院信息系统的实际情况，整个基于医院信息系统的决策支持系统的总体架构如图1：

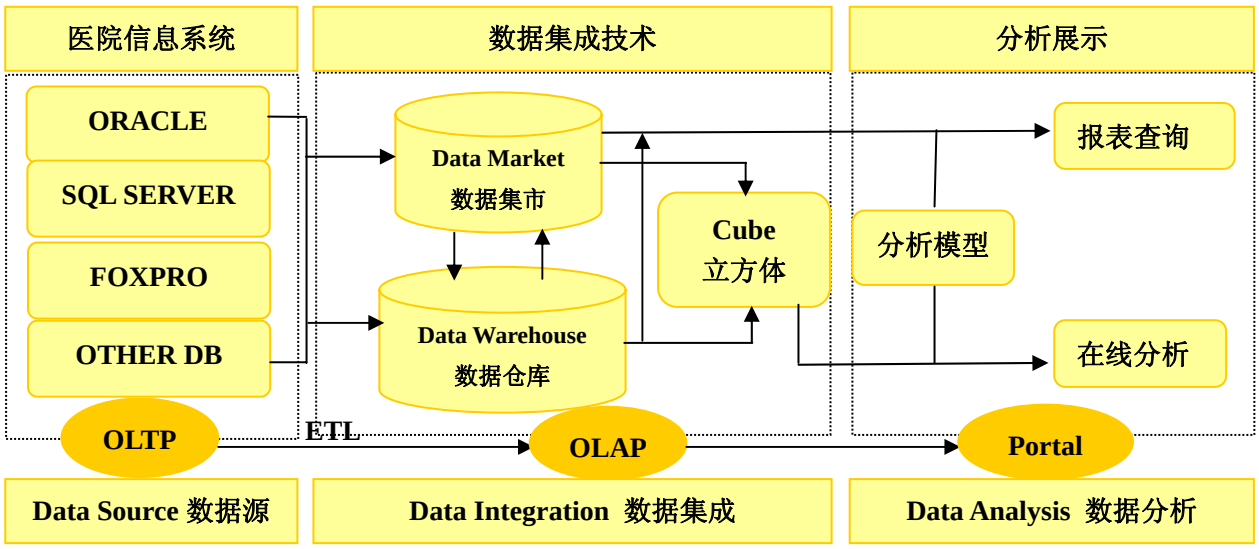

图1　医院综合数据仓库系统总体框架图

4　开发范例

数据仓库的建立是一个循环往复的过程，建设过程涉及数据的选择、变换、建模、评估、解释模型、运用和巩固模型等步骤。下面以医院住院收入多因素分析为范例详细说明医院信息系统数据仓库的建立过程。

4.1 数据现有资源

应用系统：HIS；生产数据库：SQL-Server for Window （在线数据）；历史数据平台： SQL-Server for Window 和 Foxpro for DOS 。

4.2 确定主题

数据仓库的中的数据是按分析主题来组织数据的，因此确定主题是数据仓库建设的首要目标。同时还要根据主题确定子主题、维度和数据来源等（如表2所示以医院业务收入统计为主题确定子主题及维度，数据来源等）。

表2 数据主题表

主题	子主题	维度	数据来源
业务收入	门诊业务	时间、科室	HIS
	住院业务	时间、科室	HIS、统计报表
费用	门诊费用	科室、费用属性	HIS
	住院费用	科室、费用属性	HIS
疾病	治疗方案	时间、科室、医生、医嘱	HIS
	疾病分布	时间、地区	HIS、病案系统

4.3 建立模型

本范例主要是应用成熟的统计模型，对医院业务收入进行分解分析，并用不同的维度表示出来。（1）模型公式：应用多元函数变量因素分析对医院住院业务收入进行动态分析。将住院业务收入(U)分解为出院人次(x)、出院者平均住院日(y)和出院者平均人日费用(z)3 个因素；多元函数的变量关系式如下：

住院业务收入：$u=xyz$

变量表达式为：$\triangle u=y_0z_0(x_1-x_0)+ x_0 z_0 (y_1- y_0)+ x_0 y_0 (z_1-z_0)+ \delta p$

$$\delta p= \triangle u-(y_0 z_0\triangle x+ x_0 z_0\triangle y+ x_0 y_0\triangle z)$$

（2）动态比较：应用环比，对医院业务收入按照年、季度、月进行动态比较。

（3）统计图形：饼图（多因素分析）、条图（环比分析）

4.4 数据准备

数据准备的好坏将影响到数据挖掘的效率和准确度以及最终模型的有效性，数据准备工作包括数据的选择、探索、修正、变换，在数据准备阶段，医院的数据(尤其是部门、各种数据字典等基本资料)分散，数据表关系复杂，整理转换难度很大，对应关系复杂(对数据表有一对多，多对一关系同时出现等)。

1. 数据仓库表结构

基于以上的分析模型，建立数据仓库，并根据数据仓库与医院的数据源关系，建立 ETL 提取和抽取数据过程，数据仓库的表结构如下（图 2）：

图 2　住院业务收入表

	列名	数据类型	长度	允许空
▶	年份	float	8	✓
	月份	float	8	✓
	出院人次	float	8	✓
	出院者平均住院日	float	8	✓
	出院者人均日费用	float	8	✓
	住院业务收入	float	8	✓
	月出院人次增长额	float	8	✓
	月出院者平均住院日增长额	float	8	✓
	月出院者人均日费用增长额	float	8	✓
	月出院人次增长率	float	8	✓
	出院者平均住院日增长率	float	8	✓
	月出院者人均日费用增长率	float	8	✓
	季度	float	8	✓

2. 数据仓库抽取、转换和装载

第一步　数据抽取装载过程：

1)历史数据抽取

由于历史数据不存在变动问题，只要一次性导入即可,导入历史数据的难点是历史数据与当前数据汇总的一致性问题和数据转换问题。

2)当前数据抽取

建立当前数据抽取完整的过程，根据需要，可以实现每天抽取或每周（每月）抽取，由服务器自动完成。

第二步：数据清洗和转换

数据清洗主要是不规则数据的合并和无意义数据的清除，就本范例，数据清洗问题比较关键，主要是因为医院科室和人员设置复杂，就要按照医院决策者的需要，把不具有分析意义的数据合并或者清除，达到数据分析的目的。

导入历史数据除了要完成数据清洗外，还要完成数据转换问题，主要是科室字典和费用代码归属的转换，根据两个系统数据字典的不同，首先建立数据转换表，此主题涉及科室代码转换表和费用代码转换表（表 3）：

表 3 科室代码转换表

旧科室代码	旧科室名称	新科室代码	新科室名称
10103013	心血管内科	101051	心血管内科
10105002	综合门诊	102011	综合内科
10106013	内科普通诊	101081	内科普通诊
10107013	内分泌内科	101031	内分泌内科
10108013	呼吸内科	101041	呼吸内科

4.5 OLAP分析实现

Broland Delphi5 或以上版本提供的DecisionCube控件组对OLAP分析具有良好的支持，用户可以通过BDE与数据仓库相连，使用DecisionCube、DecisionQuery、DecisionSource控件获取数据并构造多维数据立方体，用DecisionGrid、DecisionPivot控制来控制数据的切片、切块、上卷、下钻等操作。

4.6 评估、解释模型

医院业务收入动态表(表 4)：

表.4　住院业务收入动态表

年　份	出院人次（人次）x	出院者平均住院日（日）　y	出院者人均日费用（元 / 人日）　z	住院业务收入（万元）u
2004 年	11313	17.0	592.6	11397
2005 年	10969	16.6	610.4	11114
增长额	-344	-0.4	17.8	-283
增长率(%)	-3.04	-2.35	3.00	-2.48

模型解释：

住院业务收入动态分析

① 出院人次变动影响。影响额：$\triangle u_x=y_0z_0(x_1-x_0)=-346.5$；

影响程度：$\triangle u_x/x_0y_0z_0)=-3.04\%$。

② 出院者平均住院日变动影响。影响额：$\triangle u_y=x_0z_0(y_1-y_0)=-268.2$；

影响程度：$\triangle u_y/x_0y_0z_0)=-2.35\%$。

③ 出院者平均人日均费用变动影响。影响额：$\triangle u_z=x_0y_0(z_1-z_0)=342.3$；

影响程度：$\triangle u_z/x_0y_0z_0)=3.00\%$。

④ 三因素变动共同影响。影响额：$\triangle u_{xyz}=\triangle u-(y_0z_0\triangle x+x_0z_0\triangle y+x_0y_0\triangle z)=-10.6$；

影响程度：$\triangle u_{xyz}/x_0y_0z_0)= -0.09\%$。

计算得出经济数量关系满足下列等式：

增加额：$\triangle u= y_0z_0\triangle x+x_0z_0\triangle y+x_0y_0\triangle z+\triangle u_{xyz}$

即：-283=-346.5-268.2+342.3-10.6

增长率：$\triangle u/(x_0y_0z_0)= \triangle u_x/(x_0y_0z_0)+\triangle u_y/(x_0y_0z_0)+\triangle uz_{mp}/(x_0y_0z_0)+\triangle u_{xyz}/(x_0y_0z_0)$

即：-2.48%=-3.04%-2.35%+3.00%-0.09%

提示 2005 年住院业务收入比 2004 年减少了 283 万元。其中：出院人次减少影响程度为 3.04%，

减少收入 346.5 万元。出院者平均住院日影响程度为-2.35%，减少收入 268.2 万元。出院者平均住院日费用影响程度为 3.00%，增加收入 342.3 万元。三因素变动的综合影响程度为-0.09，减少-10.6 万元。

对生成的模型进行比较和评估，直到生成一个相对最佳模型，再对此模型用业务的语言加以解释，如果没有问题，可以对模型加以试验型的应用，如果有问题，再重复上面的数据准备和建立模型的过程，直到建立满意的模型为止。

4.7 结果展示

通过数据整合、窗体技术、动态报表、固定报表、多维报表、智能查询、任务调度、ETL 等技术，完成系统从数据整合、模型制作、数据抽取、系统展示等完整的建设过程，系统最终如下展示(图 3、图 4)：

图 3 住院收入仪表盘及多维分析系统

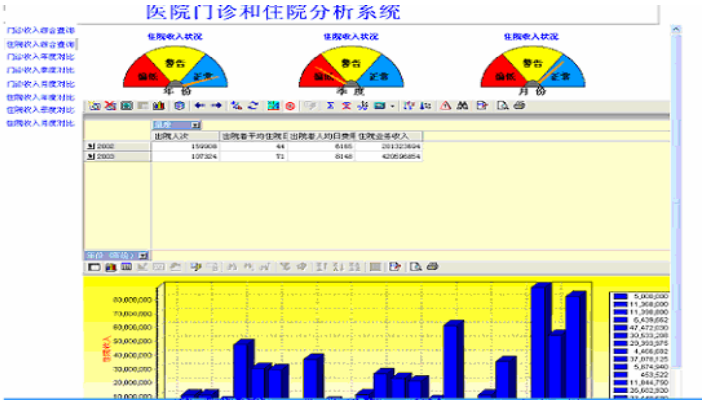

图 4 住院收入年度对比分析

以上的结果分析，显示了按照所列模型建立的数据仓库系统的结果，系统可以按照使用者的需要，可以任意调用不同的时间维度得到所需要报表、图示以及模型分析结果，不但能够快捷地查询医院收入的整体状况，同时还能够对影响收入因素的分解，对医院门诊和住院业务收入波动情况作出具有统计学意义的合理评价，并且从不同的时间维度反应医院业务收入情况，真正帮助医院管理人员对业务收入进行详细分析，找出问题所在，从而针对这些问题进行管理跟进。并且可以向下钻取，按照时间维度得到全年每月或者每季度的分析情况等等，有很强的实用价值。

5 结论

医院信息系统是一个以数据库为核心，以网络为技术支撑环境，具有一定规模的计算机化的系统，在系统内部按一定原则划分若干子系统，各子系统、分系统之间互有接口。

在整个系统建立过程中，数据仓库的建立是难点，因为医院信息系统经过近十年的发展，已经具有一定的规模，系统庞杂，平台众多、数据不一致和关系数据库之间的关联复杂等众多的因素，这些都是建立数据仓库必须要克服的困难。

统计分析是围绕特定的主题，对原始数据进行加工处理，是数据的深层次利用。利用统计模型得到的决策支持信息因为具有统计学意义，使分析结果更值得信赖。充分利用统计学模型是本研究的重点，正是因为充分利用了统计分析模型，使整个数据仓库系统脱离了简单的查询、报表和图形模式，有了更深层次的意义。

参考文献

[1] 尹辉,尹政,李政军.数据仓库技术及其应用[J].福建电脑, 2004,3:14-16

[2] 陈京民.数据仓库与数据挖掘技术[M].北京:电子工业出版社, 2002,95-96.

[3] 数据仓库技术在统计信息系统中的应用[J].科技情报开发与经济, 2005,8:223-224

[4] 邹俊卿,傅万明.电子病案的管理[J].医学研究生学报,2003,16(11):844-845.

[5] 数据仓库与数据挖掘. 华南农业大学信息学院计算机科学与工程系技术报告,2004.14-15

[6] David Hand,Heikki Mannila,Padhraic Smyth.数据挖掘原理[M].北京:机械工业出版社,2003.135-191

[7]Jose Samos, Felix Saltor, Jaume Sistac and Agusti Bardes. Database architecture for data warehousing: an evolutionary approach. In DEXA 1998: 746-756.