第3章　多维数据分析与组织

本章介绍了联机分析处理的定义、特点和一般的评价准则，从概念模型、逻辑模型、物理模型三个层面阐述了多维数据模型与结构；介绍了多维数据分析的基本操作和相关工具，以及不同的多维分析工具的特点；结合联机分析处理和数据挖掘的优势，提出了联机分析挖掘的概念及特征。

3.1　多维数据分析概述


视频 3-1

3.1.1　联机分析处理的定义和特点

1. 联机分析处理的定义

联机分析处理（OLAP）的概念最早是由关系数据库之父 E.F.Codd 于 1993 年提出的。Codd 认为联机事务处理（OLTP）已不能满足终端用户对数据库查询分析的要求，SQL 对大型数据库的简单查询也不能满足用户分析的需求。用户的决策分析需要对关系数据库进行大量计算才能得到结果，而查询的结果并不能满足决策者提出的需求。因此，Codd 提出了多维数据库和多维分析的概念，即 OLAP。

OLAP 是针对特定问题的联机数据访问和分析。通过对信息（维数据）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。OLAP 委员会对联机分析处理的定义为：使分析人员、管理人员或执行人员能够从多种角度对从原始数据中转化出来的、能够真正为用户所理解的，并真实反映企业特性的信息进行快速、一致、交互的存取，从而获得对数据的更深入了解的一类软件技术。

2. 联机分析处理技术的特点

OLAP 技术的主要特点有两个：一是在线性（On-line），表现为对用户请求的快速响应和交互操作；二是多维分析（Multi-dimension Analysis），也是 OLAP 技术的核心所在。具体特征可分为以下四点：

（1）多维性（Multi-dimensional）：多维性是 OLAP 的关键属性。系统必须提供对数据的多维视图分析，包括对层次维和多重层次维的完全支持。OLAP 最显著的特征是它能提供数据的多维概念视图。在 OLAP 数据模型中，多维信息被抽象为一个立方体，它包括维和度量。维就是观察角度，而度量则是指标值。多维结构是 OLAP 的核心，OLAP 展现在用户面前的就是一幅幅多维视图。这些多维视图能使最终用户从多角度、多侧面、多层次直观地考察数据仓库中的数据，从而深入地理解包含在数据中的信息及其内涵。以多维视图的形式把数据提供给用户，既迎合了用户的思维模式又减少了概念上的混淆，同时也降低了出现错误解释的可能性。

（2）快速性（Fast）：用户对 OLAP 的快速反应能力有很高的要求。一般认为 OLAP 系

统应在几秒内对用户的分析请求做出响应。如果终端用户在 30 秒内没有得到系统响应就会变得不耐烦，因而可能失去分析主线索，影响分析质量。对于大量的数据分析要达到这个速度并不容易，因此就更需要一些技术上的支持，如专门的数据存储格式、大量的事先运算、特别的硬件设计等。

(3) 可分析性(Analyzability)：OLAP 系统应能处理与应用有关的任何逻辑分析和统计分析。尽管系统可以事先编程，但并不意味着系统定义了所有的应用。在应用 OLAP 的过程中，用户无须编程就可以定义专门计算，并将其作为分析的一部分，以用户所希望的方式给出报告。用户可在 OLAP 平台上进行数据分析，也可连接到其他外部分析工具上。

(4) 信息性(Information)：不论数据量有多大，也不管数据存储在何处，OLAP 系统应能及时获得信息，并且管理大容量信息。这里有许多因素需要考虑，如数据的可复制性、可利用的磁盘空间、OLAP 产品的性能以及数据仓库的结合度等。

随着 OLAP 技术的应用范围日渐广泛，出现了一些新的技术，如面向对象的联机分析处理(Object-oriented OLAP，OOLAP)、对象关系的联机分析处理(Object Relational OLAP，OROLAP)、分布式联机分析处理(Distributed OLAP，DOLAP)、时态联机分析处理(Temporal OLAP，TOLAP)。

3.1.2 联机分析处理的评价准则

E.F.Codd 同时提出了关于 OLAP 的 12 条准则来描述 OLAP 系统。

准则 1：OLAP 模型必须提供多维概念模型。从用户分析员的角度来看，整个企业的视图本质上是多维的，OLAP 模型必须提供多维概念视图，因此 OLAP 的概念模型也应是多维的。

准则 2：透明性准则。无论 OLAP 是否是前端产品的一部分，对用户来说它都是透明的，如果在客户/服务器结构中提供 OLAP 产品，那么对最终分析员来说，它同样也应透明。

准则 3：存取能力准则。OLAP 系统不仅能进行开放的存取，而且还提供高效的存取策略。OLAP 用户分析员不仅能在公共概念视图的基础上对关系数据库中的企业数据进行分析，而且在公共分析模型的基础上还可以对关系数据库、非关系数据库和外部存储的数据进行分析。

准则 4：稳定的报表性能。当数据维数和数据的综合层次增加时，提供给最终分析员的报表能力和响应速度不应该有明显的降低和减慢，这时维护 OLAP 产品的易用性和低复杂性至关重要。

准则 5：客户/服务器体系结构。OLAP 是建立在客户/服务器体系结构上的，它要求多维数据库能够被不同的应用和工具访问到，服务器智能地以最小的代价完成多种服务器之间的映射，并确定它们的一致性，从而保证透明性和建立统一的公共概念模式、逻辑模式和物理模式。

准则 6：维的等同性准则。每一个数据维在数据结构和操作能力上都是等同的，系统可以将附加的操作能力授给所选维，但必须保证该操作能力可以授给任意的其他维，即要求维上的操作是公共的。

准则 7：动态稀疏矩阵处理准则。OLAP 工具的物理模型必须充分适应指定的分析模

型，提供最优的稀疏矩阵处理，这是 OLAP 工具所应遵循的最重要的准则之一。

准则 8：多用户支持能力准则。多用户分析员可以同时工作于统一分析模型上或者在同企业数据上建立不同的分析模型，OLAP 工具必须提供并发访问、数据完整性及安全性机制。

准则 9：非受限的跨维操作。多维数据之间存在固有的关系，这就要求 OLAP 工具能自己推导而不是由最终用户明确定义出相关的计算。对于无法从固有关系中得到的计算，要求系统提供计算完备的语言来定义计算公式。

准则 10：直观的数据处理。这一准则要求数据操纵直观易懂，路径重定位、向上综合、向下挖掘和其他操作都可以通过直观、方便的点拉操作完成。

准则 11：灵活的报表生成。报表必须从各种可能的方面显示出从数据模型中综合出的数据和信息，充分反映数据分析模型的多维特征。

准则 12：非受限的维与维的层次。OLAP 工具的维数不小于 15 维，用户分析员可以在任意给定的综合路径上建立任意多个聚集层次。

然而，E.F.Codd 提出的 OLAP 的 12 条准则只提供了一种数据技术的观点，而不是基准。术语 OLAP 被用来很好地描述为推动公司决策制定、分析设计的数据库和使其所指示的数据仓库的数据能被很容易访问的工具。

3.1.3 多维数据分析的主要概念

OLAP 的目标是满足决策支持或者满足在多维环境下特定的查询和报表需求，它的技术核心是“维”，下面对这个概念和其他相关概念进行介绍。

1. 维（Dimension）

维是人们观察客观世界的角度，是一种高层次的类型划分。维一般包含着层次关系，这种层次关系有时会相当复杂。通过把一个实体的多项重要的属性定义为多个维，使用户能对不同维上的数据进行比较。OLAP 展现在用户面前的是一幅幅多维视图，因此 OLAP 也可以说是多维数据分析工具的集合。例如：企业常常关心产品销售数据随着时间推移而产生的变化情况，这是从时间的角度来观察产品的销售，所以时间是一个维（时间维）；企业也时常关心自己的产品在不同地区的销售分布情况，这是从地理分布的角度来观察产品的销售，所以地理分布也是一个维（地理维），其他还有产品维、顾客维等。

2. 维的层次（Level）

人们观察数据的某个特定角度（即某个维）还可以存在细节程度不同的各个描述方面，我们称多个描述方面为维的层次。一个维往往具有多个层次，例如描述时间维时，可以从日期、月份、季度、年等不同层次来描述，那么日期、月份、季度、年等就是时间维的层次。同样，城市、地区、国家等构成了地理维的层次。

3. 维成员（Member）

维的一个取值称为该维的一个维成员，是数据项在某维中位置的描述。如果一个维是多层次的，那么该维的维成员由各个不同维层次的取值组合而成。例如，我们考虑时间维具有日期、月份、年这三个层次，分别在日期、月份、年上各取一个值组合起来，就得到了时间维

的一个维成员，即“某年某月某日”。一个维成员并不一定在每个维层次上都要取值，例如“某年某月”“某月某日”“某年”等都是时间维的维成员。例如对一个销售数据来说，时间维的维成员“某年某月某日”就表示该销售数据是“某年某月某日”的销售数据。

4. 观察变量

变量是数据的实际意义，即描述数据是“什么”。例如，数据 10 000 本身并没有意义或意义未定，它可能是一个学校的学生人数，也可能是某产品的单价，还可能是某商品的销售量等。在 OLAP 中的观察变量是一个数值型数据。

5. 多维数组

一个多维数组可以表示为(维 1，维 2，…，维 n，变量)。例如：若日用品销售数据是按时间、地区和销售渠道组织起来的三维立方体，加上变量销售额，就组成了一个多维数组(地区，时间，销售渠道，销售额)，如果在此基础上再扩展一个产品维，就得到一个四维的结构，其多维数组为(产品，地区，时间，销售渠道，销售额)。

6. 数据单元(单元格)

多维数组的取值称为数据单元。当多维数据的各个维都选中一个维成员，这些维成员的组合就唯一确定了一个变量的值。那么数据单元就可以表示为(维 1，维 2，…，维 n，变量的值)。例如在产品、地区、时间和销售渠道上各取维成员“笔记本电脑”“上海”“2000 年 1 月”和“批发”后就唯一确定了变量“销售额”的一个值，假设其为 100 000，则该数据单元表示为(笔记本电脑，上海，2000 年 1 月，批发，100 000)。

7. 多维数据集的度量值

前面的变量在实际应用中叫做多维数据集的度量值，这些值应该是数字。度量值是多维数据集的核心值，是最终用户在数据仓库应用中所需要查看的数据，这些数据一般是销售量、成本和费用等。


视频 3-2

3.2 多维数据模型与结构

3.2.1 多维数据的概念模型

多维数据概念模型涉及的核心任务是通过信息包图确定数据仓库的主题和大部分元数据。所要完成的任务是：界定系统边界；确定主要的主题域及其内容。概念模型设计的成果是在原有数据库的基础上建立一个较为稳固的概念模型。

概念模型设计也就是通常所说的需求分析，在与用户交流的过程中，确定数据仓库所需要访问的信息，这些信息包括当前、将来以及与历史相关的数据。在需求分析阶段确定操作数据、数据源以及一些附加数据，设计容易理解的数据模型，有效地完成查询和数据之间的映射。

由于数据仓库的多维性，利用传统的数据流程图进行需求分析已不能满足需要。超立方(Hypercube)用超出三维的表示来描述一个对象，显然具备多维特性，完全可以满足数据仓库的多维特性。利用自上而下方法设计一个超立方体的步骤如下。

（1）确定模型中需要抓住的商业过程，例如销售活动或销售过程。

（2）确定需要捕获的值，例如销售数量或成本，这些信息通常是一些数值。

（3）确定数据的粒度，亦即需要捕获的最低一级的详细信息。

由于超立方体在表现上缺乏直观性，尤其当维度超出三维后，数据的采集和表示都比较困难，因此可以采用一种称为信息包图的方法在平面上展开超立方体，即用二维表格反映多维特征。信息包图提供了一个用多维空间建立用户信息模型的方法，它提供了超立方体的可视化表示。信息包图拥有三个重要对象：指标、维度和类别。指标表明在维度空间衡量商务信息的一种方法，而类别是在一个维度内为了提供详细分类而定义的，其中的成员是为了辨别和区分特定数据而设。

信息包图集中在用户对信息包的需要，它定义主题内容和主要性能测试指标之间的关系，其目标就是为了满足用户需要。利用信息包图设计概念模型需要确定以下三大内容。

（1）确定指标。指标是访问数据仓库的关键所在，是用户最关心的信息。成功的信息包可以保证用户从信息包中获取需要的各个性能指标参数。

（2）确定维度。维度提供了用户访问数据仓库信息的途径，对应超立方体的每一面，位于信息包图的第一行的每一个栏目中。图 3.1 给出了一个合适的贷款分析的信息包图。每一维度作为信息包图上的一个列出现，类别作为信息包图的行给出，图 3.1 共六列（六个相关因素），因此该主题属于六维问题。通过对物流配送业务的需求分析，发现在物流配送业务中，主要关注的问题是货物的调配与运输费用。通过对运输时间、配送车辆的选择、配送货物种类和数量进行分析，可以得到很多重要的信息。因此在与其对应的信息包（见图 3.2）中给出了时间维度、货物维度和车辆维度。

合适的贷款分析（主题）　　维度 →

类别						
时期	地区	货款人	资产负债表	损益表	货款特点
年	省	货款人名字索引 A~Z	年初、年末	净利润	风险利率
季	市	某货款人（某企业）			
月	区				
旬	县				
指标／实际情况、货款额外负担、是否发生货款					

图 3.1　合适的贷款分析对应的信息包图

维度 →

类别	全部时间	全部货物	全部车辆
年	货物分类	车辆类型
月	单个品种	单车
日		
时		
分		
度量指标：运送量、运送费用		

图 3.2　货物调配分析对应的信息包图

（3）确定类别。类别表示一个维度包含的详细信息，一个维度内最底层的可用分类又称为详细类别。

如果在一张平面表格上描述元素的多维性，其中的每一个维度用平面表格的某列表示，通常的维度是时间、地点、产品和顾客，而细化本列的对象就是类别。例如时间维度的类别可以细化到年、月、日，甚至小时。平面表格中的一个元素（对应超立方体中的一个单元格）可以表示：某年某月，在某商店的某类产品的销售额。创建信息包图时需要确定最高层和最底层的信息需求，以便最终设计出包含各个层次需要的数据仓库。对于复杂的商业要求进行需求分析时，有时一张信息包图不能反映所有情况，可能需要设计不同的信息包图来满足全部需求，此时应该保证多个信息包图中出现的维度信息和类别信息完全一致。

3.2.2 多维数据的逻辑模型

数据仓库逻辑模型描述了数据仓库主题的逻辑实现，目前数据仓库的逻辑建模主要采用维度建模。维度建模采用一种直观的标准框架结构来表现数据，并允许进行高性能存取，具有非常好的可扩展性。

以信息包图为核心的多维数据概念模型为多维数据的逻辑设计提供了完备的概念基础。同信息包图中的三个对象对应，星型模型拥有三个逻辑实体：维度、指标和类别。位于星型图中心的实体是指标实体，对应信息包图中的指标对象；位于星型图星角上的实体是维度实体，对应信息包图中的维度对象；而详细类别实体，它对应信息包图中的类别对象。一个维度内的一个单元就是一个类别，代表该维度内的一个单独层次。

1. 星型模型

星型模型（Star Schema）是一种多维的数据模型，它由一个事实表（Fact Table）和一组逻辑上围绕这个事实表的维表（Dimension Table）组成。处在中间的是事实表，事实表是星型模型的核心，用于存放大量的具有业务性质的事实数据，事实表中包含了度量属性和指向周围维表的外码，即事实和外码组合成的事实表主码；维表位于事实表周围，包含一个维的描述信息；事实表中的一个事实指向每个维表中的一个元组。事实表中存放的大量数据，是同主题密切相关的、用户最关心的、对象的度量数据。用户依赖于维表中的维度属性，对事实表中的事实数据进行查询、分析，从而得到支持决策的数据。星型模型的结构图如图 3.3 所示。

通过分析某零售百货连锁店的数据仓库，可以得到其星型模型结构图如图 3.4 所示。

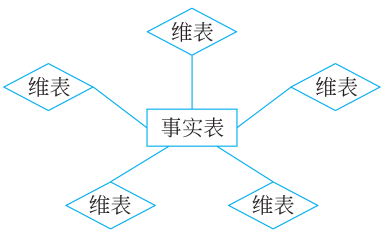

图 3.3 星型模型结构图

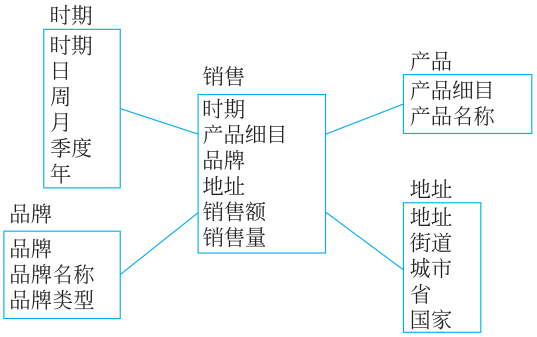

图 3.4 连锁店销售数据仓库星型模型

使用星型模型主要有两方面的原因：

（1）提高查询的效率。采用星型模型设计的数据仓库的优点是由于数据的组织已经过预处理，主要数据都在庞大的事实表中，所以只要扫描事实就可以进行查询，而不必把多个庞大的表连接起来，查询访问效率较高。同时由于维表一般都很小，甚至可以放在高速缓存中，与事实表作连接时其速度较快。

（2）便于用户理解。对于非计算机专业的用户而言，星型模型比较直观，通过分析星型模型，很容易组合出各种查询。

2. 雪花模型

雪花模型（Snowflake Schema）是星型模型的扩展和进一步规范化，结构模式图形类似雪花的形状，维表分解成与事实表直接关联的主维表和与主维表关联的次维表。即维表除了具有星型模型中的维表功能外，还连接上对事实表进行详细描述的详细类别表，通过对事实表在有关维上的详细描述，达到缩小事实表、提高查询效率的目的。雪花模型比星型模型增加了层次结构，体现了维的不同粒度的划分。雪花模型的结构图如图 3.5 所示。

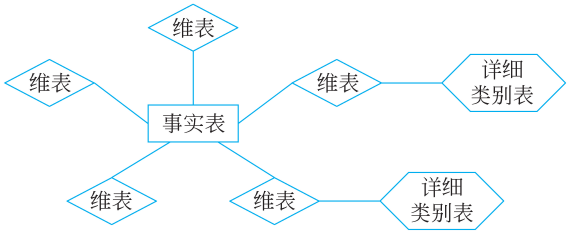

图 3.5　雪花模型结构图

通过分析某零售百货连锁店的数据仓库，可以得到其雪花模型结构图如图 3.6 所示。

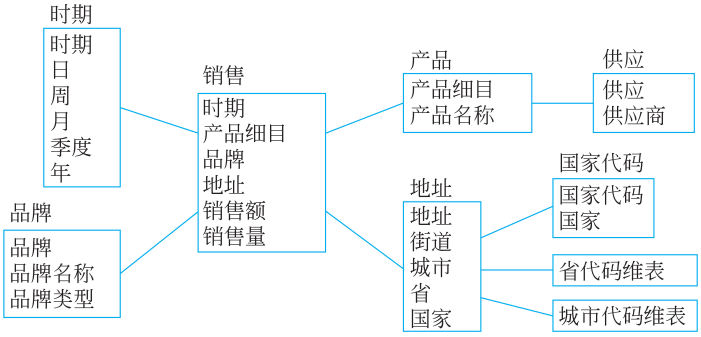

图 3.6　连锁店销售数据仓库雪花模型

雪花模型的优点是：

（1）在一定程度上减少了存储空间；

（2）规范化的结构更容易更新和维护。

雪花模型也存在以下缺点：

（1）雪花模型比较复杂，用户不容易理解；

(2) 浏览内容相对困难；

(3) 额外的连接会使查询性能下降。

3. 星系模型

星系模型(Galaxy Schema)：当多个主题之间具有公共的维时，可以把围绕这些主题组织的星型模型通过共享维表，把事实表相互连接起来。这种多个事实表共享维表的星型模型集称为星系模型。星系模型结构图如图 3.7 所示。

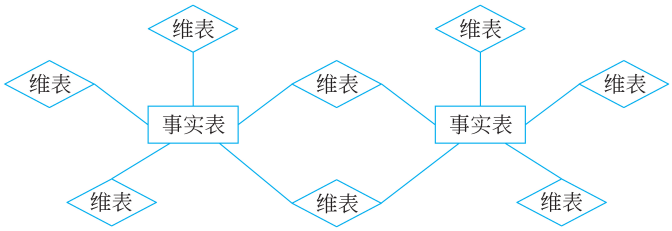

图 3.7　星系模型结构图

通过分析货物销售与配送的过程，得到其星系模型图如图 3.8 所示。

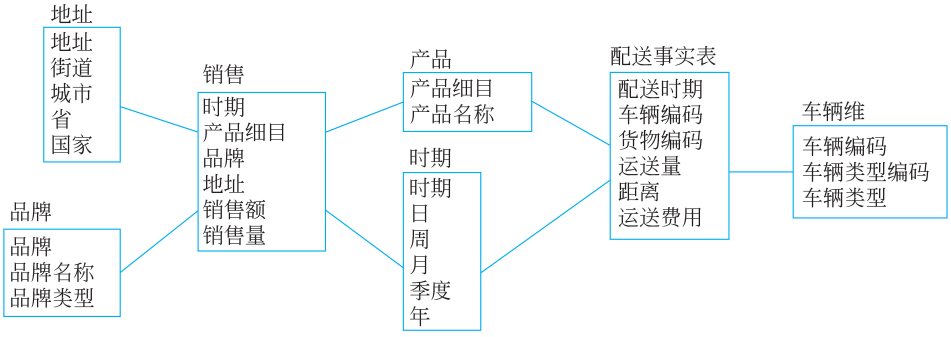

图 3.8　连锁店销售与配送多维数据的星系模型图

虽然星型模型、雪花模型和星系模型这些结构化的多维数据模型都考虑了如何表示多维数据模型中的多维层次结构的问题，但仍具有局限性，雪花模型可表示维层次结构，但要求维层次的路径长度都一样，且同一层次树上的同层节点具有相同的属性集。为了更好地表示数据仓库系统中多维数据的层次结构，需要采用支持不平衡、异构的维层次结构的多维数据模型，充分表达数据仓库的复杂数据结构，并将其作为一种具有普遍适用性和灵活性的多维数据组织的形式化定义与知识描述方法，具体请参考本书 1.3 节“数据仓库系统中多维数据组织的形式化定义与描述”的相关内容。

3.2.3　多维数据的物理模型

物理模型设计的主要任务是确定数据的存储结构、索引策略、数据存放位置及存储分配等。确定数据仓库实现的物理模型，要求设计人员必须做到：全面了解所选用的数据库管理系统，特别是存储结构和存取方法；了解数据环境、数据的使用频度、使用方式、数据规模以及响应时间要求等，这些是对时间和空间效率进行平衡和优化的重要依据；了解外部存储

设备的特性，如分块原则、块大小的规定、设备的 I/O 特性等。

1. OLAP 多维数据结构

OLAP 系统按照其存储器的数据存储格式可以分为关系 OLAP（Relational OLAP，ROLAP）、多维 OLAP（Multi-dimensional OLAP，MOLAP）和混合型 OLAP（Hybrid OLAP，HOLAP）三种类型。

（1）ROLAP 表示基于关系数据库的 OLAP 实现。以关系数据库为核心，以关系型结构进行多维数据的表示和存储。ROLAP 将多维数据库的多维结构划分为两类：一类是事实表，用来存储数据和维关键字；另一类是维表，即对每个维至少使用一个表来存放维的层次、成员类别等维的描述信息。维表和事实表通过主关键字和外关键字联系在一起，形成了“星型模型”。对于层次复杂的维，为避免冗余数据占用过大的存储空间，可以使用多个表来描述，即形成“雪花模型”。ROLAP 将分析用的多维数据存储在关系数据库中并根据应用的需要有选择地定义一批实视图作为表也存储在关系数据库中。不必将每一个 SQL 查询都作为实视图保存，只定义那些应用频率比较高、计算工作量比较大的查询作为实视图。对每个针对 OLAP 服务器的查询，优先利用已经计算好的实视图来生成查询结果以提高查询效率。同时用作 ROLAP 存储器的关系数据库管理系统（Relational DataBase Management System，RDBMS）也针对 OLAP 作相应的优化，比如并行存储、并行查询、并行数据管理、基于成本的查询优化、位图索引、SQL 的 OLAP 扩展（Cube，Rollup）等。

（2）MOLAP 表示基于多维数据组织的 OLAP 实现。以多维数据组织方式为核心，也就是说，MOLAP 使用多维数组存储数据。多维数据在存储中将形成“立方块（Cube）”的结构，在 MOLAP 中对“立方块”的“旋转”“切块”“切片”是产生多维数据报表的主要技术。MOLAP 将 OLAP 分析所用到的多维数据在物理上存储为多维数组的形式，形成“立方块”的结构。维的属性值被映射成多维数组的下标值或下标的范围，而总结数据作为多维数组的值存储在数组的单元中。由于 MOLAP 采用了新的存储结构，从物理层起实现，因此又称为物理 OLAP（Physical OLAP）；而 ROLAP 主要通过一些软件工具或中间软件实现，物理层仍采用关系数据库的存储结构，因此称为虚拟 OLAP（Virtual OLAP）。

（3）HOLAP 表示基于混合数据组织的 OLAP 实现。如低层是关系型的，高层是多维矩阵型的。这种方式具有更好的灵活性。由于 MOLAP 和 ROLAP 有着各自的优点和缺点，且它们的结构迥然不同，给分析人员设计 OLAP 结构提出了难题。因此一个新的 OLAP 结构-混合型 OLAP 被提出，它能把 MOLAP 和 ROLAP 两种结构的优点结合起来。迄今为止，对 HOLAP 还没有一个正式的定义。但很明显，HOLAP 结构不应该是 MOLAP 与 ROLAP 结构的简单组合，而是这两种结构技术优点的有机结合，能满足用户各种复杂的分析请求。

实现 HOLAP 的方法有三种：

① 同时提供多维数据库（Multi-Dimensional DataBase，MDDB）和 RDBMS，让开发人员选择。采用这种方法，开发人员可以选择把信息存放在 MDDB 中或 RDBMS 中，但不能同时存放在 MDDB 和 RDBMS 中。

② 在运行时把对关系数据库的查询结果存入多维数据库。HOLAP 系统利用开发人员定义的一个静态结构的多维模型来暂存在运行时检索出的数据。当客户端提交一个分析

请求时，系统先检查这个多维结构缓存中是否有分析所需要的数据，如果没有，则产生 SQL 语句从 RDBMS 中把相应的数据载入多维数据的缓存中。

③ 利用一个多维数据库存储高级别的综合数据，同时用 RDBMS 存储细节数据。这种方法是如今被认为实现 HOLAP 结构较为理想的方法，它结合了 MOLAP 和 ROLAP 的优点。在该方法中，客户端用户提交一个分析请求，由系统从 MDDB 中提取经过综合的数据或从 RDBMS 提取细节数据。

2. OLAP 多维数据结构的比较

1）存储结构上的比较

在 ROLAP 中对数据进行单项查询时，比较容易处理；但对数据进行钻取时，就比较麻烦了，需要对 ROLAP 的所有数据进行查询，并进行汇总，系统的效率必然降低。而 MOLAP 则只需要对库按行或列进行统计即可，其性能远优于 MOLAP。MOLAP 在 OLAP 系统中的优势，表现在查询速度高和结构清晰明了。但当维数扩展到三维或更高的维度时，成了超立方体的结构，其数据的存储是由许多类似于数组的对象来完成的，这些对象中包含经过压缩的索引和指针，利用这些索引和指针将许多存储数据的单元块联结在一起。实际中，有多维数据的稀疏矩阵问题。MOLAP 在实际应用中的数量存储往往增长较快，尤其在所创建的多维模式中拥有多个维时。但在所增加的空间中有的可能没有实际值出现，会使多维表形成一个稀疏矩阵，因此而浪费大量空间。即使采用各种方法来压缩，也不能根本解决，这势必将造成空间需求爆炸性增长。而 ROLAP 中使用的关系数据库，一般不会出现稀疏矩阵的情况，在实际应用中，只要磁盘空间足够大，ROLAP 数据库可以支持无限增长的数据存储要求，且大多数的多维数据库的容量不能无限增长。由于 ROLAP 中的事实表和维表都要使用二维关系表存放，在多维数据集的构造中，必须通过维表和事实表的联结来实现。

2）数据更新上的比较

MOLAP 需要在建立多维数据库前确定各个维度以及维度的层次关系。在多维数据库建立之后，如果要增加新的维度，则多维数据库通常需要重新建立。而 ROLAP 增加一个维度只是增加一张维表并修改事实表，系统中其他维表不需要修改，因此 ROLAP 对于维度的变更有很好的适应性。由于多维数据通过预综合处理来提高速度，当数据频繁地变化时，MOLAP 需要进行大量的重新计算，甚至重新建立索引，乃至重构多维数据库。而在 ROLAP 中预综合处理通常由设计者根据需求制定，因此灵活性较好，对于数据变化的适应性强。

3）性能上的比较

在 ROLAP 中，多维数据立方体并没有真正存在，通常在接收 OLAP 请求后，ROLAP 服务器需要将 SQL 语句转化为多维存取语句，并利用连接运算拼合出（部分）多维数据立方体，因此，ROLAP 的响应时间较长。MOLAP 是专为 OLAP 设计的，能够自动建立索引，在存取速度上占优势。但是，MOLAP 在预计算，系统响应时间上的优点是通过牺牲存储空间换来的。对于 HOLAP 来说，常用的维度和维层次，使用多维数据表来记录；对于不常用的维度和数据，采用类似于 ROLAP 星型模型来存储。它在存储容量上小于 MOLAP 方式，数据传输速率又低于 MOLAP。其在性能上都介于 MOLAP 和 ROLAP 之间，技术复杂度

高于 ROLAP 和 MOLAP。HOLAP 技术从理论上来说较成熟，而实践中只能根据具体情况来决定应用哪种结构。其决定因素很多，应用规模是一个主要因素。如果需要建立一个大型的、功能复杂的企业级数据仓库，那就可能选择 ROLAP。如果希望建立一个目标单一、维数不是很多的分析型数据集市，那么 MOLAP 可能是一个较佳的选择。

3.3 多维数据分析应用与工具

3.3.1 多维数据分析的基本操作

数据仓库中的多维数据根据其维度可以用立方体或者超立方体表示。如果数据的维度超过三个，我们可以利用立方体的思想建立“超立方体”来表示。多维分析是指对以多维形式组织起来的数据采取多种分析操作，以求剖析数据，使分析者、决策者能从多个角度、多侧面地观察数据库中的数据，从而深入地了解包含在数据中的信息、内涵。这些操作包括切片(Slice)、切块(Dice)、旋转(Rotate)、钻取(Drill)等。多维分析方式迎合了人的思维模式，因此，减少了混淆并且降低了出现错误解释的可能性。

(1) 切片和切块是在一部分维上选定值后，关心度量数据在剩余维上的分布。在多维分析过程中，如果要对多维数据集的某个维选定一维成员，这种选择操作，就可以称为切片。如果对两个或两个以上的维选定维成员，这种选择操作可以称为切块。实际上，切块操作也可以看成是进行多次切片操作以后，将每次切片操作所得到的切片重叠在一起而形成的。在多维数据结构中，按二维进行切片，按三维进行切块，可得到所需要的数据。如在“商店、产品、时间”三维立方体中进行切块和切片，可得到各商店、各产品的销售情况。其中有两个重要的概念必须掌握：一个是多维数据集的切片数量多少是由所选定的那个维的维成员数量的多寡所决定的；另一个是进行切片操作的目的是使人们能够更好地了解多维数据集，通过切片的操作可以降低多维数据集及其维度，使人们能将注意力集中在较少的维度上进行观察。图 3.9 给出了三维数据的切片与切块的示意图。

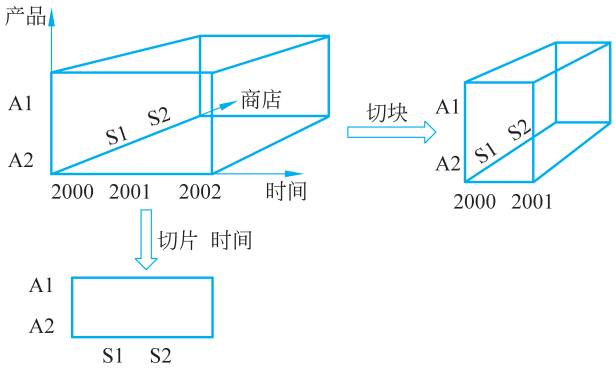

图 3.9 三维数据的切片与切块

(2) 钻取是改变维的层次，变换分析的粒度。维层次实际上反映了数据的综合程度。层次越高，代表数据综合度越高，细节越少。钻取包含向下钻取(Drill-down)和向上钻取

(Drill-up)/上卷(Roll-up)操作,钻取的深度与维所划分的层次相对应。Drill-up 是在某一维上将低层次的细节数据概括到高层次的汇总数据,或者减少维数;而 Drill-down 则相反,它从汇总数据深入到细节数据进行观察或增加新维(见图 3.10)。

地区	销售额
	2002 年
中国	20 000
美国	10 000
日本	24 000

操作
向下钻取 (Drill-down)
向上钻取 (Drill-up)

地区	销售额			
	一季度	二季度	三季度	四季度
中国	5000	6000	4000	5000
美国	2000	3000	3300	1700
日本	4000	7000	6000	7000

图 3.10　OLAP 的钻取操作

(3) 旋转是变换维的方向,即在表格中重新安排维的放置(例如行列互换)。通过旋转可以得到不同视角的数据。

3.3.2　多维数据分析的工具及特点

下面对于 OLAP 工具的特点进行详细介绍。

(1) Cognos 公司的 PowerPlay。①商务绩效评估(Business Performance Measurement,BPM)提供全面的报告和分析环境。②向决策者提供企业运行效率的各种关键数据,进行各种各样的分析。③只用鼠标单击、拖拉就可以浏览多维数据。④自动利用 Web 发布得到的分析报告。⑤支持多种 OLAP Server,如 Microsoft OLAP Services、Hyperion Essbase、SAP BW、IBM OLAP for DB2,拥有完备的授权和安全体系。

(2) Business Objects 公司的 Business Objects(B.O.)。①易用的 BI 工具,允许用户存取、分析和共享数据。②可应用多种数据源,如 RDB、ERP、OLAP、Excel 等。③可应用 VBA 和开放式对象模型来进行开发定制。

(3) Microsoft 公司的 SQL Server OLAP Service。①可以使用任何关系数据库或平面文件作为数据源,其中的 PivotTable Service 提供了客户端的数据缓存和计算能力。②实现 Client/Server 数据管理,提高响应速度,降低网络流量。③通过 OLE DB for OLAP,允许不同的客户端访问。

(4) MicroStrategy 公司的 MicroStrategy 7。①新一代的智能平台(Intelligence Platform),面向电子商务应用 e-business 和电子客户关系管理(electronic Customer Relationship Management,eCRM)。②具有强大的分析能力。③以 Web 为中心的界面。④支持上百万的用户和 TB 的数据,拥有快速开发能力,可直接利用已有的数据模式。

(5) Oracle DW 公司的 Express Serve。①Oracle 支持 GB～TB 数量级。②采用类似数组的结构,避免了连接操作,提高分析性,能提供一组存储过程语言来支持对数据的抽取;用户可通过 Web 和电子表格使用;具有灵活的数据组织方式,数据可以存放在 Express Server 内,也可直接在 RDB 上使用;有内建的分析函数和 4GL,用户可自己定制查询。

(6) IBM 公司的 DB2 OLAP Server。①强大的多维分析工具,把 Hyperion Essbase 的 OLAP 引擎和 DB2 的关系数据库集成在一起。②与 Essbase API 完全兼容。③数据用星型模型存放在关系数据库 DB2 中。

(7) Essbase 公司的 Hyperion Essbase。①以服务器为中心的分布式体系结构。②有超过 100 个的应用程序。③有 300 多个用 Essbase 作为平台的开发商。④具有几百个计算公式,支持多种计算。⑤用户可以自己构建复杂的查询。⑥快速的响应时间,支持多用户同时读写。⑦有 30 多个前端工具可供选择。⑧支持多种财务标准。⑨能与 ERP 或其他数据源集成。

(8) Informix 公司的 Informix Metacube。①采用 meta cube 技术,通过 OLE 和 ODBC 对外开放。②采用中间表技术实现多维分析引擎,提高响应时间和分析能力。③开放的体系结构可以方便地与其他数据库及前台工具进行集成。

(9) Sybase 公司的 Power dimension。①数据垂直分割(按"列"存储)。②采用了突破性的数据存取方法-bit-wise 索引技术。③在数据压缩和并行处理方面有独到之处。④提供有效的预连接(Pro-Join)技术。

(10) Brio.Enterprise 公司的 Brio Enterprise。①强大的易用的 BI 工具,提供查询、OLAP 分析和报告的能力。②支持多种语言,包括中文。③Brio.Report 是强大的企业级报告工具。

3.4 从联机分析处理到联机分析挖掘

联机分析挖掘(On-Line Analysis Mining,OLAM)是联机分析处理技术与数据挖掘技术在数据库或数据仓库应用中的结合,是联系分析处理技术的新发展,也是近年来数据库领域的研究重点和热点。

3.4.1 联机分析挖掘形成原因

OLAP 与 DM 虽同为数据库或数据仓库分析工具,但两者的侧重点不同。同时,随着 OLAP 与 DM 技术的应用和发展,数据库领域在 OLAP 基础上对深层次分析的需求与人工智能领域的数据挖掘技术的融合最终促成了联机分析挖掘技术。

一方面,分析工具 OLAP 功能虽强大,能为客户端应用程序提供完善的查询和分析,但它也存在不足,由于 OLAP 是一种验证性分析工具,是由用户驱动的,这很大程度上受到用户的假设能力的限制。OLAP 分析事先需要对用户的需求有全面而深入的了解,然而用户的需求是不确定的,难以把握,所以 OLAP 分析常常采用试凑法搜索数据仓库,耗时多而且易产生一些无用的结果。另一方面,数据挖掘可以使用复杂算法来分析数据和创建模型来表示有关数据的信息,用户不必提出确切的要求,系统就能够根据数据本身的规律性,自动挖掘数据潜在的模式,或通过联想建立新的业务模型以辅助决策。但数据挖掘存在一些缺点:如 DM 由数据驱动,用户需要事先提出挖掘的任务,但很多时候是不能预先知道要挖掘什么样的知识的。若用户仅提出挖掘任务,DM 工具就遍历整个数据库,将导致搜索空间太大。即使挖掘出了潜在有价值的信息,但它究竟用来做什么分析用,用户也可能不太清楚。

可将 OLAP 与 DM 结合使用。OLAP 的分析结果可以补充到系统知识库中,为数据挖掘提供分析依据;数据挖掘发现的知识可以指导 OLAP 的分析,拓展 OLAP 分析的深度,以便发现 OLAP 所不能发现的更为复杂、细致的信息。不可否认,两者各有长处,也各有不

足。OLAP 缺乏灵活性、准确性，而数据挖掘实施代价高昂、实现困难。针对两者的优缺点，人们提出了 OLAM。OLAM 综合了 OLAP 和数据挖掘的功能，兼有 OLAP 多维分析的在线性、灵活性和数据挖掘对数据处理的深入性。借助 OLAM，用户既可在多维数据库的不同部位和不同抽象级别交互地执行挖掘，又可以灵活选择所需要的数据挖掘功能，并动态交换数据挖掘任务。

3.4.2 联机分析挖掘概念及特征

1. 联机分析挖掘的概念

联机分析挖掘将联机分析处理与数据挖掘以及在多维数据库中发现的知识集成在一起，提供在不同的数据子集和不同的抽象层上进行数据挖掘的工具。联机分析挖掘为用户选择所期望的数据挖掘功能、动态修改挖掘任务提供了灵活性。在数据仓库的基础上提供更有效的决策支持，鉴于 OLAP 与 DM 技术在决策分析中的这种互补性，促成了 OLAM 技术的形成，其中所包含的关键技术可用如下公式表达：联机分析挖掘(OLAM)＝数据仓库(DW)＋联机分析处理(OLAP)＋数据挖掘(DM)。

但 OLAM 不是这三种技术的单纯叠加，而是多种技术的无缝集成，这种集成将带来 OLAM 技术与其构件技术在基本概念、原理、技术、方法、机制、结构、使用等方面本质上的不同。OLAM 建立在多维数据视图的基础之上，基于超立方体的挖掘算法是其核心所在。超立方体计算与传统挖掘算法的结合使得数据挖掘有了极大的灵活性和交互性。这里所说的立方体计算方法一般指切片、切块、钻取、旋转等操作；而挖掘算法则是指关联、分类、聚类等基于关系型或事务型的挖掘算法。

根据立方体计算和数据挖掘所进行的次序不同组合可以有以下一些模式：

(1) 先进行立方体计算、后进行数据挖掘。在进行数据挖掘以前，先对多维数据进行一定的立方体计算，以选择合适的数据范围和恰当的抽象级别。

(2) 先对多维数据作数据挖掘，然后利用立方体计算算法对挖掘出来的结果作进一步的深入分析。

(3) 立方体计算与数据挖掘同时进行。在挖掘的过程中，可以根据需要对数据视图作相应的多维操作。这也意味着同一个挖掘算法可以应用于多维数据视图的不同部分。

(4) 回溯操作。OLAM 的挖掘过程是对多维数据视图的一个不断深入的过程。OLAM 的标签的回溯特性，允许用户回溯一步或几步，或回溯至标志处，然后沿着另外的途径进行挖掘，这样用户在挖掘分析中可以交互式地进行立方体计算和数据挖掘。

联机分析处理概念正式提出是在 1997 年，Jiawei Han 教授等在数据立方体的基础上提出多维数据挖掘的概念。这实际上是在 OLAP 系统的基础上，把数据分析算法、数据挖掘算法引进来，解决多维数据环境的数据挖掘问题。因此这时的 OLAM 实际上还是 OLAP 和 DM 的松散结合。之后，国内外研发人员在这方面展开了积极的工作，试图将 OLAP 和 DM 技术有机结合起来形成真正的 OLAM 技术和产品。其分析和挖掘的数据基础也扩大到包括多维数据模型和关系数据模型等在内的多种模型的异构环境，研究重点是如何实现 OLAP 和 DM 技术紧密集成，即针对在异构大数据量的环境中快速响应用户的数据分析和数据挖掘请求的问题进行深入研究。

2. 联机数据挖掘的功能特征

OLAM 融合了三种技术，兼有 OLAP 和 DM 的优点，在 DW 上的数据挖掘和分析更具有灵活性和交互性。其功能特征包括：

（1）相对 OLAP 和 DW 技术，OLAM 具有较高的执行效率和较快的响应速度。

（2）OLAM 能对任何它想要的数据进行挖掘。OLAM 建立在 OLAP 基础上，因此能方便地对任何一部分数据或不同抽象级别的数据进行挖掘，甚至还可以直接访问存储在底层数据库里的数据。

（3）在 OLAM 中，用户可以动态选择或添加挖掘算法，并可以动态切换挖掘任务。

（4）OLAM 中挖掘任务具有多样性，算法具有复杂性，因此应具有标签和回溯的功能。标签功能即标记用户的操作状态功能，回溯指的是退回上次操作状态。OLAM 这种功能可以避免用户因算法的复杂性而在超立方体中“迷失方向”。

（5）OLAM 具有灵活的可视化工具。可视化工具以丰富的图文有效地显示分析和挖掘结果给用户，从而实现交互式处理。

（6）良好的扩展性。这是指 OLAM 应该高度模块化，能与其他多个子系统集成。

（7）友好的人际交互能力。OLAM 的决策分析过程是要在人的指导下进行，人作为系统的组成部分和系统应用密不可分。人与计算机分别承担各自最擅长的工作，实现资源的合理配置。

思 政 园 地

用户画像算法与思政教育的融合

用户画像算法是一种通过收集和分析用户多维度数据，以构建用户模型的技术手段，形成一个立体、多维的用户形象，为思政教育提供新的思路和方法。

精准定位与个性化教育：用户画像算法能够基于学生的多维度数据（如学习行为、兴趣偏好、社交活动等），构建每个学生的个性化模型，这有助于教育者更精准地识别学生的需求和特点，因材施教，从而提供更具针对性的教育内容和方式。

数据分析与行为预测：用户画像算法通过对学生数据的深入分析，可以揭示学生的学习习惯、兴趣爱好、心理状态等方面的规律和特征，还能在一定程度上预测学生的未来行为，这有助于更好地了解学生和评估教育效果，及时采取措施进行干预和引导。

思政元素融入与强化：将思政元素（如爱国情怀、社会责任、道德观念等）融入学生综合素养的用户画像构建之中，量化和分析思政表现，更加直观地评估学生的思政素养水平，对于薄弱的方面加强相关主题的教育和引导，对于表现突出的方面树立为榜样进行表彰和宣传。

课 后 习 题

1. 判断题

（1）联机分析处理（OLAP）主要用于支持复杂的查询处理、数据分析以及对大量历史数据的快速报表制作，它侧重于快速准确地回答用户对数据的多维查询需求，而不是事务处

理或数据录入。（　　）

（2）在评价联机分析处理（OLAP）系统时，响应时间是最关键的评价指标，其他因素如数据一致性、易用性、可扩展性等相对不重要。（　　）

（3）在多维数据的逻辑模型中，每个维度只能有一个层级结构，用来组织不同细节级别的数据。（　　）

（4）在多维数据的物理模型中，星型模型总是比雪花模型更优，因为它提供了更快的数据访问速度和更简单的查询处理。（　　）

（5）在多维数据分析中，“旋转”操作是指改变数据立方体中维度的方向，而不改变其内容和聚合度量值。（　　）

2. 单选题

（1）OLAP 系统中最常见的数据访问模式是（　　）。

A. 在线事务处理（OLTP）　　B. 批量数据加载

C. 快速多维分析　　D. 实时数据录入

（2）在 OLAP 系统中，哪个特性特别强调了系统能够支持用户在不同抽象级别上查看数据的能力？（　　）

A. 可钻取性（Drill-down/Up）　　B. 切片与切块（Slice and Dice）

C. 旋转（Pivot）　　D. 多维性（Multidimensionality）

（3）在多维数据模型中，哪个核心组件代表了用户所关心的度量或数值型数据，如销售额、利润等？（　　）

A. 维度（Dimension）　　B. 层级（Hierarchy）

C. 计量器（Measure）　　D. 事实（Fact）

（4）在多维数据仓库的物理实现中，哪种存储模式特别适用于存储大量的预计算汇总数据，以提高查询性能？（　　）

A. 关系数据库　　B. ROLAP

C. MOLAP　　D. Hybrid OLAP（HOLAP）

（5）下列哪项操作是多维数据分析中用于细化数据视图，即从汇总数据深入到更细节级别的数据？（　　）

A. 切片　　B. 切块　　C. 钻取　　D. 转轴

3. 多选题

（1）联机分析处理（OLAP）技术的核心特点包括（　　）。

A. 支持多维数据分析，允许用户从不同角度查看数据

B. 强调数据的实时性和事务处理速度

C. 便于进行复杂查询和数据分析，而非简单的数据检索

D. 适用于决策支持和商业智能，提供汇总和明细级别数据

E. 数据存储采用星型或雪花模型，优化查询性能

（2）以下哪些是评价 OLAP 系统性能和效用的重要标准？（　　）

A. 查询响应时间　　B. 数据更新的及时性

C. 系统的可扩展性和稳定性　　D. 用户界面的友好程度

E. 多维度分析的灵活性

(3) 多维数据模型中,以下哪些元素是构成其基本架构的关键部分?(　　)

A. 维度(Dimensions)　　B. 事实表(Fact Tables)

C. 度量(Measures)　　D. 级别(Levels)

E. 关键性能指标(KPIs)

(4) 多维数据物理模型中,以下哪些组件是构建星型模型时通常会包含的?(　　)

A. 事实表　　B. 维度表　　C. 桥接表　　D. 链接表

E. 属性表

(5) 在多维数据分析中,以下哪些是基本的操作,可以帮助用户从不同角度和层次探索数据?(　　)

A. 切片(Slice)　　B. 切块(Dice)

C. 钻取(Drill-down/Drill-up)　　D. 旋转(Pivot)

E. 平移(Shift)

4. 简答题

(1) 说明 OLAP 技术的定义、特点和评价准则。

(2) 解释 OLAP 多维数据结构的三种类型和比较。

(3) 列举流行的 OLAP 工具和对应的特点。

(4) 阐述联机分析挖掘产生的原因、概念和特征。

(5) 阐述多维数据分析的基本操作。

5. 案例题

请根据连锁店销售数据仓库画出星型模型、雪花模型和星系模型图。

第 4 章　预测模型研究与应用

本章对预测模型展开深入的探讨，指出预测方法的分类和建模的一般步骤；重点阐述了四类典型的预测方法的数学模型和实例应用，包括一元线性回归、多元线性回归、非线性回归预测模型，珀尔、冈珀茨、华德诺尔三种趋势外推预测模型，移动平均、指数平滑和季节指数三类时间序列预测模型，基于神经网络的预测模型，马尔可夫预测模型。

4.1　预测模型的基础理论

4.1.1　预测方法的分类

按预测目标范围的不同，可分为宏观预测和微观预测，宏观经济预测是指对整个国民经济或一个地区、一个部门的经济发展前景的预测；而微观经济预测是以单个经济单位的经济活动前景作为考察的对象。按预测期限长短不同，可分为长期预测、中期预测和短期预测；按预测结果的性质不同，可分为定性预测与定量预测。

1. 定性预测

主要是根据事物的性质和特点以及过去和现在的有关数据，对事物做非数量化的分析，然后根据这种分析对事物的发展趋势做出判断和预测。定性预测在很大程度上取决于经验和专家的努力，依靠人们的主观判断来取得预测结果。其特点为：简单易行、花费时间少、应用历史较久。当缺乏统计数据，不能构成数学模型或环境变化很大，历史统计数据的规律无法反映事物变化规律时一般用定性预测。主要有以下几种方法：用户意见法（对象调查法）、员工意见法、个人判断、专家会议、特尔菲法、主观概率法、类推法、目标分解法等。这些方法在一定程度上存在片面性、准确度不太高的缺点，可以作为定性预测的辅助方法。

2. 定量预测

定量预测主要利用历史统计数据并通过一定的数学方法建立模型，以模型为主对事物的未来做出判断和预测的数量化分析，也称客观预测。本书所采用的定量预测模型体系结构如图 4.1 所示。

本章后几节将详细介绍定量预测方法中的回归分析、时间序列分析、趋势外推法、马尔可夫预测等方法。

4.1.2　预测方法的一般步骤

（1）预测目标分析和确定预测期限：确定预测目标和预测期限是进行预测工作的前提。

（2）进行调研，收集资料：预测以一定的资料和信息为基础，以预测目标为中心收集充分、详尽、可靠的资料。同时要去伪存真，去掉不真实和与预测对象关系不密切的资料。

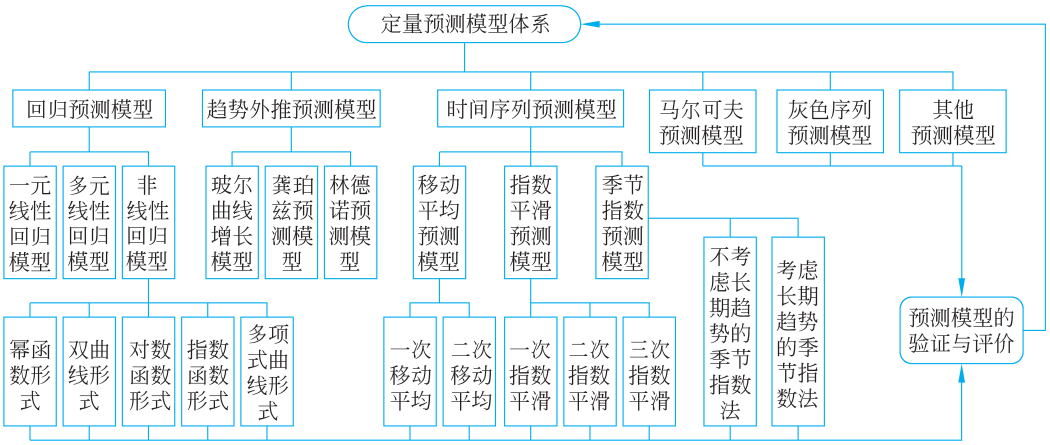

图 4.1　定量预测模型体系结构

(3) 选择合适的预测方法：分别研究当前预测理论领域的各种预测模型和预测方法。预测方法的选取应服从预测的目的和资料、信息的条件。同时使用多种预测方法独立地进行预测，并对各种预测值分别进行合理性分析与判断。

(4) 考虑模型运行平台：依据预测理论和预测方法，选择合适的数据库和编程语言实现预测模型系统。

(5) 对预测的结果进行分析和评估：考核预测结果是否满足预测目标的要求，对各种预测模型进行相关检验，比较预测精确度。根据不同模型的拟合效果和精度，选取精度较高和拟合效果好的模型。

(6) 模型的更新：应该根据最新的管理、经济动态和新到来的信息数据，重新调整原来的预测模型以提高预测的准确性。

4.2　回归分析预测模型

视频 4-1

4.2.1　一元线性回归预测模型

一元线性回归分析是处理两个变量 x(自变量)和 y(因变量)之间关系的最简单模型，研究的是这两个变量之间的线性相关关系。通过该模型的讨论，不仅可以掌握有关一元线性回归的理论知识，而且可以从中了解回归分析方法的数学模型、基本思想、方法及应用。

1. 数学模型

1) 一元回归公式

以影响预测的各因素作为自变量或解释变量 x 和因变量或被解释变量 y 有如下关系：

$$y_i = a + bx_i + u_i \quad i = 1,2,\cdots,n \tag{4.1}$$

式(4.1)称为一元线性回归模型(One Variable Linear Regression Model)，其中 u 是一个随机变量称为随机项；a、b 是两个常数，称为回归系数(参数)；i 表示变量的第 i 个观察值，共有 n 组样本观察值。

2）建立模型与相关检验

（1）参数的最小二乘估计。

相应于 y_i 的估计值 $\hat{y}_i=\hat{a}+\hat{b}\,x_i$，$y_i$ 与 $\hat{y}_i$ 之差称为估计误差或残差，以 ℓ_i 表示，$\ell_i=y_i-\hat{y}_i$。显然，误差 ℓ_i 的大小是衡量估计量 $\hat{a}$、$\hat{b}$ 好坏的重要标志，以误差平方和最小作为衡量总误差最小的准则，并依据这一准则对参数 a、b 作出估计。令

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}\ell_i^2=\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}\,x_i)^2 \tag{4.2}$$

使 Q 达到最小以估计出 $\hat{a}$、$\hat{b}$ 的方法称为最小二乘法（Method of Least-Squares）。由多元微分学可知，使 Q 达到最小的参数的 $\hat{a}$、$\hat{b}$ 的最小二乘估计量（Least-Squares Estimator of Regression Coefficient）必须满足：

$$\begin{cases}\dfrac{\partial Q}{\partial \hat{a}}=-2\sum\limits_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)=0\\[2ex]\dfrac{\partial Q}{\partial \hat{b}}=-2\sum\limits_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)x_i=0\end{cases}\qquad(i=1,2,\cdots,n) \tag{4.3}$$

解上述方程组得

$$\hat{b}=\frac{\sum\limits_{i=1}^{n}x_iy_i-\bar{x}\sum\limits_{i=1}^{n}y_i}{\sum\limits_{i=1}^{n}x_i^2-\bar{x}\sum\limits_{i=1}^{n}x_i}=\frac{\sum\limits_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2},\quad \hat{a}=\bar{y}-\hat{b}\,\bar{x} \tag{4.4}$$

其中：$\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$，$\bar{y}=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i$。

若令

$$l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2,l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2,l_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

则式(4.4)可以写成

$$\begin{cases}\hat{a}=\bar{y}-\hat{b}\bar{x}\\[1ex]\hat{b}=\dfrac{l_{xy}}{l_{xx}}\end{cases}$$

（2）相关性检验。

一般情况下，在一元线性回归时，用相关性检验较好，样本相关系数（Sample Correlation Coefficient）R 是描述变量 x 与 y 之间线性关系密切程度的一个数量指标。

$$R=\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}}\quad(-1\leqslant R\leqslant 1) \tag{4.5}$$

查相关系数临界值表，若 $R>R_\alpha(n-2)$，则线性相关关系显著，通过检验，可以进行预测；反之，没有通过检验，该一元回归方程不可以作为预测模型。