

基于数据仓库的指标体系研究

刘黎志

(武汉工程大学 计算机学院, 湖北 武汉 430073)

摘 要:政府及行业一般用报表的方式收集数据,报表中的各个数据项就是指标,不同的政府部门、行业所需要的指标的个数、种类是不一样的。指标体系用一种统一的方式来描述指标结构,因此可以为不同的应用生成不同的指标。以报表方式收集的指标数据往往是海量的,并且与时间密切相关,因此将指标数据存放在数据仓库中是合理的。数据仓库中存储的海量指标数据又为 OLAP 分析及进一步的数据挖掘奠定了基础。文中讨论构成指标体系的关系结构,描述了实用指标的形成过程,说明了指标体系多维数据集中的共享维度和事实数据表,最后给出了基于数据仓库的指标体系的应用。

关键词:指标体系;实用指标;数据仓库

中图分类号: TP317.1

文献标识码: A

文章编号: 1673-629X(2007)01-0196-03

Research of Indicator System Based on Data Warehouse

LIU Li-zhi

(Computer School of Wuhan Engineering University, Wuhan 430073, China)

Abstract: Government and industry always use form to collect data, the data item of form is called indicator. Different government departments and industry need different types of indicators. Indicator system describes indicator structure in a general way, so it can create different indicators for different applications. The indicator data size is often in great capacity and closely related to the time, so puts the data in the data warehouse is reasonable. The great indicator data stored in data warehouse also sets up the base for OLAP and data mining. In this paper, discuss the relational structure of the indicator system and the combining process of practical indicator. Then explain the shared dimension and fact data table within indicator system multi-dimension dataset. Finally give the applications of indicator system based on data warehouse.

Key words: indicator system; practical indicator; data warehouse

0 引言

政府及行业的决策支持、商业智能、信息分析预测系统的基本数据来源是指标,指标作为一种统计值为这些系统提供决策依据。实际上指标在上述各种应用中是一个共性对象,因此构建一个能在不同系统中通用的指标体系是十分有意义的。

数据仓库是支持管理决策过程的、集成的、与时间有关的、持久的数据集。数据仓库为不同来源的数据提供了一致的数据视图,与数据挖掘、联机分析处理等数据分析技术相结合,可为用户提供灵活自主的信息访问和丰富的数据分析与报表功能,使数据得到充分利用^[1,2]。利用数据仓库技术使得指标体系可以提供准确可靠的统计数据,快速地进行指标数据汇总,对大量指标数据进行有效的时空序列分析处理,形象直观地表现数据间的规律,利用数据挖掘进行指标分析和趋势预测。

1 指标体系

指标体系是构成实用指标维度的底层关系结构,指标体系包含构建它所需的基本关系实体,并在基本关系实体的基础上形成实用指标。

1.1 基本关系实体

(1)指标:度量国家经济发展、生产效益、人口情况、自然资源等各方面的统计值。

(2)指标分类:根据指标的统计领域确定其分类,一般分为工业、农业、国民经济、人口等。

(3)指标小类:在某个指标类别中,为更好归纳不同统计目的指标而自定义的分类。

(4)指标度量单位:指衡量指标值的单位,如米、美元、升等。

(5)指标度量单位分类:按度量单位的用途而进行的分类,如长度、质量、面积等。

(6)阶码:指同一度量单位的不同度量范围。对度量单位元,若阶码为3,则表示千元,阶码为9,则表示亿元。

(7)分组:指标反映的是某个统计值的综合信息,如“亏损企业数”指标,可能反映的是某个地区{特大型企业、大型企业、中型企业、小型企业}的亏损数的合计值,也可

收稿日期:2006-04-12

基金项目:科技部科技型中小企业创新基金(03C26214201044)

作者简介:刘黎志(1973-),男,湖北武汉人,讲师,研究方向为基于网络的计算机应用、商业智能、数据仓库及数据挖掘。

能反映的是某年{国有企业、私有企业、三资企业}的亏损数的合计值。具体确定指标统计对象的集合称为分组, {特大型企业、大型企业、中型企业、小型企业……}为企业规模分组, {国有企业、私有企业、三资企业……}为工商登记类型分组,可能的分组还有性别、学历、营业状况、本末期值等。

(8)实用指标:确定指标的具体统计对象及度量范围的统计值,实用指标是指标、分组、阶码的有效笛卡儿集。

1.2 实用指标的形成

定义:

指标元组 $i = \langle \text{指标码}, \text{指标名称}, \text{指标父结点码}, \text{指标兄弟结点码}, \text{指标度量单位码}, \text{指标所属小类} \rangle$

指标 $I = \{i_1, i_2, i_3, \dots, i_j\}$

分组元组 $g = \langle \text{分组码}, \text{分组名称}, \text{分组依据}, \text{分组父结点码} \rangle$

分组 $G = \{g_1, g_2, g_3, \dots, g_k\}$

阶码元组 $n = \langle \text{度量单位码}, \text{阶码值}, \text{阶码名称} \rangle$

阶码 $N = \{n_1, n_2, n_3, \dots, n_m\}$

实用指标元组 $pi = \langle \text{实用指标码}, \text{指标码}, \text{分组码}, \text{阶码} \rangle$

实用指标 $PI = \{pi \mid pi \in \langle i_1, g_1, n_1 \rangle, \langle i_2, g_2, n_2 \rangle, \dots, \langle i_j, g_k, n_m \rangle\}$, 其中 $\langle i_i, g_k, n_m \rangle$ 为有效组合, j, k, m 为自然数。

将实用指标定义为指标、分组、阶码的有效笛卡儿集提高了构造指标体系的灵活性,特别是分组的引入可以极大地减少系统中指标的个数。实用指标必须是指标、分组、阶码的有意义的组合,实用指标一般是在系统初始化时根据需要收集的具体指标统计值,由系统管理员生成,报表设计工具可根据实用指标设计收集指标数据及汇总指标数据的表单格式。由于指标体系结构的灵活性,使得实用指标在使用中很方便系统管理员维护。

2 建立指标体系数据仓库

2.1 指标体系维度

(1)时间维:时间维的设计根据系统的不同应用域而定,对于应用指标体系的系统存在一个时间粒度控制的问题。粒度是维划分的单位,体现着数据单元的详细程度和级别。数据越详细,粒度越小,级别越低。数据综合程度越高,粒度越大,级别越高^[3]。时间维粒度的确定一般和报表的报送制度一致,若报表每月报送一次则时间维可为年、季、月,若报表每天报送一次则时间维可为年、季、月、天。时间维是典型的星型层次架构。

(2)地区维:反映指标体系的地区信息,地区是具有父子层次架构的维度模型。指标体系的地区维度一般表现系统所应用区域的树型结构。将指标体系应用到信息产业,地区维表现为企业、地市、省、信息产业部;应用到农村扶贫统计,地区维表现为村、乡镇、地市、省。数据仓库按

地区的层次结构逐级汇总得出指标的聚合值,但其默认方式只是从叶结点开始逐级向上汇总。若应用系统需要在非叶结点输入实用指标值,例如扶贫统计中乡镇同样也要上报指标数据报表,则需设置地区维,允许其在非叶结点上接收数据。

(3)实用指标维:指标、分组、阶码构成实用指标维的雪花型层次模型,以指标名称为第一级别、分组名称为第二级别。实用指标维的雪花型层次模型如图1所示。

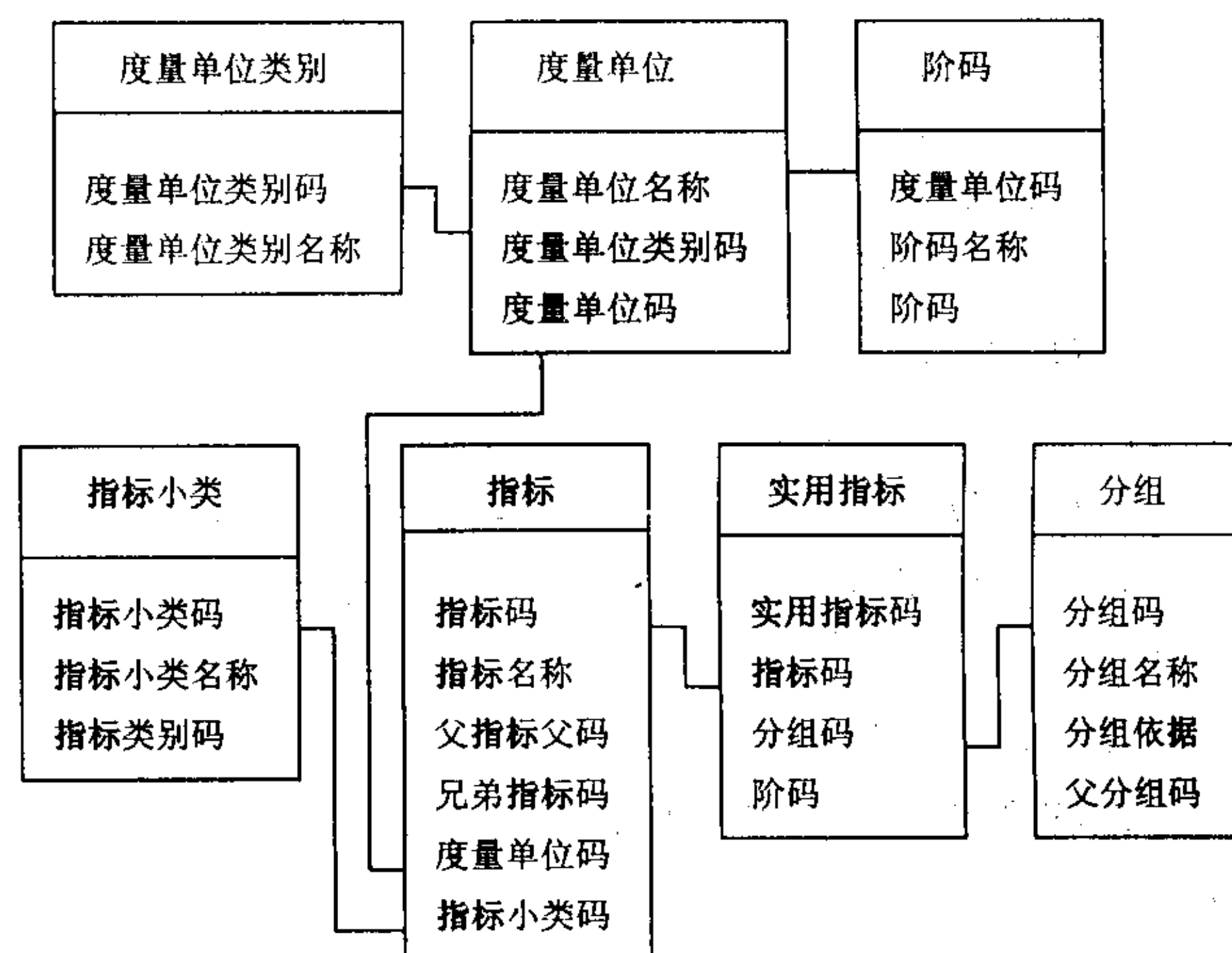


图1 实用指标维雪花型层次模型

(4)指标类别维:确定指标所属的类别,可以设定为共享维度或虚拟维度。若设定为共享维度,则需指定其底层指标类别数据表;若设定为虚拟维度,可为实用指标维度指定指标类别成员属性。

2.2 指标体系多维数据集

数据仓库多维数据集有星型模型和雪花模型两种。星型模型的设计比较简单,是基于关系型数据库的、面向OLAP的一种多维数据模型的数据组织形式。星型模型由事实表和多个维度表组成,通过使用一个包括主题的事实表和多个包含事实的非正规化描述的维度表来执行指标体系查询,由于数据仓库在存储事实数据表时会自动计算指标数据的聚合值,因此可获得比一般SQL语句分组查询更高的查询性能。星型模型的中心是指标数据,对应实用指标事实数据表,四周是访问的角度,对应维度表,每一个维度表通过一个关键字直接与事实表关联。事实表中每条记录都包含指向各个维度表的外键和实用指标度量值。雪花模型与星型模型的区别仅在于对维度表的描述是正规的,对实用指标维的描述就是正规的。指标体系雪花多维数据集如图2所示。

对事实数据表和维度表所作的修改在对其数据进行刷新后可以从底层关系表中反映到多维数据集中,对数据集的刷新往往有一定的逻辑规则及时间规则。DTS(数据转换服务)以工作流的方式实现数据集刷新的逻辑规则,数据仓库管理员可以规定多维数据集的刷新次序,并在数据集刷新完成后收到DTS发送的完成消息^[4]。数据集刷新的逻辑规则以DTS包的形式组织,数据仓库管理员设置DTS包的时间规则完成对DTS包的调度。

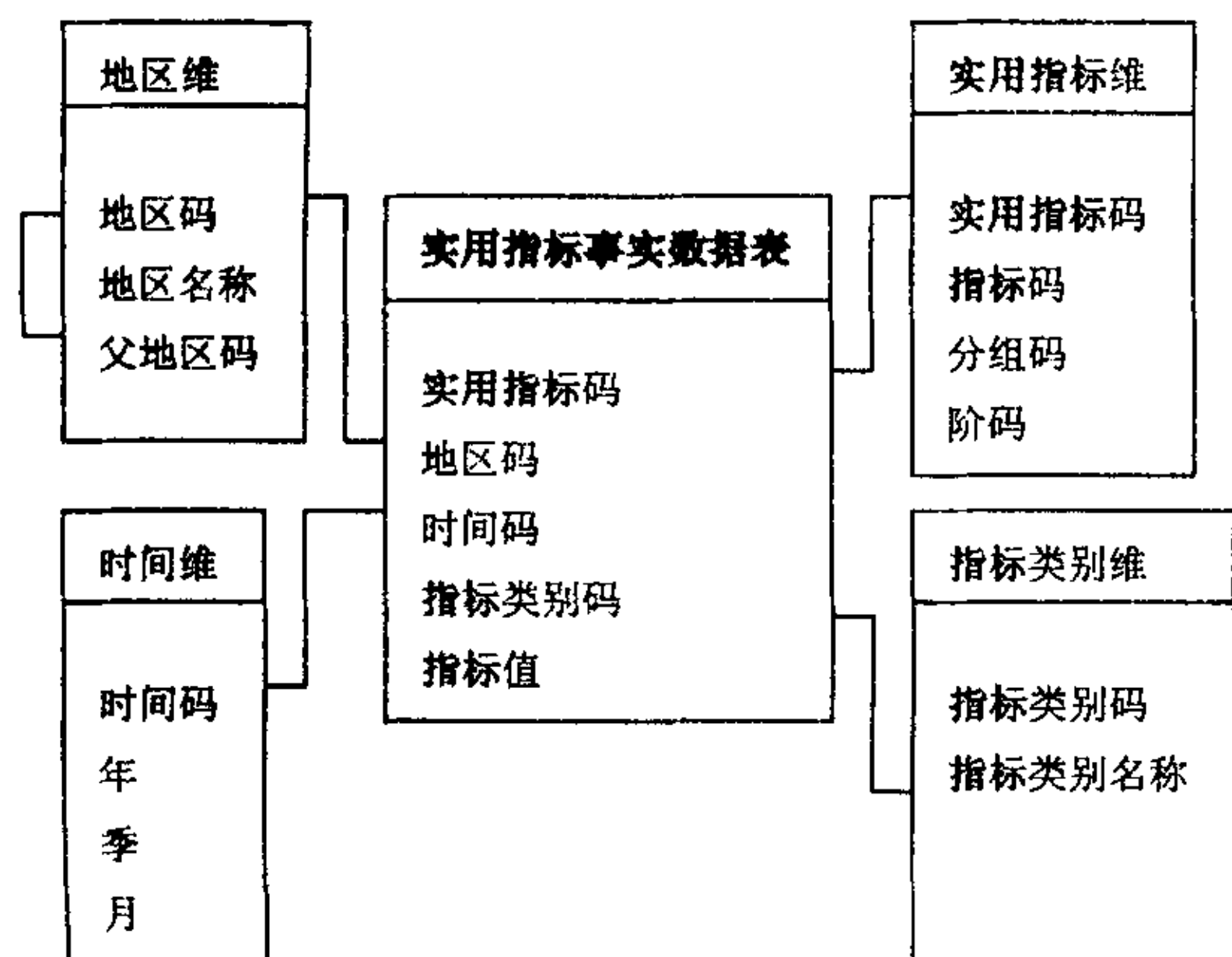


图 2 指标体系多维数据集

2.3 阶码转换

由于实用指标的阶码可能与汇总报表中的阶码不一致,所以向多维数据集添加数据记录的指标值时,程序逻辑要根据实用指标的阶码对用户输入的指标值进行转换,公式为:

令: $M = \text{实用指标阶码}$

存储到事实数据表的指标值 = 用户输入的指标值 $\times 10^M$

从多维数据集中查询的指标值在向用户展示时同样也要进行阶码转换,公式为:

展示给用户的指标值 = 存储到事实数据表的指标值 $\times 10^{-N}$ $N = \text{用户指定的阶码}$

3 基于指标体系的应用

使用指标体系的系统的工作流程一般为:客户端利用报表设计工具设计收集指标数据的报表,报表的格式通常有行业标准;将报表通过 Web 服务器发布;地区维的叶结点、非叶结点在规定报告期内填报数据,从而完成数据收集;在一个规定报告期结束后,对数据进行汇总,按规定出汇总报表;对系统历史数据进行查询、分析、比较、预测,为决策提供依据。指标体系为整个应用过程构建了数据存储层,更为重要的是对于系统开发而言,指标体系可以设计为通用组件进行维护。由于指标系统具有针对不同的应用领域生成不同实用指标的能力,所有可以大大减少系统开发、维护工作量,从而降低开发成本。

基于数据仓库的指标多维数据集允许对数据进行切片和切块,可以很容易地得到某个地区、某个时间段、某个指标类别的某个指标的汇总和明细信息。利用 MDX(多维表达式)来查询指标体系多维数据集时,地区、时间、指标类别维度一般作为切片器维度,而实用指标维度一般作

为轴维度。由 MDX 查询返回的数据元组集合可以和基于 Web 的多维数据集浏览控件绑定,从而使得用户可以根据维度进行查询条件筛选,可以深化以看到数据的细节,还可以浅化以看到汇总数据。通过 MDX 查询返回的二维数据集向用户展示直观的指标数据关系,并为制作汇总指标报表及直方图提供数据源。由于多维数据集在存储时使用聚合预先计算汇总指标数据,所以可以极大地提高汇总指标查询的效率和缩短响应时间,这对于系统应用来说是非常重要的。利用钻取功能,又可以反向获得构成汇总指标的底层实用指标记录,从而得到汇总指标的详细成分。

构建基于数据仓库的指标体系的意义不仅在于查询汇总指标,更重要的是在海量历史指标数据的基础上进行指标分析和趋势预测^[5]。关联分析可以得到指标值之间的隐藏关联网;预测可以在历史指标数据中找出变化规律,建立模型,并由此模型对所关心的指标未来特征进行预测;偏差分析可以发现指标存在的异常情况并利用偏差检验寻找观察结果与参照之间的差别。利用数据挖掘技术发现指标间的关联性和发展趋势,为应用指标体系的业务系统提供了更好的决策支持和分析能力。

4 结束语

指标体系的存储、管理、共享在政府决策支持、商业智能、经济发展分析与预测等各方面将发挥越来越重要的作用,基于以上领域的计算机应用系统都无一例外以大量的统计指标作为数据基础。通过近几年的实践总结出的基于数据仓库的指标体系在指标的底层存储、指标体系多维数据集的构建、基于指标体系的应用等方面做了一些有益的探索。

参考文献:

- [1] 张志军,夏传良. 基于数据仓库的企业管理决策支持系统[J]. 计算机应用与软件,2005,22(6):65-66.
- [2] Inmon W H. Building the Data Warehouse[M]. 2nd Edition. New York: John Wiley & Sons Inc,1996.
- [3] 冉春玉,谷川. 税务决策支持系统数据仓库的设计[J]. 计算机应用,2005,25:158-159.
- [4] Srivastava J, Chen P. Warehouse Creation: A Potential Roadblock to Data Warehousing[J]. IEEE Transactions on Knowledge and Data Engineering,1999,11(1):118-126.
- [5] Kantardzic M. Data Mining: Methods, Tools and Techniques[M]. [s.l.]: IEEE Press and John Wiley,2002.

(上接第 195 页)

IEEE Transactions on Knowledge And Data Engineering, 2003,15(4):840-854.

- [4] 贾瑞新,刘永军,赵晓颖. P2P 网络模型下发现机制的研究和实现[J]. 北京工业大学学报,2005,10(5):35-37.

- [5] 陈瑛. 一种分布式流媒体系统的应用初探[J]. 大众科技, 2006,15(6):97-98.

- [6] 李纲,岑雄鹰,陈叶芳. 一种基于 P2P 点组技术的流媒体协作计算[J]. 计算机应用与软件,2006,18(2):38-44.