上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

数据挖掘和OLAP

饶艳超 副教授

上海财经大学会计学院

raoyanchao@qq.com

学习目标

- 熟悉数据挖掘的定义、特征、分类
- 熟悉数据挖掘技术工具和相关产品
- 学会分析应用数据挖掘技术
- 熟悉OLAP的定义、特征、基本操作和实现方法
- 学会分析应用OLAP

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

1

数据挖掘

- **数据挖掘概述**
  - 数据挖掘作为一项从海量数据中提取知识的信息技术是一个“以发现为驱动”的过程，已经引起了学术界和产业界的极大重视。
  - 数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。

- **数据挖掘概述**
  - **数据挖掘的定义和特征**
    - 数据挖掘是一门受到来自各种不同领域的研究者关注的交叉性学科，因此导致了很多不同的术语名称。
    - 最常用术语是“知识发现”和“数据挖掘”。
    - **数据挖掘**主要流行于统计界（最早出现于统计文献中）、数据分析、数据库和管理信息系统界；
    - **知识发现**则主要流行于人工智能和机器学习界。

- **数据挖掘概述**
  - 数据挖掘是从大量数据中寻找其规律的技术，主要有数据准备、规律寻找和规律表示三个步骤。
    - **数据准备**是从各种数据源中选取和集成用于数据挖掘的数据；
    - **规律寻找**是用某种方法将数据中的规律找出来；
    - **规律表示**是用尽可能符合用户习惯的方式（如可视化）将找出的规律表示出来。
  - 但在具体实施数据挖掘应用时，还要有一个步骤就是结果评价。
    - 因为数据算法寻找出来的是数据的规律，其中有些是人们感兴趣的有用的，还有一些可能是不感兴趣的没有用的，这就要对寻找出的规律进行评估。

- 数据挖掘概述
  - 数据挖掘的分类
    - 根据数据挖掘的任务分
      - 分类或预测模型数据挖掘、数据总结、数据聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等等。
    - 根据数据挖掘的对象分
      - 关系数据库、面向对象数据库、空间数据库、时态数据库、文本数据源、多媒体数据、异质数据库、遗留系统（legacy system）数据库，以及Web数据源

- 数据挖掘概述
  - 数据挖掘的分类
    - 根据数据挖掘的方法分
      - 统计方法、机器学习方法、神经网络方法和数据库方法。
      - （1）统计方法则可细分为：回归分析（多元回归、自回归等）、判别分析（贝叶斯判别、费歇尔判别、非参数判别等）、聚类分析（系统聚类、动态聚类等）、探索性分析（主元分析法、相关分析法等）、以及模糊集、粗糙集、支持向量集等。

- 数据挖掘概述
  - 数据挖掘的分类
    - 根据数据挖掘的方法分
      - （2）机器学习方法可细分为：归纳学习方法（决策树、规则归纳等）、基于范例的推理CBR、遗传算法、贝叶斯信念网络等。
      - （3）神经网络方法可细分为：前向神经网络（BP算法等）、自组织神经网络（自组织特征映射、竞争学习等）等。
      - （4）数据库方法主要是基于可视化的多维数据分析或OLAP方法，另外还有面向属性的归纳方法。

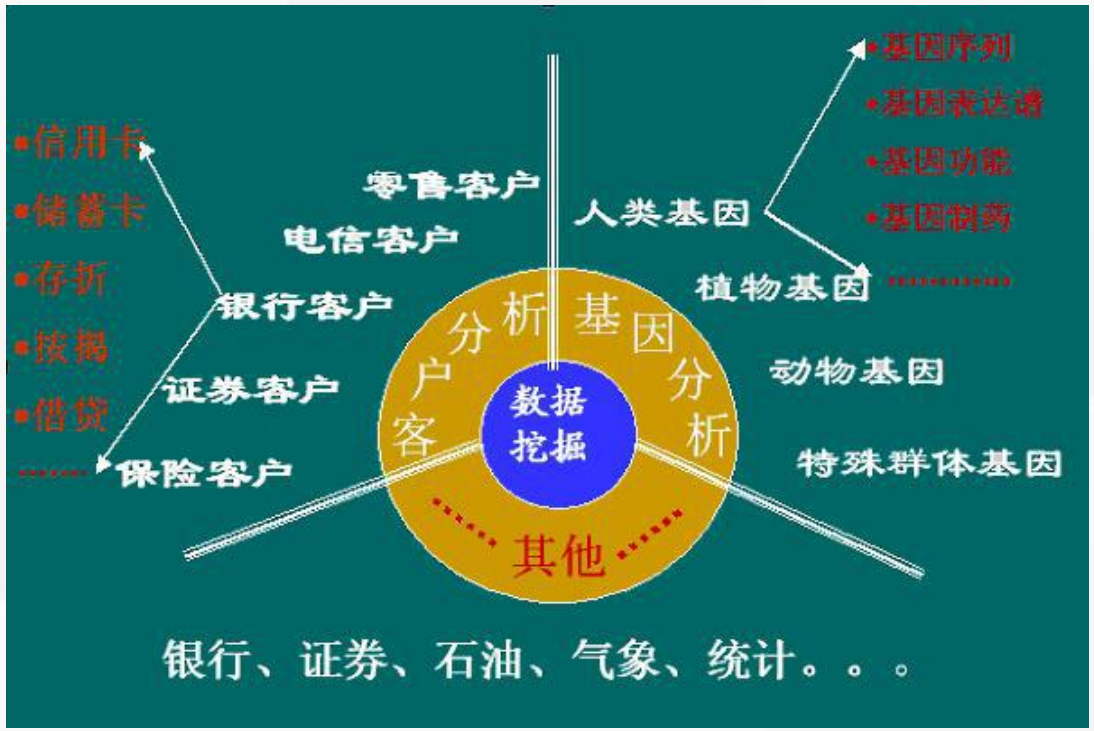
- 数据挖掘技术和工具
  - 数据挖掘技术
    - 关联分析
      - 寻找数据项之间感兴趣的关联关系。
    - 演变分析
      - 描述时间序列数据随时间变化的数据的规律或趋势，并对其建模。包括时间序列趋势分析、周期模式匹配等。
    - 分类分析
      - 找出描述并区分数据类的模型（可以是显式或隐式），以便能够使用模型预测给定数据所属的数据类。

- 数据挖掘技术和工具
  - 数据挖掘技术
    - 聚类分析
      - 根据最大化类内的相似性、最小化类间的相似性的原则将数据对象聚类或分组，所形成的每个簇（聚类）可以看作一个数据对象类，用显式或隐式的方法描述它们.
      - 聚类和分类有着很大的区别：分类时，我们总是事先知道哪些属性是重要的，如企业总是将重要的、有影响力的属性作为分类的依据；而聚类时，企业事先根本不知道哪些属性在起作用，聚类分析的任务之一就是要找到那些起关键作用的属性。

- 数据挖掘技术和工具
  - 数据挖掘技术
    - 异常分析
      - 一个数据集中往往包含一些特别的数据，其行为和模式与一般的数据不同，这些数据称为异常。
      - 对异常数据的分析称为异常分析，在欺诈甄别、网络入侵检测等领域有着广泛的应用。

- 数据挖掘技术和工具
  - 数据挖掘工具
    - 单算法挖掘工具
    - 数据挖掘算法工具集
    - 数据挖掘解决方案
  - 主要数据挖掘工具产品
    - 主要的数据挖掘工具实验系统
    - 主要的商业数据挖掘工具

- **数据挖掘技术的应用**
  - 数据挖掘应用领域非常广阔，先期主要在数据积累比较充分的领域如银行、证券、电信等领域应用，然后在各行各业各领域中逐步获得应用。



- **数据挖掘概述**
  - **以银行为例**
    - 银行业的信息基础建设，网络平台；
    - **数据大集中** ：系统内的所有的交易和管理集中
    - 在数据大集中的基础上，利用数据挖掘技术建立起有效的数据集成、管理、利用机制，建立商业银行数据挖掘软件系统，充分挖掘数据价值，为银行发展新的业务服务和科学化管理提供决策支持。
      - **客户分析系统**
        - 风险管理活动
        - 信用评分活动
        - 争夺客户活动
        - 客户流失分析

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

2

OLAP

- OLAP
  - 联机分析处理（OLAP）的概念
    - 由关系数据库之父E. F. Codd 于1993年提出的，OLAP是使分析人员、管理人员或执行人员能够从多角度对信息进行快速、一致、交互地存取，从而获得对数据的更深入了解的一类软件技术。
    - 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” 。
    - 通过把一个实体的多项重要的属性定义为“多个维”，可以使用户能对不同维上的数据进行比较。

- OLAP
  - 证券公司股票交易量统计分析
    - **计算股票交易量的前提条件**很多，不同营业部、不同委托方式、不同的交易所、不同的币种、不同的证券类别、不同的客户规模、不同的客户年龄段、不同的月份日期等等，计算出来的交易量都不一样，如何区分这些不同的前提条件下的不同结果？
    - **多维数据库：**一个多维的数据库将每一项数据特征（例如营业部、委托方式、交易所、币种等）都看作一维。

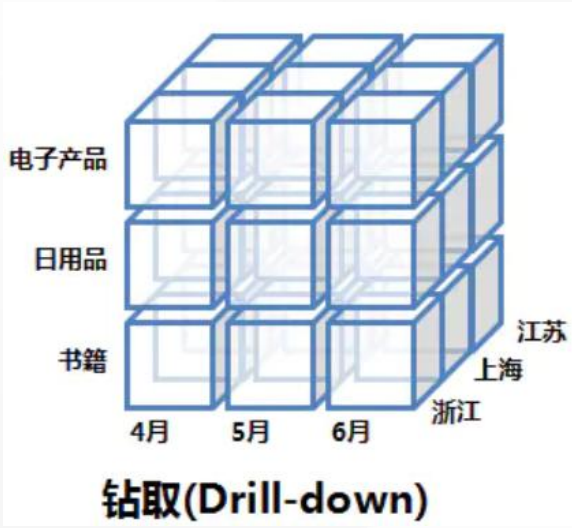
- OLAP
  - OLAP的特点
    - （1）实时性要求不是很高
    - （2）数据量大
    - （3）OLAP的重点在于决策支持

- OLAP
  - OLAP的特点
    - 联机分析处理（OLAP）使得用户可以简单并有选择性的从不同的视点提取并查看数据。
    - OLAP数据被存储在多维的数据库中，OLAP软件可以找到不同维之间的交集（例如一定时间以内在东部区域以高于某一价格售出的产品）并显示出来。
    - OLAP可以被用于数据挖掘以及发掘数据项之间未发现的关系。

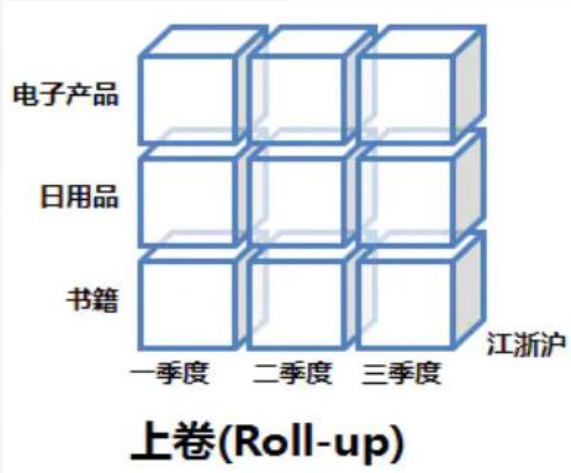
- OLAP
  - OLTP vs. OLAP

	OLTP	OLAP
用户	操作人员,低层管理人员	决策人员,高级管理人员
功能	日常操作处理	分析决策
DB 设计	面向应用	面向主题
数据	当前的, 最新的细节的, 二维的分立的	历史的, 聚集的, 多维的集成的, 统一的
存取	读/写数十条记录	读上百万条记录
工作单位	简单的事务	复杂的查询
用户数	上千个	上百个
DB 大小	100MB-GB	100GB-TB

- OLAP
  - OLAP的基本操作
    - 钻取（drill down）：在维的不同层次间的变化，从上层降到下一层，或者说是将汇总数据拆分到更细节的数据，比如通过对2010年第二季度的总销售数据进行钻取来查看2010年第二季度4、5、6每个月的消费数据；当然也可以钻取浙江省来查看杭州市、宁波市、温州市……这些城市的销售数据。
    - 上卷（roll up）：钻取的逆操作，即从细粒度数据向高层的聚合，如将江苏省、上海市和浙江省的销售数据进行汇总来查看江浙沪地区的销售数据

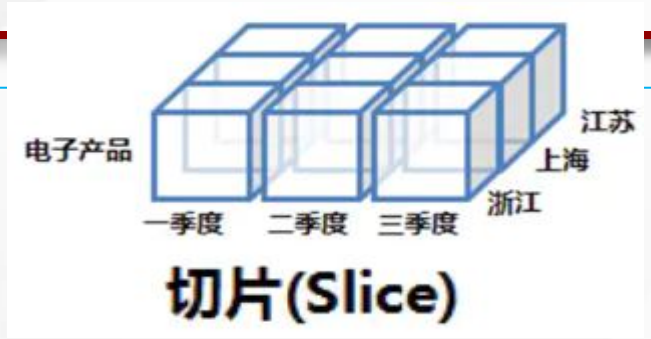


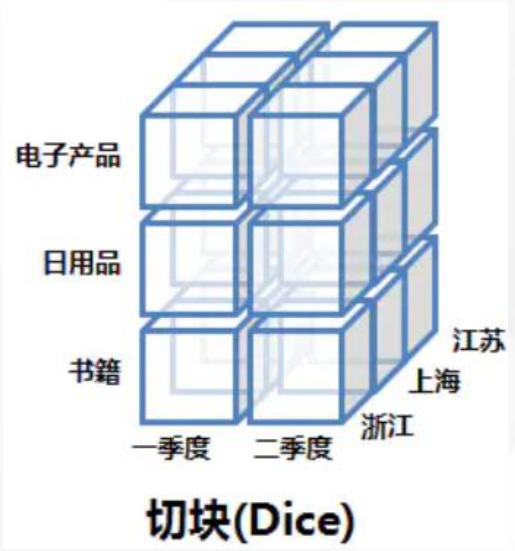
钻取(Drill-down)

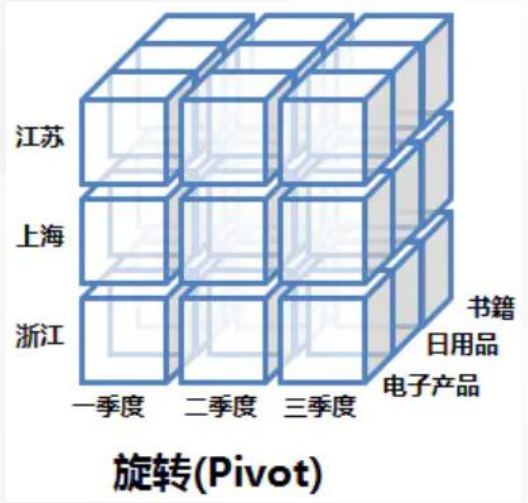


上卷(Roll-up)

- OLAP
  - OLAP的基本操作
    - 切片（slice）：选择维中特定的值进行分析，比如只选择电子产品的销售数据，或者2010年第二季度的数据。
    - 切块（dice）：选择维中特定区间的数据或者某批特定值进行分析，比如选择2010年第一季度到2010年第二季度的销售数据，或者是电子产品和日用品的销售数据。从所有的维度中切出一个范围较小、维度相同的一个小立方体。
    - 旋转（Pivot）：即维的位置的互换，就像是二维表的行列转换，如通过旋转实现产品维和地域维的互换。







- OLAP
  - 交叉探查（drill across）：如“销售表”和“库存表”中有相同的维度，“日期维度”、“产品维度”和“商场维度”，如果有个需求是想按共有维度来对比查看销售和库存，这时就需要发出两个SQL，分别查出按维度统计出的销售数据和库存数据。然后再基于共有的维度进行外连接，将数据合并。这种发出多路SQL再进行合并的操作就是交叉探查。
  - 穿透钻取（drill through）：在一个会话内从一个报表跳转到另一个报表，同时将焦点保持在相同的数据上。如在销售表中选择一种产品，并跳转到有关该产品的库存表中。
    - IBM Cognos Analytics使用穿透钻取访问;
    - 例：Cognos Report Studio报表之间穿透钻取功能的实现

- OLAP
  - OLAP的实现方法
    - OLAP有多种实现方法，根据存储数据的方式不同可以分为：

名称	描述	细节数据存储位置	聚合后的数据存储位置
ROLAP(Relational OLAP)	基于关系数据库的OLAP实现	关系型数据库	关系型数据库
MOLAP(Multidimensional OLAP)	基于多维数据组织的OLAP实现	数据立方体	数据立方体
HOLAP(Hybrid OLAP)	基于混合数据组织的OLAP实现	关系型数据库	数据立方体

- OLAP应用和工具
  - OLAP应用分析要点
    - 1．从动态的多维角度分析数据
      - OLAP将数据分为两种特征
      - **表现特征**——比如一个销售分析模型中的销售额、毛利等；是被观察的对象，OLAP术语称之为“**度量数据**”。
      - **角度特征**——比如销售分析中的时间周期、产品类型、销售模式、销售区域等。为观察视角，OLAP术语称之为“**维数据**”。

- OLAP应用和工具
  - OLAP应用分析要点
    - 2．对数据进行钻取，以获得更为精确的信息
      - 在现有数据基础上，将数据进一步细化
      - 比如，在销售分析中——以产品类型和销售地区为维、以销售额为度量进行分析
      - 希望进一步观察某类产品的不同销售模式在各个销售地区的表现，可以在产品大类这个数据维下面，再加上一个销售模式维，从而获得相应的信息。

- OLAP应用和工具
  - OLAP应用分析要点
    - 3．创建数据CUBE
    - 一个对于1000万条记录的分析模型，如果一次提取4个维度进行组合分析，那么实际的运算次数将达到4的1000次方的数量：这样的运算量将导致数十分钟乃至更长的等待时间。
    - 如果用户对维组合次序进行调整，或者增加减少某些维度的话，又将是一个重新的计算过程。
    - 而如果不能解决OLAP运算效率问题的话，OLAP将是一个毫无实用价值的概念。

- OLAP应用和工具
  - OLAP应用分析要点
    - 3．创建数据CUBE
    - 作为一个成熟产品是如何解决这个问题？
    - OLAP中一个非常重要的技术——**数据CUBE预运算**。
    - 一个OLAP模型中，度量数据和维数据我们应该事先确定，一旦两者确定下来，就可以对数据进行预先的处理，在正式发布之前，将数据根据维度进行最大限度的聚类运算，运算中会考虑到各种维组合情况，**运算结果将生成一个数据CUBE，并保存在服务器上。**
    - 当最终用户在调阅这个**分析模型**的时候，就可以直接使用这个CUBE，在此基础上根据用户的维度选择和维度组合进行复运算，从而达到实时响应的这么一个效果。

- OLAP应用和工具
  - OLAP与报表的区别
    - **报表是指定数据的固定形态展现。**
    - 报表的主要属性特征表现为：**数据结构**固定、**数据表现样式**固定、**数据提取范围**可以由浏览者临时通过报表参数进行控制，也可以预先固定（比如生产日报，既可以固定体现当天的生产数据，也可以由浏览者确定是要调阅某一指定日期的生产数据）。

- OLAP应用和工具
  - OLAP与报表的区别
    - OLAP是一种**分析模型**，它更为关心的是如何帮助浏览者对**数据的内在规律**进行分析，找出数据表象下的内在因素。
    - OLAP更加侧重于维度的任意灵活组合，以及大数据量下的运算效率。
    - 在应用中，**OLAP不是一种习惯性的数据观察，**使用OLAP需要首先确定一**个分析目标**，然后才是根据目标，对OLAP进行进一步操作。

上海财经大学

SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

THANK YOU