

营销系统

自我诊断

数据库

企业 (19)

计算机科学2000Vol. 27No. 11

# 基于数据仓库的营销系统自我诊断方法研究<sup>\*</sup>

袁晓洁 董震 康义楠

(南开大学信息技术科学学院 天津300071)

**Abstract** This paper discusses the application of the DW's multidimensional technology and presents the idea & method about using the technology to diagnose an enterprise's sale system. Then, the paper introduces how to build a model of sale DW and design analysis algorithms of the model. In the end, multidimensional multi-granularity fast analysis decision tools and the sale system's diagnosis prototype are proposed.

**Keywords** Data warehouse (DW), Dimension, Measure, Sale diagnosis, Purchasing power

## 一、引言

传统的决策支持系统(DSS)大都建立在数据库技术上,但随着数据量的迅速增长和用户查询要求的复杂化,数据库技术越来越难以胜任。这是因为数据库系统主要用于联机事务处理(OLTP),它面对的是大量细节数据,对分析处理的支持一直很弱。在事物型环境中直接构造分析型应用是一种失败的尝试。要提高分析和决策的效率和有效性,必须把分析型数据提取出来,按DSS重新组织,建立一个综合的、便于分析的处理环境,对大量异构数据进行综合的、多角度的、多层次的分析处理,为高层人员(管理者、执行者、分析家)提供决策信息,这就是数据仓库这一概念引入的原因。

迄今,业界尚未对术语“数据仓库”有一个统一的定义。在这里,我们引用公认的数据仓库概念的创始人——W. H. Inmon 给出的定义:“数据仓库就是面向主题的、集成的、稳定的、不同时间的数据集合,用以支持经营管理中的决策制订过程。”

数据仓库支持联机事务分析(OLAP)操作,OLAP是针对特定问题的联机数据访问和分析,它通过对信息的多种可能的观察形式进行快速、稳定一致和交互性的存取,允许管理决策人员对数据进行深入观察。为实现复杂的分析和可视化工作,数据仓库中的数据常以多维模式建立。维是人们观察数据的特定角度,维通常是有层次性的。常用的多维分析操作包括切片、切块、旋转、滚加(rollup)及分解(drilldown)。在数据仓库上实现OLAP服务,具有较高的性能,设计重点在于如何组织数据仓库中的综合性数据,如何满足前端用户的多维数据分析需要。

数据仓库技术具有很强的应用背景,它尤其适合采用计算机实施管理的企业、事业或商业部门。针对国家863计划支持项目“制造类型企业的自我诊断与评价方法研究”,我们开展了基于数据仓库的营销系统自我诊断研究,利用多维模式分析了企业销售信息,建立了销售数据仓库模型,设计了模型转化算法,提供了多维度、多粒度快速分析决策工具,实现了营销系统自我诊断软件原型系统。

## 二、营销诊断系统分析

### 1. 诊断思想

企业多年来积累下来的销售数据,只能为企业管理人员提供详尽的细节信息,帮助企业完成日常销售管理工作,但如何从这些细节数据中挖掘出潜在的决策信息,却一直不被人们重视,浪费了宝贵的信息资源。事实上企业销售历史数据中隐藏了大量的知识,通过对销售数据分析,完全可以发现诸如购买力趋势、价格因素影响、销售趋势等知识,它可以帮助企业管理者发现问题、调整销售策略、预测未来、作出各种决策,数据仓库技术为我们实现这一目标奠定了基础,通过建立销售数据仓库模型,可以将大量销售信息经预处理存入仓库中,利用OLAP快速处理历史数据,执行各种聚集操作,显示分析结果,给出趋势预测,发现有用的决策知识。

从企业营销角度来看,数据主要集中在产品信息、购货单位、销售信息三个方面,从企业提取原始数据信息,定义营销诊断数据仓库,按时间、产品、销售三个方面建立维表,在销售量、产品价格和销售利润三个方面建立测度,为用户提供维的粒度划分工具,用户可自行

<sup>\*</sup> 本课题得到国家“863”CIMS主题“应用基础研究”项目基金(No. 863-151-944-010)资助。袁晓洁 副教授,博士生,主要研究方向为数据仓库及CIMS技术。

对维进行层次和类的划分,从而实现多层次、多角度的查询、汇总,以便使企业了解销售状况的综合信息、发现销售弊病、提高产品销量。通过以上信息可为用户列出查询后得出的数据集,并绘制图表。为用户提供参考建议,以便于用户进行进一步的查询和分析。

### 2. 诊断流程

根据企业销售数据和数据仓库的建模思想,整个诊断过程应按如下流程设计:

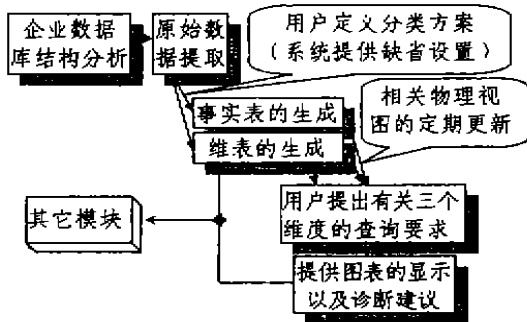


图1 营销诊断系统模型

首先分析企业原始数据库的格式,从产品、购货单位和销售信息三个方面提取事实表所需的数据,事实表的格式为(日期 ID,产品 ID,购货单位 ID,销售量,价格,毛利润,其它)。

其次提供接口可让用户自定义有关产品和购货单位信息的分类方案,并根据用户提供的分类方案产生维表。然后根据维表定期对各种粒度上的一些常用的物理视图进行刷新,系统按照各种表的使用频率,将其分别安置在各种不同性质的存储介质上。

根据下面列出的各种查询流程,引导用户进行分步的各种查询操作,并根据查询结果的建议,指示用户进入到相应的模块。

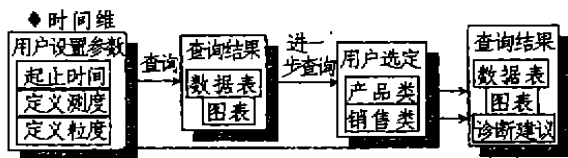


图2 基于时间维的查询

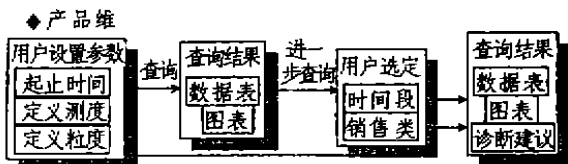


图3 基于产品维的查询

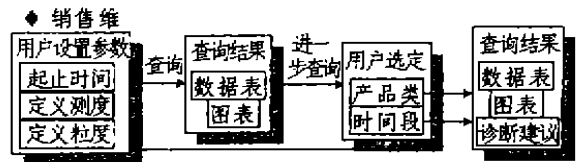


图4 基于销售维的查询

如以上各图所示,对于用户提出的各种查询要求(包括维度定义和测试的定义等),根据相应粒度的各种物理视图,得出聚合后的数据表和图表。若没有相应的物理视图,则转到原始的事实表中去查询。

最后根据得到的结果,建议用户进行进一步的查询,用户可以进一步选定产品类、时间段和销售类等,根据查询结果进行更深层次的分析,依诊断建议,或跳转到其它模块进行诊断。

### 3. 诊断知识

对于各种细化查询所得出的结果,需要进行相应的分析,才能得出诊断建议。我们考虑运用以下模型对查询结果进行分析。

考察购买力(Purchasing Power, 缩写为  $\lambda$ )。用字母 P 代表某一产品, T 代表某一段时间, R 代表某地区。用字母  $\Phi$  代表聚合操作。

(1) 地区维诊断: 对销售区域进行诊断

$$R_m \text{ 地区购买力 } \lambda_m = \frac{R_m \text{ 地区 } \Phi(R_m, T, P)}{\text{所有地区 } \Phi(R, T, P)}$$

$$R_m \text{ 地区平均购买力 } \bar{\lambda}_m =$$

$$\frac{\sum_{\text{对于每一种产品}} R_m \text{ 地区购买力 } \lambda_m}{n}, n \text{ 为产品种类数}$$

$$R_m \text{ 地区平均购买力 } 2 \bar{\lambda}_m =$$

$$\frac{\sum_{\text{对于一段时间}} R_m \text{ 地区购买力 } \lambda_m}{n}, n \text{ 为时间段数目}$$

若存在 P,  $R_m$  地区购买力  $\lambda_m + \lambda < R_m$  地区平均购买力  $\bar{\lambda}_m$ ,  $\lambda$  为浮动变量。

诊断结果: 产品 P 在地区  $R_m$  销售不畅。比较产品 P 在地区  $R_m$  和其它地区的单价、销售成本之间的差异。若单价偏高, 诊断是否在  $R_m$  地区的销售成本制定得不合理……

(2) 时间维: 进行季节性销售分析

$$T_r \text{ 时间购买力 } \lambda_r = \frac{T_r \text{ 时间 } \Phi(R, T_r, P)}{\text{所有时间 } \Phi(R, T, P)}$$

$$T_r \text{ 时间平均购买力 } 1 \bar{\lambda}_r =$$

$$\frac{\sum_{\text{对于所有地区}} T_r \text{ 时间购买力 } \lambda_r}{n}, n \text{ 为地区个数}$$

$$T_r \text{ 时间平均购买力 } 2 \bar{\lambda}_r =$$

$$\frac{\sum_{\text{对于所有产品}} T_i \text{ 时间购买力 } \lambda_i}{n}, n \text{ 为产品种类数}$$

若存在  $T_1, T_2$  时间购买力  $\lambda_1 - \lambda_2 < T_1$  时间平均购买力  $\bar{\lambda}$ 。

诊断结果:在  $T$  时间内,产品  $P$  在地区  $R$  销售不畅。

(3) 产品维:对产品销售进行分析

$$P_p \text{ 产品购买力 } \lambda_p = \frac{P_p \text{ 产品 } \Phi(R_i, T_j, P_p)}{\text{全部地区 } \Phi(R_i, T_j, P)}$$

$P_p$  产品的平均购买力1  $\lambda_p =$

$$\frac{\sum_{\text{对于所有地区}} P_p \text{ 产品购买力 } \lambda_p}{n}, n \text{ 为地区总数}$$

$P_p$  产品的平均购买力2  $\lambda_p =$

$$\frac{\sum_{\text{对于一段时期}} P_p \text{ 产品购买力 } \lambda_p}{n}, n \text{ 为时间段数}$$

若存在  $R_m, R_n$  地区购买力  $\lambda_p + \lambda < P_p$  产品平均购买力  $\bar{\lambda}_p$ 。

诊断结果:在  $R$  地区内,产品  $P$  在  $T$  时间内销售不畅。

根据以上的公式,可以挖掘出各种查询结果所代表的更深层次的含义,从而帮助决策人员进行正确的判断;并以此为依据,建议用户转到其它模块进行进一步的诊断,从而使企业诊断的整个体系进一步完整化。

#### 4. 用户接口设计

为使诊断系统更加灵活,为用户提供如下接口设计:

> 测度定义接口:可定义销售量、销售金额、销售毛利等;

> 时间定义接口:可定义起止时间、各个层次的时间粒度(年、月、旬等);

> 产品分类定义接口:可定义产品大类、小类等各种分类粒度;

> 销售区域定义接口:可定义地区大类、地区小类、企业性质等。

### 三、营销诊断系统设计

#### 1. 逻辑结构设计

根据以上系统分析,采用数据仓库中的雪片模式组织销售信息最为自然,其逻辑结构如图5所示。

从逻辑结构图可以看出,它是多层维表,一个事实表,而且在购货单位这一维上,体现了对于同一维度的不同分类方法。

虽然上述组织方式结构清晰、用户接口设计简单,但也有很大的缺陷,即当实现系统分析阶段的诊断思想时,系统开销相当大,主要时间和空间均浪费在实现

聚合操作的连接操作上。

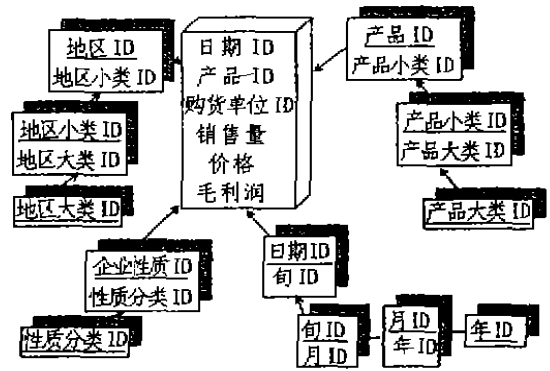


图5 雪片模式图

#### 2. 模型转换

为解决系统响应速度慢、开销大的问题,我们采用如下优化策略对系统进行了改进:

(1) 重新组织维表中的ID 对各层维表中的ID值进行组合并映射到事实表的ID值上,例如:产品维表结构如下:

产品维表(产品ID,产品名称,生产成本,其它)

其中产品ID与事实表中的产品ID相关联。而在用户对分类定义完成之后,系统自动生成如下结构的产品分类表:

表1

A	家电
A01	洗衣机
A01001	滚筒洗衣机
A01002	双缸洗衣机
A02	空调机
B	餐具
B01	燃气灶
.....	.....

由此,在“事实表-产品ID”字段中添加相应的位数,并根据产品类别和产品分类表来填充这些位。而对于购货单位的性质、所属地区、销售日期等维度,亦按照类似的办法处理。

这样一来,事实表中时间ID、产品ID、购货单位ID三个子段的长度将有所增加;但是此后所需进行的各种分类聚合操作的过程将大大地简化,只需按要求截取相应ID字段中的相应位,简单的分组求和即可。

(2) 将雪片模式转变为星型模式 由于对ID进行了映射,因此可将雪片逻辑结构转变为图6所示的星型结构,并将维表用“分类定义表”取代。

通过上述改进,提高了检索时的效率,并且简化了

系统的工作模式,从逻辑上讲,仍然实现了雪片模式的功能。

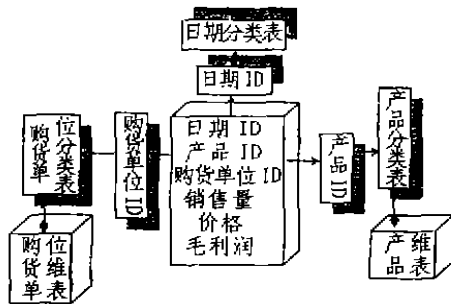


图6 星型模式图

由图6可以看出,各种粒度上的物理视图总共有  $4 \times 3 \times 3 \times 2 = 72$  张,而高频使用的物理视图共有  $3 \times 2 \times 2 \times 1 = 12$  张,据此,可以考虑将它们分置于各种不同的存储器中。

### 3. 决策支持工具

为了方便用户进行多维度、多粒度的检索,我们提供了一个可视化的分析描述工具,并制作成控件,其主要功能如图7。

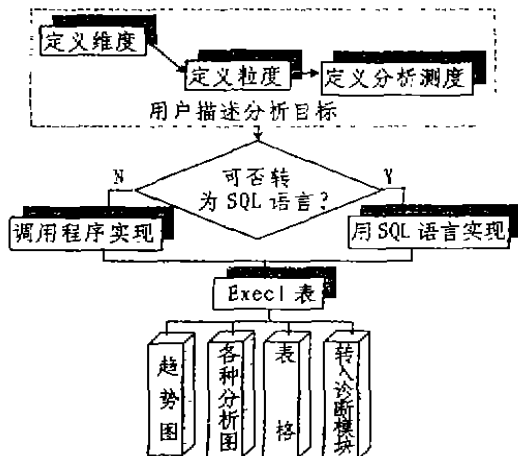


图7 分析描述工具流程图

### 4. 系统出错处理设计

为保证系统的安全性和完整性,只给予用户有限的操作权限,所有的物理表的结构对用户透明。表的内容和结构均为只读属性,三个维度的分类定义的编号亦对用户透明。

如果出现误操作,则系统给予提示,并不予执行,而不影响其它模块。

系统给予提示,定期追加事实表的内容,定期刷新物理视图,定期做系统安全性检查。

总结 通过数据仓库和OLAP技术,分析诊断企业现有的数据,为管理者提供决策支持是一项既有理论意义,又有很强实用性的工作。本文以企业营销为背景,探索了从系统分析到模型实现的全过程,其中接受用户自定义分层粒度、模型转换以及联机查询接口等思想发展了现有的数据仓库技术,所设计的原型系统已与合作单位配合开始验证与改进。我们有理由相信,基于数据仓库技术能够十分有效地反应企业目前的发展状况,它对决策支持系统的发展起着关键的作用,这项研究会现代化企业中发挥更大的作用。

### 参考文献

- 1 Piatetskey-Shapiro G, Frawley W J. Knowledge Discovery in Database. AAAI/MIT Press, 1991
- 2 Chaudhuri S, Dayal U. An Overview of Data Warehousing and OLAP Technology, VLDB'96
- 3 Chen M-S, et al. Data Mining: An Overview from Database Perspective IEEE Trans Knowledge and Data Engineering, 1996
- 4 Zhuge Y, Wiener J, Molina H G. Multiple view consistency for data warehousing. In: Proc. of the Int. Conf. on Data Engineering, Birmingham, 1997
- 5 韩路. 企业诊断指南. 新华出版社, 1989. 9

**《计算机科学》成为中国计算机学会会刊**  
**是《计算机科学》杂志社的幸事**  
**是广大读者的幸事**  
**是作者的幸事**