面向市场分析预测的数据仓库技术应用与研究*

刘卫东 冯建华 王令赤 郑 彤

(清华大学计算机系 北京 100084)

摘要 本文针对市场分析预测对数据的要求,探讨了数据仓库技术在"玉烟系统"中的应用,详细说明了"玉烟系统"中数据仓库的建立、数据转换与校验、数据规范化方法及其对分析预测的支持策略,提出了应用数据仓库的一些基本要求.

关键词 分析预测,数据仓库,转换和校验,规范化.

"玉烟系统"是一个面向市场进行分析预测的多媒体智能数据库系统.它根据专家知识和与市场有关的数据,分析市场的现状,预测市场发展趋势,以语音、图形、简表、卡片等多种形式形象地输出,为决策提供支持.同时,系统提供丰富的查询功能,帮助决策者了解用于分析预测的详细数据.

"玉烟系统"的市场分析预测建立在大量数据的基础上.这些数据来源多种多样,有各种市场调查数据,也有各地经济情况的数据.由于数据来源的多样性,造成数据形式、数据格式的多样性.

系统数据的这些性质给分析预测使用数据带来了极大的困难.同时,数据随时间的推移不断增加,有些甚至带来数据内容的不一致,因此,必须对数据进行重新组织和校验.而且,分析预测子系统应能面向所有数据进行分析和预测,并适应数据的不断增加和更新,这就要求系统和数据相对独立.

在"玉烟系统"中,我们采用数据仓库技术,较好地解决了上述问题.

数据仓库是为管理者决策过程提供支持的数据集合,它面向专题和时间组织数据,并对数据进行集成.[1,2]它的主要宗旨是通过通畅、合理、全面的信息管理,来达到对管理决策的支持.与联机事务处理相比,是完全另一种类型的信息管理方式.

很清楚,数据仓库要能够从源数据库中提取数据,并与来自其他源的数据合并,这个过程就是从来自不同源的数据里分离出管理决策需要的业务信息,供前端用户使用.[4,5]所以,一般的数据仓库体系结构的主要构件之一就是一个用于决策支持的只读数据库,它有下面

* 作者刘卫东,1968年生,讲师,主要研究领域为数据库应用及信息管理系统.冯建华,1967年生,讲师,主要研究领域为数据库应用及信息管理系统.王令赤,1966年生,讲师,主要研究领域为数据库应用及信息管理系统.郑彤,1971年生,硕士生,主要研究领域为数据库应用.

本文通讯联系人:刘卫东,北京 100084,清华大学计算机系

本文1996-01-30收到修改稿

4 个特征:(1) 数据是从源系统、源数据库和源文件里提取出来的;(2) 来自源的数据在进入数据仓库之前要进行转换和校验;(3)用于决策支持的数据放在一个独立的只读数据库中,可以将其简单地称为"数据仓库";(4) 用户通过前端工具和应用程序来访问数据仓库.[3]在这里,需要强调的 2 点是:(1) 不论数据来源于何处,进入数据仓库之后都具有统一的数据结构和编码规则,数据仓库中的数据具有一致性特点;(2) 数据仓库是一个信息源,它只是为在其上开发的决策支持系统等提供数据服务,因此它应是只读数据库,一般不轻易做改动,只能定期刷新.

数据仓库的组织方式有下面 2 种:(1) 按横向对数据进行分类存储.数据仓库存储的信息是面向专题来组织的,它根据所需要的信息,分不同类、不同角度等专题把数据整理之后存储起来;(2) 按纵向对数据进行分类存储.数据仓库中要有一处专门用来存储 5～10 年或更久的历史数据,以满足分析预测之用的数据需求.

数据仓库中的信息存储,是根据对数据的不同深度处理来分成不同层次的.其结构一般划分为以下几个方面:(1) 详细数据层.包括历史数据和最新数据,它们是分析预测的基础数据;(2) 不同程度的归纳总结信息层.可包含多个层次,根据所需分类和归纳的不同深度而定.如按周、月、季和年统计的数据;(3)专业分析信息层.专业分析的结果,如统计分析、运筹分析、时间序列分析以及表面数据的内在规律分析等;(4) 结构信息.数据仓库的内部结构信息,反应各种信息在数据仓库中的位置分布和处理方式等,以便检索查询之用.

1 "玉烟系统"数据仓库的结构和功能设计

在"玉烟系统"中,我们将其数据仓库体系结构确定如图 1.

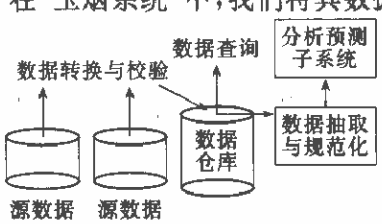

图1 玉烟系统数据仓库体系结构

图 1 中"数据转换和校验"模块完成数据采集和定期刷新功能.为保证数据的一致性,还需利用知识、规则对原始数据进行校验.经校验无误后,将数据按时间、专题的不同存放于数据仓库中.

"数据抽取与规范化"模块主要是为分析预测等应用程序提供获取数据的通用接口.它遵照应用程序的要求,从数据仓库中抽取应用程序所需数据,并规范化为应用程序所需要的格式,提供给应用程序使用."数据查询"模块是获取数据的通用接口,它为前端用户提供访问数据仓库的工具,并将查询结果以多媒体的形式输出.

2 数据转换和校验

企业数据仓库中的归档数据可能来自企业内部的各个部门,也可能来自企业外部.这些数据可能存储在数据库中,也可能存储在其他类型的数据源中.由于数据来源、数据类型和存储方式的不一致,来自各种数据源的数据必然存在着数据冗余和不一致,甚至是自相矛盾的情况.简单地把各个数据源中的数据转换到数据仓库中就存在着上述问题.这就要求在数据转换和校验过程中发现这些冗余和不一致,进而消除之,并形成单一、一致的数据.建立在这种数据基础上的分析预测才是可靠的.

数据转换和校验的总体结构如图2所示.数据转换程序根据在数据中所描述的映射关系和各种整理规则,将数据由各种数据源转换到数据仓库中,并对数据进行规范化处理;数据校验程序根据在元数据中所描述的校验规则和专家知识,消除源数据库中存在的不一致和不合理的数据;元数据管理程序用于对数据仓库中的元数据进行管理.本节就数据转换和校验中的主要问题进行讨论.

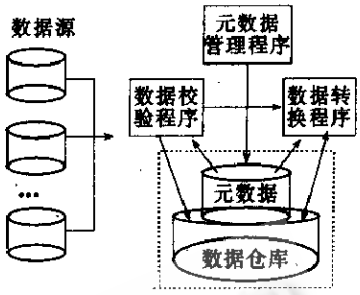

图2 数据转换和校验的体系结构

2.1 元数据

元数据可以称为数据中的数据.前端用户并不需要知道数据仓库中的元数据.元数据用于描述数据来自哪里,并且怎样变成目前的形式.

为了便于进行数据转换和校验,在元数据中需要描述以下信息:数据源信息、数据类型映射关系、编码信息和数据分类信息等.

2.2 数据转换原则

在现实世界中,数据仓库中的归档数据往往来自不同的数据源.在这些不同的数据源中,描述同一数据的数据类型或编码系统可能各不相同.所以,在把这些数据转换到数据仓库中时,还必须对被转换的数据进行整理、归并,对编码进行规范化等等.总之,数据转换程序的核心工作是进行数据整理,它主要依据下述4条原则进行:(1) 数据类型规范化.把不同数据源中描述同一类数据的不同数据类型统一映射成数据仓库中的同一数据类型;(2) 按专题和时间对数据进行归并处理.对来自异种数据源的同类数据,首先按专题进行归并,在此基础上再按时间进行归并,这将有助于进行分析预测;(3) 数据对象命名规范化.给来自各个数据源的数据对象重新命名,使依据对象名就能确定该类数据所属的专题和时间区间,这将极大地方便数据查询;(4) 编码规范化.不同的数据源中,可能采用了不同的编码系统描述数据.所以,在把数据源中的数据集成到数据仓库时,必须采用统一的编码,形成具有统一编码的数据.

2.3 数据校验规则

经数据转换程序处理后的数据基本上处于一致性状态,但仍有可能存在着不合理和不一致的数据.而这些数据只能通过数据校验规则和专家知识进行消除.校验程序的核心部分是校验规则和专家知识.我们主要采用以下规则和知识对数据进行校验:(1) 根据不同数据之间的计算关系校验.在数据仓库中某些数据之间往往存在着一种计算关系,可根据这种计算关系来纠正某些数据间的不一致性;(2) 根据同类历史数据校验.通过分析同类历史数据,可以归纳出该类数据的某些特征(比如,该类数据必须取值于某区间).如果新近被转换的数据不符合这些特征,它很可能是错误数据,需加以纠正;(3) 根据同类数据的整体特征进行校验.根据同类数据在同一时期的整体特征(比如平均值)进行校验.如果发现某一数据远远偏离该类数据的整体特征,则该数据可能是错误数据,需加以纠正;(4) 根据专家知识进行校验.在长期的实践中,一些专家总结出了某类数据必须具备某些特征(比如,某类东西的价格不会超过多少,也不会低于多少),那么违背这些特征的数据就可能是错误数据,需加以纠正.

3 数据抽取与规范化

建立数据仓库的一个重要目的，就是为应用程序使用大量数据进行分析、预测等提供方便，为用户在数据仓库中获取信息提供基础.

面对数据仓库中庞大的数据，虽已按时间、专题进行了重新组织，但若是想从中取出某些特定的数据进行分析预测，还是十分困难.①数据仓库中数据按专题和时间进行组织，但分析预测可能涉及到对不同专题的某几个因素之间的关系进行分析，或是对一段时间的某些因素的时间序列的预测，这些都涉及到跨时间、跨专题的数据抽取；②数据仓库的数据在不断刷新，如何使分析预测子系统适应这些新的数据？另外，相同的分析预测方法，可针对相同性质的不同因素进行分析预测，数据仓库应提供数据规范化手段，以满足分析预测子系统的输入接口要求，使分析预测与数据相对独立.

为满足上述要求，“玉烟系统”提供通用数据抽取接口，使分析预测子系统能方便地从数据仓库中查询到不同时间、不同专题的一批数据，并将数据规范化为应用程序的接口形式.

“玉烟系统”中的数据，经过数据转换与校验，已按时间、专题进行组织，通用数据抽取接口根据被抽取数据的时间、专题即可查询到相应数据；对查询结果的规范化，可通过描述规范规则进行.

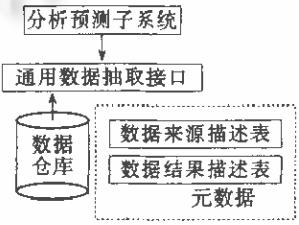

图3 通用数据抽取接口工作原理图

基于上述考虑，通用数据抽取接口的工作原理如图3所示.该通用数据抽取接口主要通过下述2个系统表进行工作：(1) 数据来源描述表.本表描述所抽取数据在数据仓库中的专题和数据域.若是抽取一定时间序列的数据，还可在本表描述抽取该项数据的个数，数据抽取接口可自动抽取多个数据.被抽取数据的时间由接口参数描述；(2) 数据结果描述表.本表描述应用程序的输入数据结构，通用数据抽取接口将按照此结构提供数据.

分析预测子系统根据本身的数据要求，利用数据仓库中的元数据，整理出对数据来源及数据结果的要求.通用数据抽取接口按照这些要求为分析预测子系统提供数据接口.

“玉烟系统”通过上述通用数据抽取接口，解决了数据仓库与分析预测程序的数据接口问题，为系统的灵活性奠定了基础.

4 数据查询

在数据仓库的体系结构中，用户通过前端工具和应用程序访问数据仓库中的数据.这种访问，要达到下面3个目标：(1) 用户可以很方便地查询到数据仓库里各个层次的数据；(2) 查询输出的结果要丰富多彩，最好是集声图文于一体的多媒体形式；(3) 输出窗口上的操作要实用、简单.

在“玉烟系统”中有一个方便、通用的数据查询系统，该系统根据用户指定的专题、子专题和时间信息来查询数据仓库里的数据.所以，在数据查询系统的界面里为用户提供了选择专题的专题菜单和选择时间的时间标尺.这样的话，用户就可以从横向和纵向2个方向查询

到数据仓库中详细数据层、专业分析信息层等各个层次的数据.之所以能够这样做是基于以下2点:(1)对数据仓库中的任何一张数据表,采用统一的规则为其命名.例如,用数据表名的前3个字符和最后2个字符表示时间,其中前3个字符表示年份,最后2个字符表示月份或季度等,而中间的其他字符表示专题、子专题等;(2)在数据仓库中有一张数据来源描述表,它描述了其他数据表的专题和时间信息,指出了其他数据表在数据仓库中的位置和处理方式等.基于以上2点,数据查询系统就能够根据用户指定的专题和时间信息唯一确定数据仓库中一张数据表的名字,从而根据表名和附加的查询条件查询其中的信息.

前面已经提到,"玉烟系统"的数据仓库中,数据来源描述表不仅描述了其他数据表的位置分布,而且还指出了对其他数据表的处理方式.根据这些信息,数据查询系统的输出集图表文于一体,共有以下5种形式:(1)简表用来显示用户查询到的数据表中所有记录的表格,由标题、表头和表身构成;(2)卡片用来显示简表中所有数值域的总计或平均值,或者显示某一条记录的所有域值;(3)统计图形以三维柱图、饼图、折线图、雷达图、面积图等图形显示简表中指定域的信息;(4)可视化模型也是一种三维图形,但它比统计图形要复杂得多.它用实体表示简表中指定域的数值大小、类别、地理位置及相互关系等.在玉烟系统中实现了下面4个可视化模型:管道模型、板条模型、水槽模型和地图—树根模型.关于这些模型的详细信息,请参阅本期的有关论文;(5)专题信息是指在屏幕的背景图上,以图符、颜色或台、柱等方式显示简表中指定域的信息.

在输出窗口中,简表是最主要的窗口,在其上为用户提供的操作要求简单、实用."玉烟系统"中主要实现了以下简表操作:简表排序、简表内列拖动、简表间列拖动、着重显示简表上某条记录、打印简表等.对于统计图形、可视化模型和专题信息,都可以进行三维漫游操作.另外,在各个窗口上还有诸如打开、关闭、滚动、缩放及移动之类的操作,在此不再赘述.

5 结　论

从我们在玉烟系统中的实际经验来看,要建造一个实用的数据仓库,必须首先解决以下4个主要问题:(1)对大量的不同格式、跨越不同软件平台的企业一般运行数据要能及时、有效地访问到;(2)对访问到的基本数据要能进行有效的分类、合并、归纳、整理以及深层次的分析和处理;(3)必须具备一个合理的数据存储结构;(4)建造的数据仓库具有开放性,使其成为众多信息系统的物理信息源.

致谢　清华大学计算机系博士生刘宁宁同学和硕士生刘磊同学也参与了本文的研究工作,在此对他们一并表示感谢.

参考文献

1　Inmon W H. Building the data warehouse. A Wiley—QED Publication, John Wiley & Sons, Inc., 1992.

2　Inmon W H, Hackathorn R D. Using the data warehouse. A Wiley—QED Publication, John Wiley & Sons, Inc., 1994.

3　Poe Vidette. Data warehouse: architecture is not infrustructure. Database Programming & Design, July 1995, 8(7):24～31.

4 The Editor. Data—warehouse or workhorse? Database and Network Journal, July 1995,25(3).

5 Jenyao Chung. Scalable paralled query server for decision support applications. Eleventh International Conference on Data Engineering, March 6～10 1995 IEEE, 1995. 186～187.

THE APPLICATION AND RESEARCH ON MARKET ANALYSIS AND FORECAST—ORIENTED DATA WAREHOUSE TECHNOLOGY

Liu Weidong Feng Jianhua Wang Lingchi Zheng Tong

(*Department of Computer Science Tsinghua University Beijing* 100084)

Abstract This paper is intended to explore solutions to the data requirements of a market analysis and forecast system though application of data warehouse technology. It presents several aspects related to the building of a data warehouse, including data extraction and check—up, data normalization, and a mechanism to support market analysis and forecast. The paper also summarizes some basic requirements for applying data warehouse.

Key words Analysis and forecast, data warehouse, transformation and check, normalization.