数据仓库与数据挖掘

SHUJU CANGKU YU SHUJU WAJUE

陈燕 著

大连海事大学出版社

数据仓库与数据挖掘

SHUJU CANGKU YU SHUJU WAJUE

数据仓库与数据挖掘

陈　燕　著

大连海事大学出版社

图书在版编目(CIP)数据

数据仓库与数据挖掘 / 陈燕著 . —大连 : 大连海事大学出版社, 2006.12
ISBN 7-5632-2021-6

Ⅰ.数… Ⅱ.陈… Ⅲ.①数据库系统—研究生—教学参考资料 ②数据采集—研究生—教学参考资料 Ⅳ.①TP311.13 ②TP274

中国版本图书馆 CIP 数据核字(2006)第 160250 号

大连海事大学出版社出版

地址:大连市凌海路 1 号 邮编:116026 电话:0411-84728394 传真:0411-84727996
http://www.dmupress.com E-mail:cbs@dmupress.com
大连金华光彩色印刷有限公司印装 大连海事大学出版社发行
2006 年 12 月第 1 版 2006 年 12 月第 1 次印刷
幅面尺寸:185 mm × 260 mm 印张:17.5
字数:399 千 印数:1 ~ 1000 册
责任编辑:董玉洁 版式设计:冰 清
封面设计:王 艳 责任校对:李雪芳
定价:29.00 元

内 容 提 要

本书较系统地介绍了数据仓库产生的背景及其技术、方法的理论和应用。主要内容包括:数据仓库的相关术语及框架体系结构;数据仓库设计与元数据研究;异构数据智能整合与建模;联机分析处理与联机分析挖掘;数据挖掘与数据库中知识发现;基于神经网络的数据挖掘模型的应用研究;预测模型与应用;基于 GMDH 原理的自组织数据挖掘模型研究;快速发现关联规则的模型及应用;粗熵的关联规则挖掘方法及其在肇事逃逸侦破中的应用;模糊层次分析法等。

前　言

随着计算机的普及与发展，计算机信息化管理模式的不断出现，使更多的用户和计算机使用者越来越依赖于 MIS 技术，尤其是在网络技术、企业组织发展朝着面向全球化、国际化的目标发展，众多的信息管理工程研究者和管理者面临一个最严峻的瓶颈问题：信息化企业陷入了海量的、分散的、复杂类型的数据海洋中，或者说是陷入了多维的、面向空间的复杂类型数据海洋里。那么，如何在这些复杂类型数据中集中组织、建模与管理、分析这些数据，如何从数据的海洋中提取企业经营中所需要的有价值的信息，并且及时从中挖掘有价值的信息，是多年来信息管理工程领域与管理工程领域所研究的重要课题。

关于上述问题的解决方案，从 20 世纪 90 年代起，国际上著名的信息工程专家就提出了利用建立数据仓库的方案解决此问题。多年来，国内外已经有许多面向各领域的数据仓库的解决方案，并且还相继出现了许多重要的研究成果。笔者在此方面的研究中，也做了 10 年的潜心研究，尤其是面向大连海事大学交通运输、物流管理专业特色领域，做了大量的有关数据仓库技术与数据挖掘模型的研究，尤其是近几年来，所研究的成果获得国家自然科学基金的资助，所从事的相关研究成果还获得省、市多项科技进步奖；还将所研究的成果用于交通、物流等领域，取得了良好的经济效益。

为了使信息管理工程理论、技术和方法向纵深发展，适合管理科学与工程学科研究生及计算机技术、信息管理、物流系统工程等领域学生学习和研究的需要，笔者撰写了《数据仓库与数据挖掘》一书。主要内容包括：各种管理方法、数据仓库产生的背景；数据仓库的相关术语及框架体系结构；数据仓库设计与元数据研究；异构数据智能整合与建模；联机分析处理与联机分析挖掘；数据挖掘与数据库中知识发现；基于神经网络的数据挖掘模型的应用研究；预测模型与应用；基于 GMDH 原理的自组织数据挖掘模型研究；快速发现关联规则的模型及应用；粗熵的关联规则挖掘方法及其在肇事逃逸侦破中的应用；模糊层次分析法等内容。在撰写本书过程中，笔者还参考了国内外近百种参考文献和相关资料，为本书的理论、技术、方法和实际应用模型的研究打下了良好的基础。

本书的许多模型在实际应用和研究中，得到了使用单位和研究者的指导和帮助，在此深表感谢。

作为本书的作者，从研究到撰写，笔者一直感受到了撰写本书的重要意义和用途。撰写本书的目的在于：利用数据仓库技术与方法，将异构的、分散的、空间的、多维的、复杂类型的数据整合在一个公共平台上来统一组织与管理，并进行预测、决策分析，也就是说，将一般的信息管理方法有效地变成从信息管理到知识管理的全过程，从而达到信息的高层管理、深加工及信息增值的目的。

由于信息技术的管理方法发展多样化，再加上本人的能力和水平有限，本书难免存在不足之处，恳请读者提出宝贵意见。

作　者

2006 年 9 月

目 录

第 1 章　绪论

1.1　MIS 的开发方法

1.1.1　MIS 的发展

MIS 是 Management Information System 的缩写。MIS 的定义是：由计算机和通信设备等组成，进行管理信息的收集、传递、储存、加工、维护和使用，用于辅助一个部门（企业）的事务处理和管理职能部门的系统。

虽然这里讲的管理信息系统是以计算机为核心的计算机信息系统，但 MIS 首先强调的是“管理”（Management），建立的信息系统是为管理服务的；其次强调的是“信息”（Information），也就是说对所处理功能的系统，信息是这个系统的“血液”、动力和归宿，没有信息的流动，系统将不存在；第三强调的是“系统”（System），也就是说所处理的对象是具有管理功能的系统。

而计算机应用（Computer Application），第一个强调的对象是计算机，第二个强调的才是应用，即如何使用（应用）计算机。在此，并没有出现 MIS 的字样。我们将一个科学题目、项目、报表、算法等都认为是计算机的应用。

MIS 首先重视的是管理，然后才是信息技术（Information Technique，简称 IT），而不是计算机技术（Computer Technique，简称 CT）。从深层次讨论的实质来看，虽然 IT≠CT，但是从实际问题的实现与所产生的效应来说，在一定效应上，IT 比 CT 要复杂得多、灵活得多。再次，因为突出了系统的概念，所以在解决这一类问题上，必须引用一系列工程的理论方法，评价方案，方案的选定以及方案的规划实施等环节来系统化、完善化。

1.1.2　MIS 开发方法的步骤

（1）MIS 开发的总体规划

①总体规划的内容和步骤

现行规划的初步调整；系统开发策略的确定（比如是采用自上而下，还是自下而上的方法）；可行性研究（应该从经济、技术、管理诸方面的基础来分析可行性）；可行性报表的编写（必须对现行系统的优、缺点进行分析，对系统的组织、新系统方案的建立等进行研究）。

②总体规划的其他方法

关键成功因素法（Critical Successful Factor，简称 CSF 法）和某部门系统规划法（BSP 法）。在开发策略的确定过程中，应用综合开发方法，开发一个信息系统的规划和实现过程，如图 1-1所示。

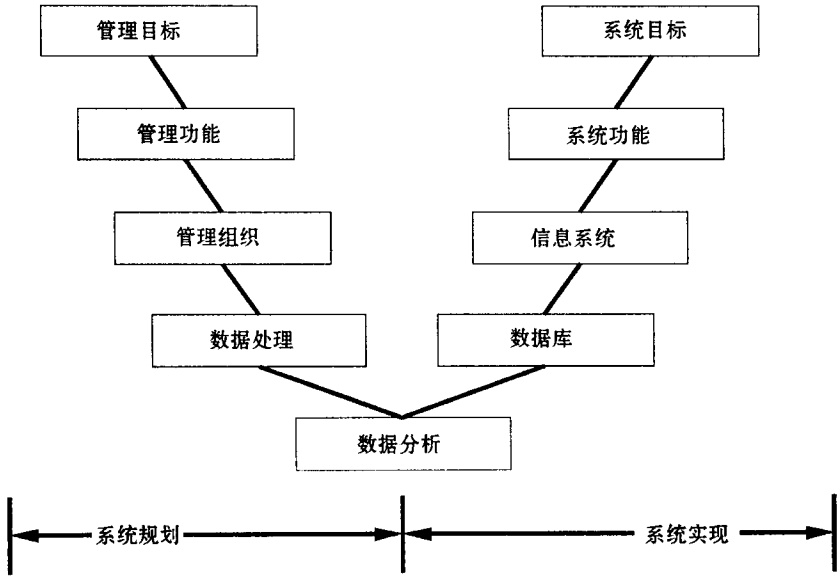

图 1-1 信息系统的规划和实现过程

(2) 系统分析

系统分析包括系统分析的任务和步骤、对现行系统的调查、数据流图、数据字典。

以 CSF 法(关键成功因素法)为例,说明系统分析。该方法是由哈佛大学 Williamzani 教授和 MIT 大学 John Bockart 教授提出的。CSF 法的目标是开发数据库,因此输出的是一个数据字典。其步骤是:

①了解企业目标;

②识别关键成功因素;

③识别性能的评价指标和标准;

④识别度量性能的数据。

CSF 模型如图 1-2 所示。

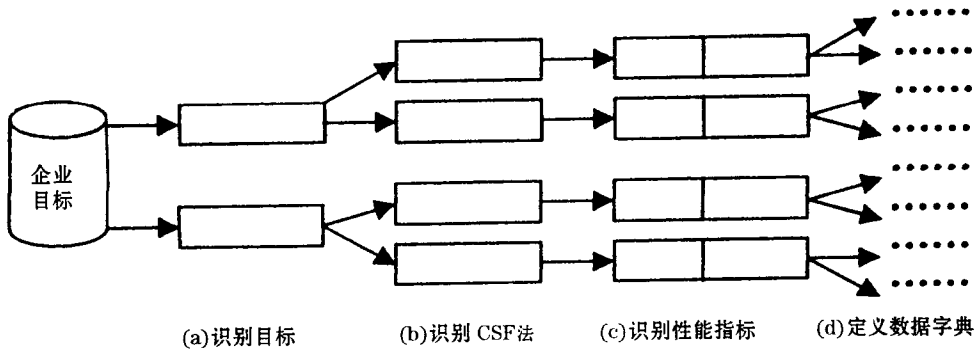

图 1-2 CSF 模型

(3) 系统设计

系统设计分如下几个步骤:

①概要。

主要从两个方面即目标和内容来考虑,其中"目标"包括该系统的可靠性、可维护性、用户界面友好、系统工作效率以及合法性等;"内容"指的是该系统的硬件配置、功能模块设计、DBS 设计、代码设计等。

②结构化系统设计方法(Structure Design,简称 SD 方法)。

由系统工程设计人员根据功能说明书进行设计。其结果是向程序员交付规格说明书,涉及软件的设计技术,如结构设计规约和结构设计图等内容。结构设计规约(Structure Design Specification)的主要目的在于:软件设计工程开始后,根据功能规约或用户提出的规约经过一系列设计过程所做的过程说明书。通常运用统一的方式描述程序的规格和构造。而在结构设计图中主要是指系统结果化设计所使用的描述方式,其描述了程序的模块结构和层次特征,反映了块间联系和块内联系。结构图中的主要成分有:模块、调用、数据。结构图方法设计总则是使每个模块执行一个功能,模块间传送数据型参数,且

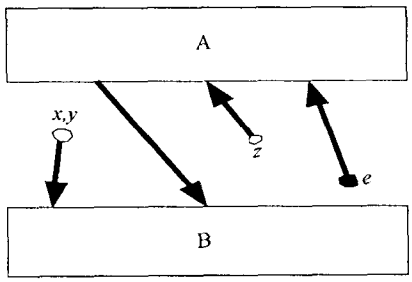

图 1-3　模块 A 与 B 之间的数据调用、返回(传递)的形式

参数尽量少。图 1-3 给出了模块 A 调用模块 B 与从模块 B 返回至模块 A 中的数据 x,y,z,e 之间的关系。其中:如模块 A 调用模块 B,将数据 x,y 传递给模块 B;而从模块 B 返回模块 A 时,将数据 z 即控制信号 e 传递给模块 A。

③功能模块设计。

④系统平台设计:分单业务 MS,综合业务 MS,集成管理系统,客户机/服务结构。

⑤代码设计。

⑥数据库设计。

⑦输入设计。

⑧输出设计。

(4) 系统实现

这一阶段的工作包括编制程序、人员培训、数据准备、试运行、对系统的评价与局部调整的 MIS 的试运行情况。

1.1.3　MIS 与 CA 的重要性

综合上述来看,如果只强调计算机应用,那么所编的软件满足不了用户的需求,这也

说明开发信息系统一开始时,就必须强调调整管理的前期工作即系统分析与设计的重要性。因为只有这样,才能够体现出“创造性劳动”这个含义。

如果从 MIS 的组合“M”+“I”+“S”来看,没有直接出现“C”,但 MIS 却包括如下方面的内容:

①MIS 是一个系统工程。所谓系统工程指的是:随着计算机的出现和发展,形成了一些既有控制功能又有管理功能的系统,于是发展起来一门新的工程技术,称为系统工程。系统工程的工作对象是能量、材料和信息,主要是信息。系统工程的主要工作是获得各项信息,进行判断、加工,然后输出必要的信息,显示给操纵管理人员作为判断的依据或直接完成对有关设备的控制。

对系统的分析、综合、模拟、优化等比较理论性的技术是狭义的系统工程理论;为了合理地进行系统的研制设计、运用等所采用的思想、程序、组织、方法等内容是广义的系统工程理论。系统工程不仅是技术问题,而且还涉及经济、经营、社会、心理等因素。

其类型有:只对数据完成收集管理以便检索的数据收集管理系统;对各项情报进行计算、控制传递的情报系统;用于计划和调度的经营管理系统;用于大型工厂生产过程的控制管理系统;用于军事方面的指挥控制系统等等。

②MIS 是以计算机和通信网络为核心的人/机系统。

③MIS 包含管理体系和组织结构。

④MIS 是覆盖整个组织,支持各级管理和决策的计算机信息系统。

⑤人是 MIS 不可分割的一部分。

⑥MIS 应该能实现系统现状,控制系统行为,预测系统未来及指出系统发展战略。

1.2 MIS 的发展过程

随着 20 世纪 80 年代初 PC 机的大量涌入市场,PC 机的价格一再下降,信息技术也发展很快,MIS 得到了迅速的发展,受控制论、信息论、系统工程和信息经济学的影响,形成了 MIS 的层次概念和系统设计方法。当时 MIS 的核心是以信息作为“血液”,在系统这个大动脉中流动,因此,结合市场需求,为提高生产效率,信息逐渐被企业看作盈利和竞争的资源,在市场中形成了因人、因地、因工程种类而异的先进管理方式。这些方式有:JIT、5S 管理、MRP、MRP II、综合管理等。

1.2.1 JIT(Just In Time)有效的物流管理系统

有效的物流管理系统在日本丰田公司最早被使用。这种即时工作法,即定时、定点、定量运送物料。“三定”送料的目的是将材料的利用率达到最大,并减少现场管理时的意外损失。JIT 在日本的管理方面(尤其是零库存)体现了最为先进的技术。现在,有许多企业的库存都是在运输的路上。

1988 年,巴克斯特公司利用原美国医院供应公司(American Hospital Supply Corp,1985 年被巴克斯特公司兼并)开发的 MIS 成为有关医院全线产品的供应商,也成为满足全美医院需求的商品总汇。做到这一点需有 120 000 种以上的库存量,而要维持海量的库存是十分昂贵的。有时对于紧缺产品、稀有产品的库存也十分昂贵。为使医院转向竞

争者的供货，而将连接到巴克斯特公司的计算机终端安装在医院里。在医院下达订单时，不需要给销售人员打电话或送订货单，只需利用医院里的巴克斯特公司的终端，就可以从巴克斯特公司的全线供货目录中订货。该订货系统生成订单、收款单、发票和库存信息，还为客户提供到货日期，在全美有 80 多个分销中心的巴克斯特公司常常在收到订单的几个小时内，把货物当天送到客户手中。这个供货系统的模式类似于日本发明的仍被美国汽车工业采用的即时送货（JIT—Just In Time，一种强调以现场控制为主导的拉式生产管理）系统。在即时供货系统的使用中，把具体的汽车部件数量和部件发货计划输入到该公司的信息处理系统中，然后这些要求被自动地输入到某个部件供应商的订单输入系统中，该供应商必须同意按规定的时间供货。因此，该汽车公司利用此管理系统能降低库存成本，减少存放部件或原料的场所与项目建设时间等。

巴克斯特公司利用此管理系统不仅将货物送到医院的库房里，而且还进一步将医院订的物品直接送到该院的手术室、护理室等直接应用的地方，这种方法真正创造了零库存的管理方式，把一切库存的责任转给经销商，这比传统库存的管理方法要省去许多复杂的环节，同时也大大减少了库存的损耗、库存的管理人员等。

1.2.2　MRP 与 MRP Ⅱ的兴起

早在 20 世纪 60 年代初，美国就出现了在库存订货点法的基础上，经过生产实践的不断积累、完善而形成的一种以计划与控制为主导，使用 CAD/CAM、GDI 和 EDI 等接口的企业思想和方法编制出物料需求计划（Material Requirements Planning，简称 MRP），MRP 被广泛应用于离散、重复和连续生产的工业企业，取得了显著的经济效益和社会效益。

（1）MRP Ⅱ的出现

传统的库存理论认为：只有降低供货率，才能减少库存费用。成功的 MRP 系统已经证明：可以在降低费用的同时，提高供货率。结合这种思想，把企业作为一个有机的整体，从整体优化的角度出发，应用科学的方法把企业各种制造资源的产、供、销、财、工程技术等各个环节合理有效地组织、控制和调整，使得它们在生产经营过程中得以协调有序，并充分发挥作用，形成一个一体化的系统，称为制造资源计划（Manufacturing Resource Planning，其缩写也为 MRP，为区别物料需求计划的 MRP，而记为 MRP Ⅱ）。MRP Ⅱ展示了制造业新的管理方法，全面覆盖了市场预测、生产计划、物料需求、能力需求、库存控制、现场管理等，直至产品销售的各个生产过程及与之相关的所有财务活动。

（2）MRP Ⅱ的特点

①计划性：分层计划，确保供需平衡，主要根据经济学理论，在市场上产品应达到供需相等的需求—供给曲线，如图 1-4 所示。

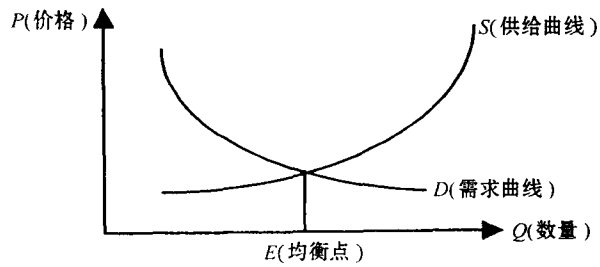

图 1-4　需求—供给曲线

②管理的系统性：将企业中所有与生产、库存有关的各子系统有机地结合起来，形成一个面向整个企业的一体化系统，其中生产和财务子系统

关系尤为密切，每个部门都从整体考虑从事本岗位工作。

③数据的共享性：MRP Ⅱ所有数据均来源于企业的中央数据库，并为下一步建立数据仓库，以公共平台为基础分析数据库、管理数据库打下了良好的基础。

④动态的反馈性：MRP Ⅱ是一个闭环系统，其工作过程是“计划→实施→评价→反馈→计划”的闭环系统。该闭环系统的过程如图 1-5 所示，能跟踪、反映随机多变的实际情况，使之协调、平衡，并根据反馈信息来判断、分析、调查、组织和生产。

⑤系统模拟的预测性：MRP Ⅱ能够反映生产管理的客观规律，由此建立的信息逻辑具有模拟功能，可以对各种情况（条件）下的数据进行模拟，可以采用不同的决策方法来模拟未来发生的结果。这些方法可按单目标、多目标的决策方法分类。一般单目标决策方法包括表1-1的内容。

表 1-1 目标决策方法

名称	细分	具体说明
盈亏决策分析		
线性规划决策分析		
确定条件下的其他决策方法	差量分析法	
	经济批量法	
	临界成本法	
风险决策分析	决策分析法	
	决策树分析法	
	矩阵分析法	
未确定型决策分析	小中大决策方法	悲观决策：运行保守，从最坏的情况来看
	大中小决策方法	又称最小最大后悔值法，亦称遗憾值法：由于未采取最佳方案而造成后悔，为尽量减少后悔而采取最小和最大后悔值
	大中取大决策方法	乐观决策方法
	折中分析法	又称赫威斯法，介于乐观与悲观之间

折中分析法设置一个处于 0～1 之间的折中系数 α，利用折中系数计算出各行动方案最小收益值和最大收益值的折中收益后，选择最大的折中收益值所属的方案为最优方案。当折中系数趋于零时，近似于悲观决策法；当折中系数趋于 1 时，近似于乐观决策法。

例：设 A_i 为第 i 种方案的折中收益值

α 为折中系数（$0 \leqslant \alpha \leqslant 1$），$a_{ij}$ 为第 i 种方案在第 j 种自然状态下的收益值，有：

$$A_i = \alpha \max(a_{i1}, a_{i2}, \cdots, a_{in}) + (1-\alpha)\min(a_{i1}, a_{i2}, \cdots, a_{in})$$

再设 $A^* = \max(A_i)$，A^* 所属方案为最优方案（$i = 1, 2, \cdots, m; j = 1, 2, \cdots, n$）

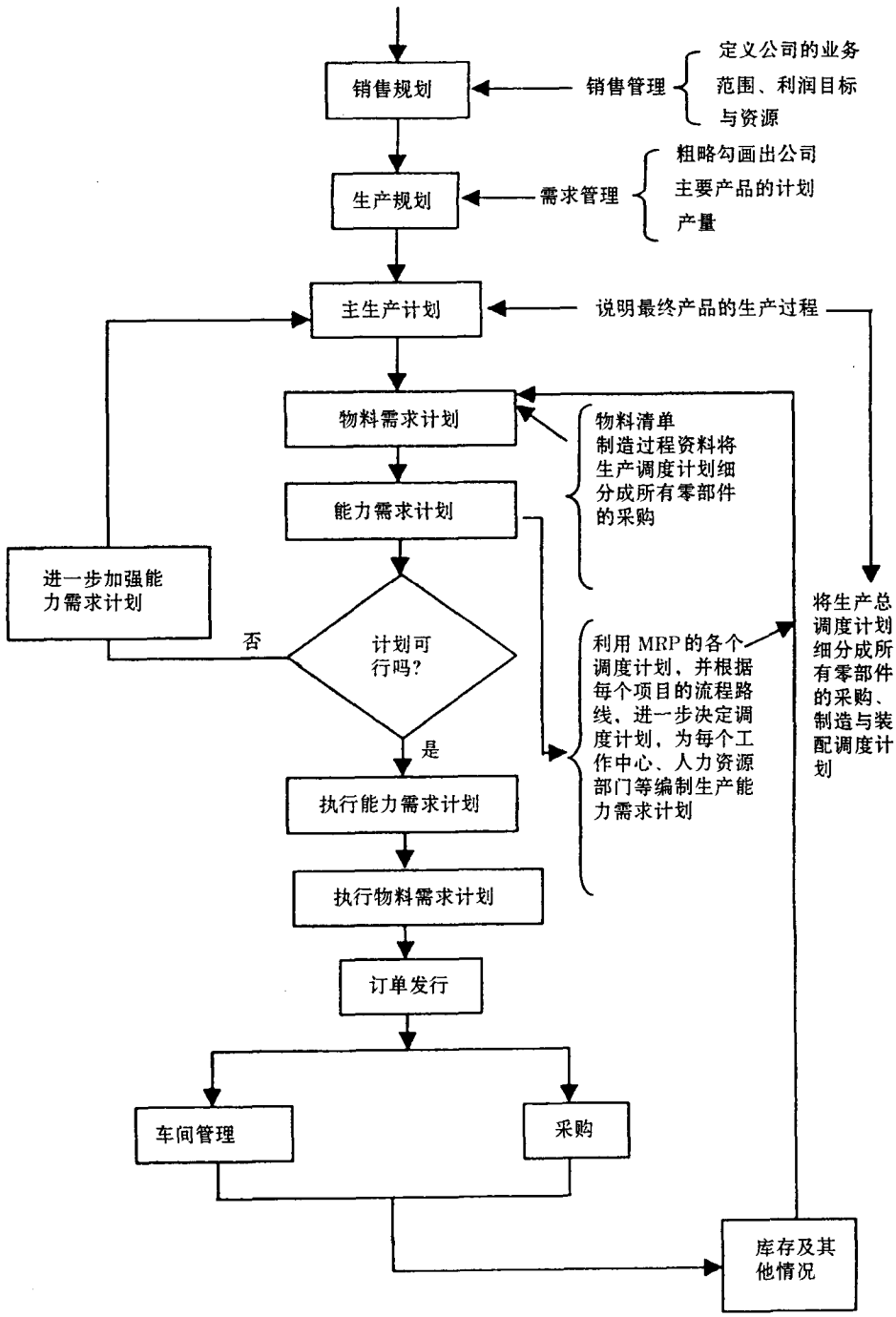

图 1-5　MRP Ⅱ闭环系统模式

根据资料(见表1-2):

表1-2 折中方法选择最优方案的数据 单位:万元

方案 / 预期收益 / 市场情况	方案I	方案II	方案III
旺盛	550	810	450
一般	400	370	250
疲软	-190	-250	-100

运用折中方法选择最优方案,设 $\alpha=0.5$ 则有:

$A_1=550\times0.5+(1-0.5)\times(-190)=180$(万元)

$A_2=810\times0.5+(1-0.5)\times(-250)=280$(万元)

$A_3=450\times0.5+(1-0.5)\times100=225$(万元)

$A^*=\max(180,280,225)=280$(万元)

一般来说,根据以各个决策方法对于不同的事物及其发展趋势而采取相应的决策方法。具体的选法可参考相应的资料。

⑥"三流"的统一。

所谓"三流"是指物流、信息流和资金流。MRP Ⅱ以物流、信息流、资金流的统一为目的,基于现有的生产能力与设备条件,有计划地、合理地使用资源,达到追求更大的生产效益的目的。

一般在这"三流"中,以信息流作为最基本、最基础的信息源,如果没有信息流,物流管理与资金流也就无从谈起。我们可以将这"三流"用一个简单的图示法给出,如图1-6所示。

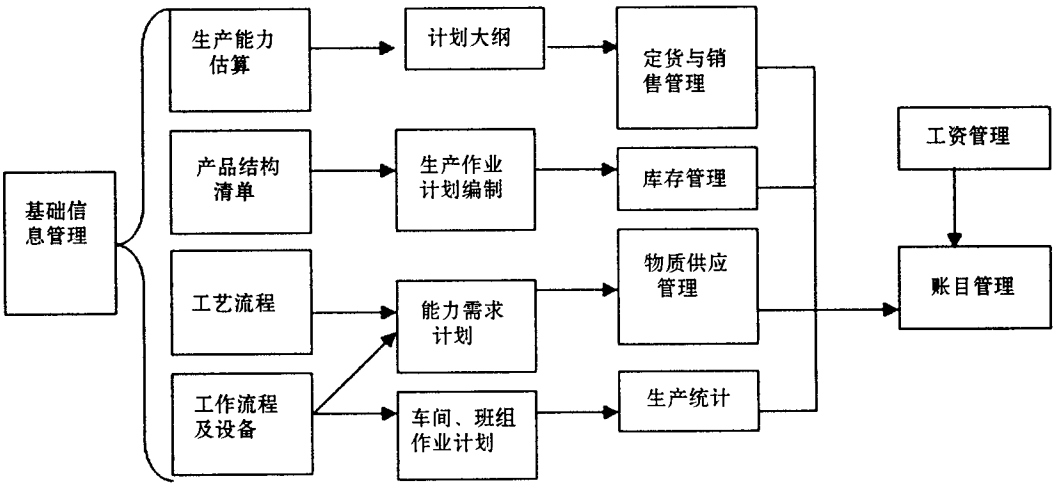

图1-6 MRP Ⅱ的"三流"统一模式

综合起来看,MRP Ⅱ是对基本生产均衡的模拟,为管理人员提供了一套强有力的计