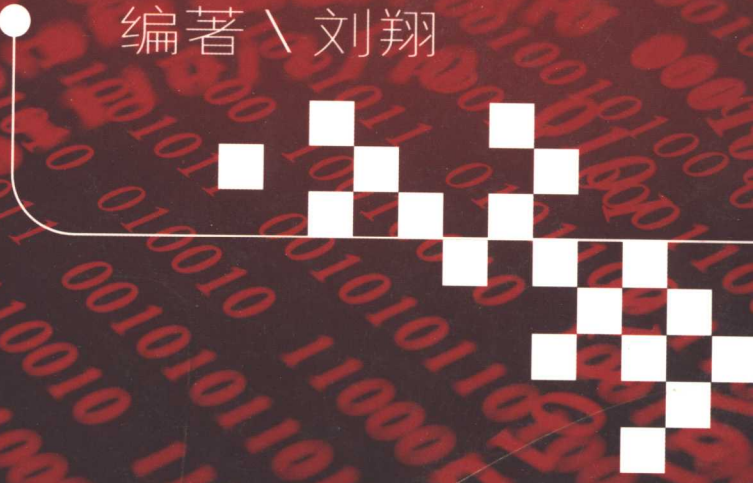


企业信息化丛书

数据仓库与 数据挖掘技术

编著 \ 刘翔



上海交通大学出版社

企业信息化丛书

数据仓库与数据挖掘技术

刘 翔 编著

上海交通大学出版社

内 容 提 要

本书以企业信息化为基点,从理论与实际应用角度,全面深入地介绍了数据仓库与数据挖掘的基本理论与应用方法。主要内容包括:数据库与数据仓库应用系统分析,微软 SQL Server 2000 数据仓库工具,数据仓库结构及其创建,数据仓库开发过程,数据转换服务(DTS),联机分析处理(OLAP)的 MDX 表示与实现,数据挖掘基础,常用的知识发现技术,微软 SQL Server 2000 数据挖掘技术。

本书注重工程实践性,实用性强,可以帮助读者全面掌握数据仓库构建与数据挖掘的方法和步骤,开发出具有实用价值的数据仓库系统。

本书适用于高等院校信息管理与信息系统专业、电子商务专业、物流管理专业等相关专业本科生教材,也可作为金融类、管理类有关专业研究生教材,同时,对企事业单位数据仓库与数据挖掘工作人员、研究人员有重要参考价值。

图书在版编目(CIP)数据

数据仓库与数据挖掘技术/刘翔编著. —上海:上海
交通大学出版社, 2005
(企业信息化丛书)
ISBN 7-313-04030-X

I.数... II.刘... III.①数据库系统②数据采集
IV.①TP311.13②TP274

中国版本图书馆CIP数据核字(2005)第080243号

数据仓库与数据挖掘技术

刘 翔 编著

上海交通大学出版社出版发行

(上海市番禺路877号 邮政编码200030)

电话:64071208 出版人:张天蔚

常熟市文化印刷有限公司印刷 全国新华书店经销

开本:787mm×960mm 1/16 印张:17.25 字数:324千字

2005年8月第1版 2005年8月第1次印刷

印数:1—5 050

ISBN7-313-04030-X/TP·619 定价:30.00元

版权所有 侵权必究

前 言

计算机应用于企业管理始于20世纪60年代,经历了60年代的时段式MRP、70年代的闭环式MRP、80年代的MRPII及90年代的ERP,目前正向基于数据仓库技术的智能协同商务系统方向发展。传统模式下的ERP系统运行在UNIX或NT上,由许多企业模块组成,整个操作模式建立在一个封闭的环境中,但是可以提供对外的连接。建立在Intranet环境中的Intranet ERP系统放在大服务器端防火墙的后面,通过Front Office和Self-Service等模块,用户只要登录到Internet,就可以通过浏览器进入到ERP系统。这样,整个ERP系统就可以扩展到世界各地,并且应用起来非常方便;同时,它还有支持移动计算、可以和掌上电脑进行信息的传输和修改等功能。Intranet ERP是在基于Internet信息化的企业管理环境下工作的,企业间外部资源的管理是其主要的管理内容之一。这种以电子方式互联的地理分散的企业管理运作方式产生了许多海量数据处理、分析、智能协同决策的问题,如在非面对面的网络虚拟环境交流的情况下,如何让群体管理人员在讨论问题时方便地交换信息和意见,并对问题的解决提供智能协同决策支持方案,诸如企业预算方案的确定、客户商务谈判方案的确定、供应商协作方案确定。数据仓库就是为了满足用户进行基于海量数据处理、分析的决策支持需要,可以解决上述问题的技术。一般而言,使用数据仓库可以执行下面的工作:

- (1) 把来自异构数据源中的数据整合成单一的数据结构。
- (2) 把数据按照简单的方式进行组织以完成高效率的数据挖掘、查询分析。
- (3) 提供用于智能分析的、有效的、一致的、集成的、格式化的、经过转换的数据。

- (4) 提供业务历史的稳定的数据。

本书以企业信息化为基点,介绍基于ERP数据仓库系统的概念、体系结构、开发方法及步骤。

全书共分9章。第1章主要介绍企业信息化发展历程、数据库与数据仓库概念、特点、区别、联系及数据仓库系统在企业应用概况;第2章介绍数据仓库开发工具——微软SQL Server 2000数据仓库体系结构及应用技术;第3章介绍数据仓库结构及其创建,包括数据仓库数据库、事实表、维度表及多维数据

集创建；第4章介绍数据仓库开发完整过程，包括项目系统规划、用户需求确定、系统分析、系统设计、系统实现、系统试用及扩充；第5章介绍DTS基本概念、DTS包的创建、设置及使用，如何利用DTS包把数据源自动转入数据创库；第6章介绍OLAP的MDX表示与实现，涉及OLAP的基本分析动作的MDX语言表示与实现及OLAP的前端展现方式，提供了丰富的MDX语言程序实例；第7章介绍数据挖掘基础，主要包括数据发掘的概念、数据发掘技术、工具、方法及步骤；第8章介绍常用的知识发现技术及数据挖掘方法，主要包括依赖性分析、聚类分析、基于神经网络的数据挖掘方法、基于遗传算法的数据挖掘方法及基于粗糙集的数据挖掘方法；第9章介绍SQL Server 2000数据挖掘技术。

因作者水平有限，书中难免存在不足与错误，欢迎读者批平指正。如对书中内容有疑问或其他意见，请与作者联系。

联系方法 E-mail: xiangliu@hqu.edu.cn

编 者

目 录

1 数据库与数据仓库应用系统.....	1
1.1 企业信息	1
1.2 决策支持系统	6
1.3 数据库到数据仓库的演变.....	17
1.4 数据仓库在企业领域应用概述.....	21
2 SQL Server 2000 数据仓库工具.....	25
2.1 关系型数据库管理系统.....	25
2.2 Analysis Services	28
3 数据仓库结构及其创建.....	33
3.1 数据仓库数据库.....	33
3.2 事实数据表	44
3.3 维度表	45
3.4 维度	47
3.5 多维数据集	52
4 企业数据仓库开发过程.....	71
4.1 项目系统规划及可行性报告	71
4.2 确定用户需求	82
4.3 系统分析	87
4.4 系统设计	97
4.5 系统实现	105
4.6 试用	106
4.7 扩充	107
5 数据转换服务	109

5.1	DTS 概述.....	109
5.2	DTS 包.....	113
5.3	数据转换.....	117
5.4	使用 DTS 包.....	121
5.5	创建 DTS 包.....	127
5.6	DTS 的高级应用.....	146
6	OLAP 的 MDX 表示与实现.....	155
6.1	OLAP 的基本分析动作.....	155
6.2	OLAP 与 MDX.....	156
6.3	OLAP 的前端展现方式.....	200
6.4	OLAP 的局限性.....	201
7	数据挖掘基础.....	203
7.1	数据发掘的概念.....	203
7.2	数据发掘技术和工具.....	204
7.3	数据挖掘方法及步骤.....	224
8	常用的知识发现技术.....	232
8.1	依赖性分析.....	232
8.2	聚类分析.....	238
8.3	基于神经网络的数据挖掘方法.....	240
8.4	基于遗传算法的数据挖掘方法.....	256
8.5	基于粗糙集的数据挖掘方法.....	261
9	SQL Server 2000 数据挖掘技术.....	265
9.1	数据挖掘数学模型.....	265
9.2	用 DTS 建立数据挖掘模型.....	267
9.3	使用 Pivot Table 服务进行数据挖掘.....	268

1 数据库与数据仓库应用系统

1.1 企业信息

1.1.1 企业信息与管理

从本体论上说,信息是事物存在方式及其运动状态的表现形式,如某时刻企业厂房、设备等的静态客观信息,企业厂房、设备等随时间而变化的动态客观信息。

从认识论上说,信息是主体所感知的或表达的事物存在的方式和运动状态,如企业日、月、年财务报表的静态表现信息及会计科目的借贷实时变化,生产制造过程中物料消耗的实时变化的动态表现信息。

数据是载荷或记录信息的并按一定规则排列组合的物理符号,如图像、文字、数值等。信息是数据载荷的内容,企业管理信息是管理者依据自己的管理知识对管理数据理解的结果,因此,企业数据+管理背景的理解=企业信息。

信息可以被有效地传播、传输。信息可以形成知识,即接收者通过对信息的提炼、推理后又经大脑思维重新组合成系统化的信息,因此,企业信息+管理者应用经验=企业知识。信息社会的三大支柱是物质、能量、信息。信息是重要的核心支柱。

20世纪近百年管理理论的发展异常活跃,出现了许多新的理论与学说,学派众多,难以统计。以西蒙为代表的当代决策理论学派认为管理就是决策,决策的过程包括信息收集、设计方案及选择方案三个阶段,信息是决策的前提,并提出有必要在组织中成立一个特别的信息服务中心。系统管理学派把管理对象视作一个由相互联系的要素组成的、具有共同目标的系统,信息在系统中起着决定性的作用,管理系统中的人、财、物由信息来协调,信息把系统的内外环境联系起来,并实现计划组织控制等基本管理职能。经验主义学派以向大企业的经理提供成功经验和管理工作方法为目标,特别强调信息联系对管理工作的重

要作用。管理科学学派以数学模型为基础处理大量信息,制定管理决策,管理科学的目的是把科学的原理、方法和工具采用各种活动,制定用于管理决策的模型,以减少不确定性,使各种资源发挥更大的效用。经理角色理论认为经理是组织信息的中枢,通过组织内外信息做出正确的决策,经理具有三大类角色,一是人际关系方面的角色,包括挂名首脑角色、领导者角色和联网者角色;二是信息方面的角色,包括信息接收者角色、信息传播者角色和发言人角色;三是决策方面的角色,包括企业家角色、故障排除角色、资源分配者角色及谈判者角色。行为管理理论学派认为管理的效率在很大程度上取决于信息交流的效率,并对信息交流网络进行了系统的研究,为提高信息交流的效率,提出管理领导者应具备六种重要形象:一是信息发送者形象,二是信息发送者对信息接收者形象,三是信息接收者对信息发送者的形象,四是信息接收者所认为的信息发送者的自我形象,五是信息发送者所认为的信息接收者的自我形象,六是信息接收者的自我形象。要提高管理效率就要从改变这些形象入手。社会系统理论学派强调信息联系的重要作用,认为信息联系是一切活动的基础。管理过程理论学派认为管理是在组织中通过别人或同别人一起完成工作的过程,强调吸收其他一切学科的知识构成完整的管理理论,管理职能的实现取决于信息。权变管理理论学派认为管理同环境信息存在一种函数关系,作为因变数的管理思想、管理方法和管理技术,要随环境自变数的变化而变化。

综观这些管理理论的变革主要表现在转向强调信息管理。各个学派正走向一个共同的方向,即企业管理信息化。

1.1.2 企业信息类型

在现代企业管理中,企业生存在信息的海洋中。从不同的角度可以给企业信息作如下分类:

1) 企业信息的内容类型。根据信息的内容,可以将企业信息划分为:

(1) 企业技术信息,是关于企业所需的技术进步或技术开发方面的信息。比如生产技术、产品技术及其标准、设计图纸、实验数据、技术动向、产品开发等。

(2) 企业管理信息,又可分为企业生产管理信息、企业经营管理信息和企业行政管理信息。企业生产管理信息是关于企业生产过程组织、质量管理、人力资源开发与管理、物资及设备管理等方面的信息。企业经营管理信息是关于企业经营思想和战略、市场营销和企业财务等方面的信息,比如原材料的价格、产品销售情况、市场动态以及企业的资产、利税、负债、会计报表等。企业行政管理信息是企业行政管理过程中产生的各种信息,比如上级的指示、政策、

文件,企业向上级的请示、报告,企业制定的管理制度等。

(3) 企业文化信息,是指企业内多数成员在长期过程中形成的、共同拥有的价值观念、行为方式、企业道德、企业精神、企业形象、企业习俗、企业规范等。

2) 企业信息的信源类型。根据企业信息的来源不同,可以将其划分为:

(1) 内源性信息。是企业内部产生的信息,诸如各种生产计划和规划、设计图纸、质量要求指标、企业管理者的指令、下级部门向上级递交的请示、报告等。这类信息是企业管理者管理企业的主要手段,一般是通过正式传播渠道传递。

(2) 外源性信息。是企业从企业经济环境中获得的信息,诸如重大的国家经济活动、政府发布的企业政策法规、市场价格和发展趋势、供应商和销售商的行为、竞争对手的情况等。这类信息通过各种渠道进入企业内部,为企业经营管理决策服务。外源性信息既是企业了解国家经济和社会运行状况的重要依据,也是一种重要的反馈信息,用以检查企业管理的情况和存在的问题,是调整和改进企业管理的依据之一。

3) 企业信息的价值程度类型。根据企业信息的价值程度,可以将其划分为:

(1) 高值信息。指使用价值很高的一类信息,是企业竭力寻求的信息。具体又可以分为:

① 机会信息。指对于企业来说是可能获得大好发展机会的信息。与企业相关的新产品、新技术、新资源、新材料、新市场等。机会信息的使用价值在于它可以导致机会收益。经济过程是不可逆的,机会的错过是企业不可挽回的极大浪费。

② 战略信息。指对于企业来说是至关重要、涉及企业长远利益和生存发展等重大问题的信息。这类信息有时也是机会信息,可能来自企业的高层领导,也可能来自参与决策的智囊人物或局外人。战略信息并不决定于信息本身,而在于信息持有者对信息的认识和理解。

③ 竞争信息。指有关已知的竞争对手、潜在的竞争对手和竞争环境等方面的信息。在今天市场竞争激化的条件下,掌握竞争信息对于提高企业竞争力具有重要意义。

④ 环境变动信息。指企业所处的自然环境和社会经济环境变动的信息,诸如社会、政治、经济、科技、文化和观念等,都是企业生存的重要外部环境。社会经济环境的不断变化给企业的发展带来了许多不确定性,所以及时地掌握环境变动信息,预测环境对企业的影响,及早采取措施,才可以使企业立于不败之地,不断地获得发展。

⑤ 反馈信息。指企业在生产、经营管理过程中实施各种决策和管理措施

之后,获得的关于这些决策和措施实施结果的信息。它是企业管理者发现管理中的经验和存在的问题,及时地坚持好的做法,修正、补充、重新设计新的决策方案的重要依据。反馈信息的获得和利用在企业管理中很重要,它直接反映企业管理者的管理控制水平。

(2) 潜值信息。指具有潜在使用价值的信息。比如,企业平时收集的各种与企业生产、发展有关的文献、资料等,各级管理者在管理工作实践中形成的经验、体会等。潜值信息和高值信息是相对的,是可以相互转化的。当信息采集到手之后,马上就认识到其具有使用价值,自然是高值信息,若没有意识到其具有使用价值,那它就是潜值信息。当认识到该信息的作用,或者经过激活使该信息具有使用价值,则潜值信息就转变为高值信息了。相反,信息在刚采集到时,虽然认识到其具有使用价值,但是如果企业还不具备实施这一信息的条件,该信息仍旧不能实施,高值信息就转变为潜值信息了。潜值信息的存在,要求企业管理者要注意信息的平时积累,只有实现了潜值信息大量的储备,在机遇出现的时候,就可以把握机遇,获得成功。

(3) 常规信息。指维持企业正常运转的那些信息,比如,企业日常活动中的通知、报告、订单、报表、广告、信函等。

(4) 无值信息和负值信息。指没有使用价值、甚至起负面作用的信息。这类信息可能是某些人故意制造的假信息,也可能是信息传播过程中由于各种障碍造成的失真信息,也可能是信息采集者的理解不当造成的信息失真,这些信息对企业管理者的决策是没有帮助的,需要信息管理者能够予以识别,并加以排除。

4) 企业信息的传递范围类型。根据企业信息的传递范围,可将企业信息划分为:

(1) 公开信息。指不限制使用范围的企业信息,可在全社会范围内传播,社会全体成员可以共享,如企业的产品、企业形象、企业总产值、销售额、利润率、固定资产、无形资产、盈利水平、信用与负债等。

(2) 内部信息。指在企业内部传播的企业信息,如企业内部文件、内部会议纪要、生产指挥信息等,不向社会公开。

(3) 保密信息。指严格限制传播范围的企业信息,如企业的专有技术、技术发展动向、战略决策等,不仅不能向社会公开,在企业内也严格控制传播范围。

1.1.3 企业信息化管理历程

1) 时段式物料需求计划(Materiel Request Plan, MRP)系统(20 世纪 60 年

代)。1965年,美国的 Joseph A.Orlicky 博士提出了“物料独立需求和相关需求”的学说,它将物料需求区分为独立需求和非独立需求,同时按具体的时段记录 and 存储物料状态数据,以解决定货时间及订货数量的物料计划编制问题。

2) 闭环式 MRP 系统(20 世纪 70 年代)。这是对时段 MRP 系统的扩展,它依据长期生产计划,经过生产能力负荷平衡分析制定短期主生产计划,然后据此编制并执行物料需求计划、能力需求计划和车间作业计划,以及采购计划的企业内部对生产中的人力、机器和材料资源进行计划与控制的封闭循环物流管理过程。

3) 制造资源计划(Manufacturing Resource Planning, MRP II)系统(20 世纪 80 年代)。这是对闭环式 MRP 系统的进一步扩展,把生产、财务、销售、采购等各个子系统集成为一体化的有机系统,实现资金流与物流的统一管理。MRP II 克服了 MRP 的不足之处,增加了生产能力计划、生产活动控制、采购和物料管理计划三方面内容。生产能力计划功能以物料需求计划的输出作为其输入,根据计划的零部件需求量和生产基本信息中的工序、工作中心等信息计算出设备与人力的需求量、各种设备的负荷量,以便判断是否有足够的生产能力。如果发现能力不足,进行设备负荷的调节和人力的补充;如果能力实在无法平衡,则可调整产品的生产计划。生产活动控制功能以调整好的物料需求计划的输出作为输入,利用计算机的模拟技术,按照作业优先执行的原则,自动地编制各种设备或工作中心的作业顺序和作业完成日期。采购与物料管理计划功能根据物料需求计划和库存管理策略编制物料请购计划,建立采购与进货的管理计划,供应商档案和供应商账务的管理计划,库存会计账务的管理计划,成本、应收账、应付账及总账的管理计划。MRP II 具有模拟功能:模拟将来物料需求而提出任何物料缺料的警告;模拟生产能力需求,发出能力不足的警告。这些警告为管理者提供了必要的信息,使管理者能及时地进行准备与安排。MRP II 包含了分销、制造和财务三大部分,是一个集成度相当高的信息系统。但 MRP II 仅集成企业内部信息,并没有考虑如何充分利用企业外部资源,提高企业综合竞争力的问题。

4) 企业资源规划(Enterprise Resource Planning, ERP)系统(20 世纪 90 年代)。由美国加特纳公司(Gartner Group Inc.)首先提出,当时 ERP 概念只是根据计算机技术的发展和供需链管理,推论各类制造业在信息时代管理信息系统的发展趋势和变革。随着 ERP 实践的发展,ERP 至今已有了更深的内涵,它是 MRP II 面向供应链管理的扩展。ERP 除了传统 MRP II 系统的制造、供销、财务功能外,在功能上增加了支持物料流通体系的运输管理、仓库管理;支持在线分析处理(OLAP)、售后服务及质量反馈,实时准确地掌握市场需求的脉搏;支持生产保障体系的质量管理、实验室管理、设备维修和备品备件管理;支持跨国经营的

多国家地区、多工厂、多语种、多币制需求；支持多种生产类型或混合型制造企业，汇合了离散型生产、流水作业生产和流程型生产的特点；支持远程通信、电子数据交换(EDI)。此外，还支持企业资本运行和投资管理、各种法规及标准管理等。事实上，当前一些 ERP 软件的功能已经远远超出了制造业的应用范围，成为一种适应性强、具有广泛意义的企业管理信息系统。ERP 软件涵盖了商业事务的各个部分。在事务级，应用程序管理着客户和企业间详细的交互过程。通过这种方式进行付费、订货、出货和产品制造。ERP 应用是所有商业运作的中心环节。但对于 ERP 来说，事务处理仅仅是信息处理的开始。数据仓库将是 ERP 进一步发挥作用的关键。

1.2 决策支持系统

决策支持系统是以信息技术为手段，应用决策科学及有关学科的理论和方法，针对某一类型的半结构化和非结构化的决策问题，通过提供背景材料、协助明确问题、修改完善模型、列举可能方案、进行分析比较等方式，为管理者做出正确决策提供帮助的人机交互式的信息系统。

1.2.1 决策问题

决策问题的范围很广，计划、调度命令、政策、法规、发展战略、体制结构、系统目标等都属于决策范畴，但它们的结构化程度不同。决策问题按结构化程度不同可划分为三种类型：

(1) 结构化决策问题。这种类型相对比较简单、直接，其决策过程和决策方法有固定的规律可以遵循，能用明确的语言和模型加以描述，并可依据一定的通用模型和决策规则实现其决策过程的基本自动化。早期的管理信息系统能够求解这类问题，例如，应用解析方法、运筹学方法等求解资源优化问题。

(2) 非结构化决策问题。这种类型决策过程复杂，其决策过程和决策方法没有固定的规律可以遵循，没有固定的决策规则和通用模型可依，决策者的主观行为(学识、经验、直觉、判断力、洞察力、个人偏好和决策风格等)对各阶段的决策效果有相当影响。往往是决策者根据掌握的情况和数据临时做出决定。

(3) 半结构化决策问题。这种类型介于上述两者之间，其决策过程和决策方法有一定规律可以遵循，但又不能完全确定，即有所了解但不全面，有所分析但不确切，有所估计但不确定。

这样的决策问题一般可适当建立模型，但无法确定最优方案。决策问题的

结构化程度并不是一成不变的,当人们掌握了足够的信息和知识时,非结构化决策问题有可能转化为半结构化决策问题,半结构化决策问题也有可能向结构化决策问题方面转化,这是人们对客观事物不断提高认识的过程。

通常认为,管理信息系统(MIS)主要解决结构化的决策问题,而决策支持系统(DSS)则以支持半结构化和非结构化问题为目的。这个论点早期由 Gony 和 Scott Morton 提出,他们把 DSS 定义为“一个在非结构或半结构环境下支持管理决策者的系统”。在此定义中“结构”和“支持”是两个关键概念。“支持”意味着帮助或提高决策者于决策过程之中而非替代决策者。非结构有两个方面的含义:一是问题无结构,这意味着从理论上讲问题本身是不可判定的,但实际上很少遇到此类问题;二是问题在一定意义下有结构而人们至今尚未找到此结构,或未找到恰当的结构。这是我们在实际中经常遇到的问题。半结构问题指问题的局部可以结构化而不能全部结构化。此类问题需介入人的判断来完成。结构问题有两个方面的含义:一是完全结构化,即结构简明,足以完全交付计算机予以自动处理;二是结构复杂,计算机难以处理(如 NP 完全问题)。这种结构复杂的问题在理论上是可以解决的,但考虑实际情况中的时间和空间复杂性就难以实现,因而此类问题必须介入人的判断。在企业组织管理决策中,组织结构中的管理活动可以划分为三个层次:

(1) 战略计划。确定组织的目标、政策和总的发展方向,以组织为整体进行分析。

(2) 管理控制。资源获取(如预算的确定和成本的确定等),其目的是实现战略计划。

(3) 作业控制。在预算内有效地利用现有资源来完成各项活动,具体实施管理控制,间接完成战略计划。

上述三个层次所处理的对象不尽相同:战略计划是面向非结构决策的,管理控制是面向半结构化信息的,而作业控制是面向结构化的。在这里,并不是说 DSS 仅仅是最高层管理惟一需要的,实际上每个层次都有决策问题。

1.2.2 决策支持系统的基本功能

图 1.1 中显示了一个完整的 DSS 系统模式以及它与“真实系统”、人和外部环境的关系。由图可见,管理者处于核心位置,他运用自己的知识,把他和 DSS 的响应输出结合起来,对他所管理的“真实系统”进行决策。管理者往往需要协助人员的帮助。就“真实系统”而言,提出的问题和操作数据是输出信息流,而人们的决策则是输入信息流。图 1.1 的下部表示了与 DSS 有关的基础数据,它包括来自“真实系统”并经过处理的信息(如 MIS 信息、统计信息等)、

环境信息、与人的行为有关信息等。图 1.1 的右边是 DSS，由模型库系统、数据库系统和人机对话系统等组成。决策者运用自己的知识和经验，结合 DSS 响应输出，对他所管理的“真实系统”进行决策。

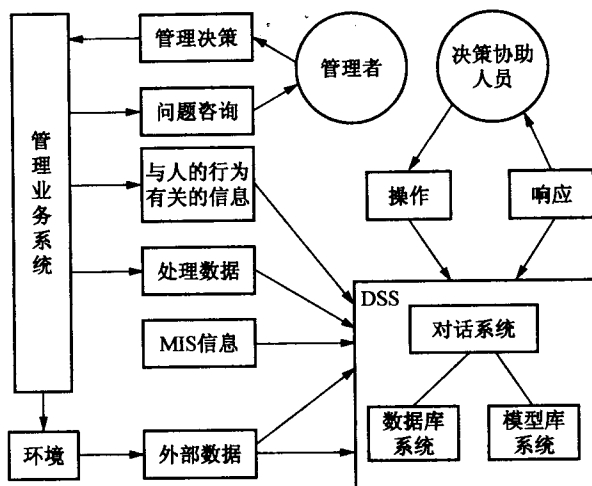


图 1.1 传统 DSS 基本模式

人作为 DSS 的用户、开发者和评价者在 DSS 的建立和运行中起着关键作用。

人的因素(如个人判断和偏好)是一个重要而实际的问题。Elam 等对人的创造性、判断及学习能力在 DSS 中的体现开始有所研究。Shaw 等人则将人工智能中机器学习的方法应用于模型管理之中。

人的因素是一个复杂的因素，它涉及人的心理与认识过程。这是将人工智能(特别是专家系统)与传统 DSS 相结合的一个重要原因。

依据以上 DSS 模式，DSS 基本功能可描述如下：

- (1) 整理并及时提供本系统与本决策问题有关的各种数据，如工厂的生产能力、库存和财务状况等。
- (2) 收集、存储并及时提供系统之外的与本决策问题有关的各种数据，如市场需求、原材料价格、新技术动态等。
- (3) 及时收集和提供有关各项行动的反馈信息，如生产计划完成情况、产品销售情况和用户反映等。
- (4) 能够用确定的方式存储与决策有关的各种模型，如库存控制模型与生产调度模型等。
- (5) 能够存储及提供常用的数学(特别是数理统计)与运筹学的方法，如统计检验方法、回归分析方法、线性规划方法等。

(6) 上述数据、模型、方法的管理都应该是容易改变、容易增添的,如数据模式的变更、模型的连接或修改、各种方法的修改等等,都可以由用户修改变更。

(7) 能够灵活地运用模型与方法对数据进行加工、汇总、分析、预测,以便得到所需要的综合信息与预测信息。

(8) 提供方便的人机对话接口或图形输出功能,不仅能够随机查询所需的数据,而且能够回答“如果……则……”(what... if...)之类的问题。

(9) 具有使用者能够忍受的加工速度与响应时间。

1.2.3 以数据仓库为基础的决策支持系统

DSS 和数据仓库的目标客户群相同,都是针对企业的中高层主管,它们执行的都是决策和趋势分析之类的应用,DSS 中一些传统的统计分析方法及一些智能型决策技术可以很好地综合到数据仓库中,使数据仓库的分析能力更为强大,并大大提高数据仓库决策支持能力。同时,数据仓库解决了以后 DSS 中数据不一致的问题,为决策支持系统开辟了一条新途径,也使 DSS 的应用范围更加广泛。随着数据仓库的广泛应用,以数据仓库为基础的企业决策支持系统应运而生。以数据仓库为基础的整体解决方案如图 1.2 所示。

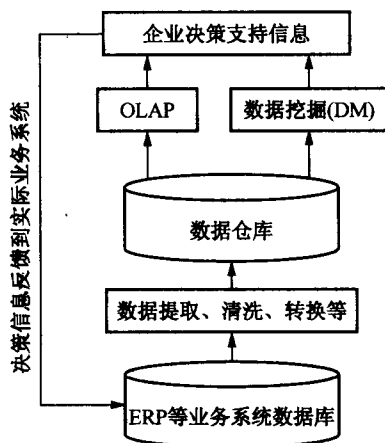


图 1.2 以数据仓库为基础的决策支持系统解决方案

图中设计的 DSS 将数据库、模型库及方法有机地结合在一个多维数据库中,并以面向对象的方式进行存储,较好地协调了数据、模型和方法三者之间的关系,使整个系统形成一个有机的整体,从而提高了系统的综合性。联机分析处理工具(OLAP)和数据挖掘(DM)在整个 DSS 整体解决方案中发挥了关键性

的作用。OLAP 从数据仓库中的综合式数据出发, 提供面对分析的多维模型, 并使用多维分析方法, 从多个角度、多个侧面及多个层次对多维数据进行分析与比较, 使客户能以更自然的方式来分析数据。而 DM 以数据仓库和多维数据库中的大量数据为基础, 自动地发现数据中的潜在模式, 并以此种模式为基础, 自动做出预测。

决策支持工具是数据仓库应用程序及工具的统称, 这些工具可以按验证和发现两种模式使用。在验证模式中, 用户做出假设, 然后试图通过存取数据仓库中的数据来证明该假设。它是面向用户的, 用于执行验证模式的工具包括查询、多维分析工具。在发现模式中, 用户事先并不知道发现的方式与关系, 工具尽力帮助用户处理大量存在的数据, 以期发现一些意外的关系。它明显是面向数据的, DM 工具是发现模式的一个范例。可以说, 此两种模式相辅相成, 共同支持决策支持工作平台。

OLAP 的功能结构是 3 层客户服务器结构, 此结构由 3 个服务器组件组成: 数据仓库服务器、OLAP 服务器及用户描述服务器。其物理结构有多维数据仓库器及关系型数据仓库器。因而, OLAP 也就形成了面向多维数据库(MDDB)的 MOLAP 和面向关系型数据库的 ROLAP。

面向多维数据库的 MOLAP 以多维数据库为重点, 它使用多维数据库管理系统来管理所需的数据或数据仓库。各个联机交易处理系统(OLTP)数据库中的数据, 通过筛选、过滤、转换与综合等步骤后, 提交给多维数据库。多维数据库依靠“维”来形成超立方体结构, 并产生旋转、切片、成块、上翻、下查等操作。用户通过客户端的应用软件接口向 OLAP 服务器提交分析要求, 由 MOLAP 服务器检索 MDDB, 然后将得到的结果显示给用户。

面向关系型数据库的 ROLAP 以关系型数据库为重心, 而不是多维数据库。在 ROLAP 结构中, 用 RDBMS 来代替 MDBMS。ROLAP 将多维数据库的多维结构划分为两种列表: 一类为事实数据表, 用来存储数据和维关键词; 另一种是维表, 即对每个维至少使用一个表来存放维的阶层、成员类别等维的描述信息。它是一种有组织的分类(或层级)阶层结构, 例如地理维可能包含国家、区域、州或省以及市层级。两者通过主键和外键结合起来。用户可通过客户端的工具向 OLAP 服务器提交分析要求, OLAP 服务器动态地将这些要求转换成“结构化查询语言”来执行, 结果以多维窗口方式显示给客户。

DM 是决策支持工具中的重要组成部分。它是数据分析的发现模式, 通过对详细数据的分析, 发现出意料之外或未知但潜在的关系和模式。例如, 产品和客户的关系、产品和客户购买模式的关系等, 其目的是要发现“富有战略性和竞争性的问题”。要发挥 DM 的功能, 首先必须在企业中建有数据仓库, 同时还必须具备执行 DM 的技术和工具。DM 的技术和工具可分为以下三大类: