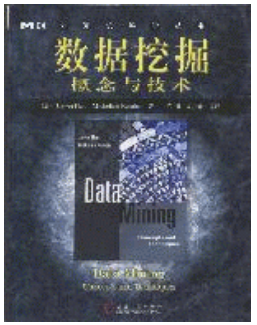
数 据 挖 掘与知识发现

主讲教师：王玲

教科书和参考书



中文版

英文影印版

- 教科书
  - 数据挖掘：概念与技术，Jiawei Han和Micheline Kamber 著，机械工业出版社 (2001)
- 参考书
  - 数据挖掘原理, David Hand, Heikki Mannila和Padhraic Smyth著，机械工业出版社 (2003)

  - **Data Mining: Practical Learning Tools and Techniques with Java Implementations 作者：（新西兰）Ian H.Witten，Eide Frank**
    - 这本书是结合开源数据挖掘工具**weka**编写，用**java**语言实现，书中描述了怎么去开发，怎么**weka**的基础上进行二次开发。
  - 机器学习**(**英文版**)** 作者**: Tom M.Mitchell**

数据挖掘的发展动力

---需要是发明之母

- 数据爆炸问题
  - 自动数据收集工具和成熟的数据库技术使得大量的数据被收集，存储在数据库、数据仓库或其他信息库中以待分析。
- 我们拥有丰富的数据，但却缺乏有用的信息
- 解决方法：数据仓库技术和数据挖掘技术
  - 数据仓库(Data Warehouse)和在线分析处理(OLAP)
  - 数据挖掘：在大量的数据中挖掘感兴趣的知识（规则，规律，模式，约束）

数据库技术的演化 (1)

- 1960s和以前:
  - 文件系统
- 1970s:
  - 层次数据库和网状数据库
- 1980s早期:
  - 关系数据模型, 关系数据库管理系统(RDBMS)的实现

数据库技术的演化 (2)

- 1980s晚期:
  - 各种高级数据库系统(扩展的关系数据库,面向对象数据库等等.)
  - 面向应用的数据库系统 (空间数据库，时序数据库，多媒体数据库等等）
- 1990s:
  - 数据挖掘, 数据仓库, 多媒体数据库和网络数据库
- 2000s
  - 流数据管理和挖掘
  - 基于各种应用的数据挖掘
  - XML数据库和整合的信息系统

什么是数据挖掘？

- 数据挖掘 (从数据中发现知识)
  - 从**大量的数据**中挖掘哪些令人感兴趣的、有用的、隐含的、先前未知的和可能有用的模式或知识
  - 挖掘的不仅仅是数据（所以“数据挖掘”并非一个精确的用词）
- 数据挖掘的替换词
  - 数据库中的知识挖掘（KDD）
  - 知识提炼、
  - 数据/模式分析
  - 数据考古
  - 数据捕捞、信息收获等等。

并非所有的东西都是数据挖掘

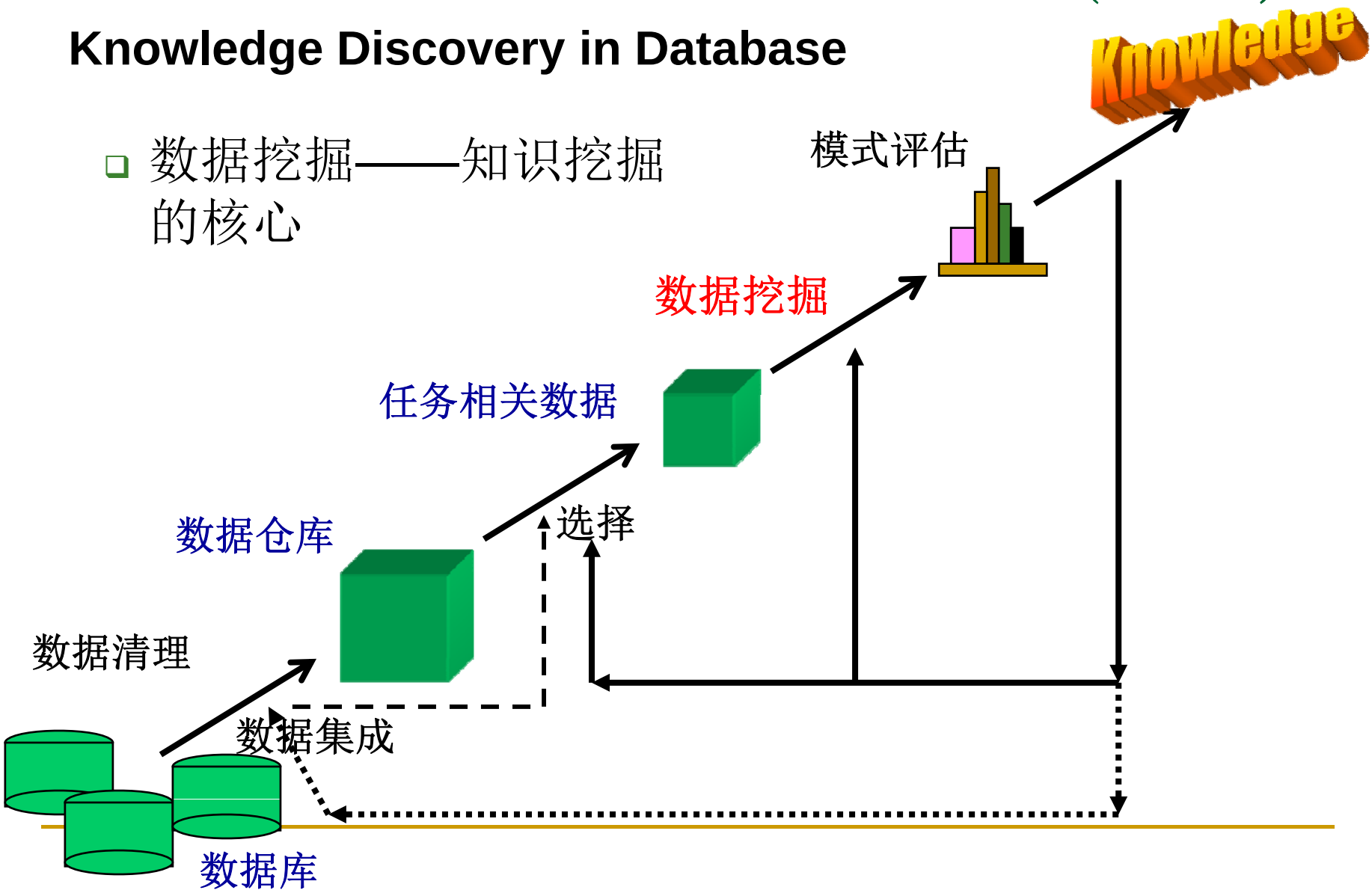
- 基于数据仓库的OLAP系统
  - OLAP系统专注于数据的汇总，而数据挖掘系统可以对数据进行多种复杂的处理。
- 机器学习系统，数据统计分析系统
  - 这些系统所处理的数据容量往往很有限。
- 信息系统
  - 专注于数据的查询处理。

- 相比于上述系统，数据挖掘系统关注更广的范围，是一个多学科的融合

数据挖掘: 数据库中的知识挖掘(KDD)

Knowledge Discovery in Database

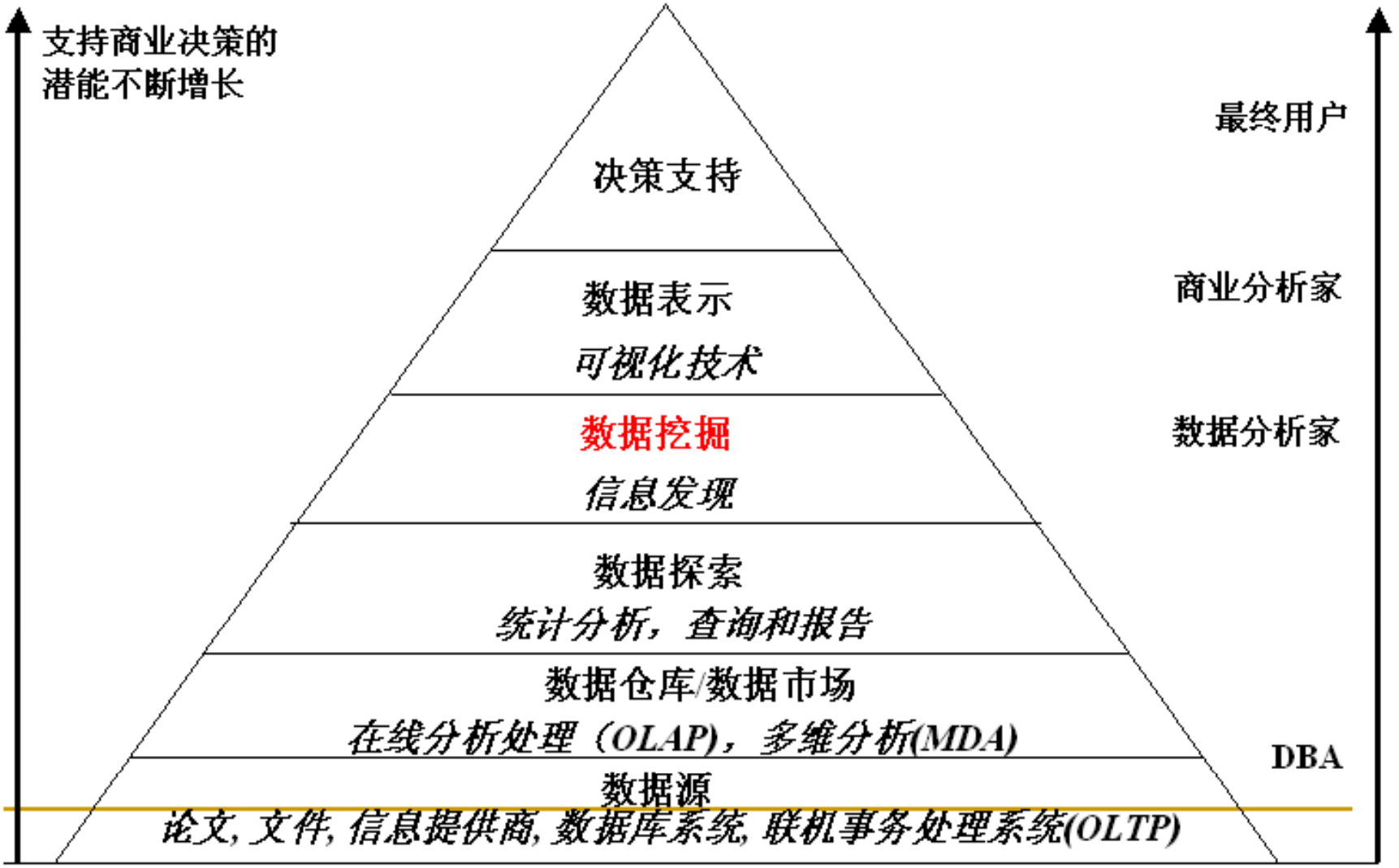
- 数据挖掘——知识挖掘的核心



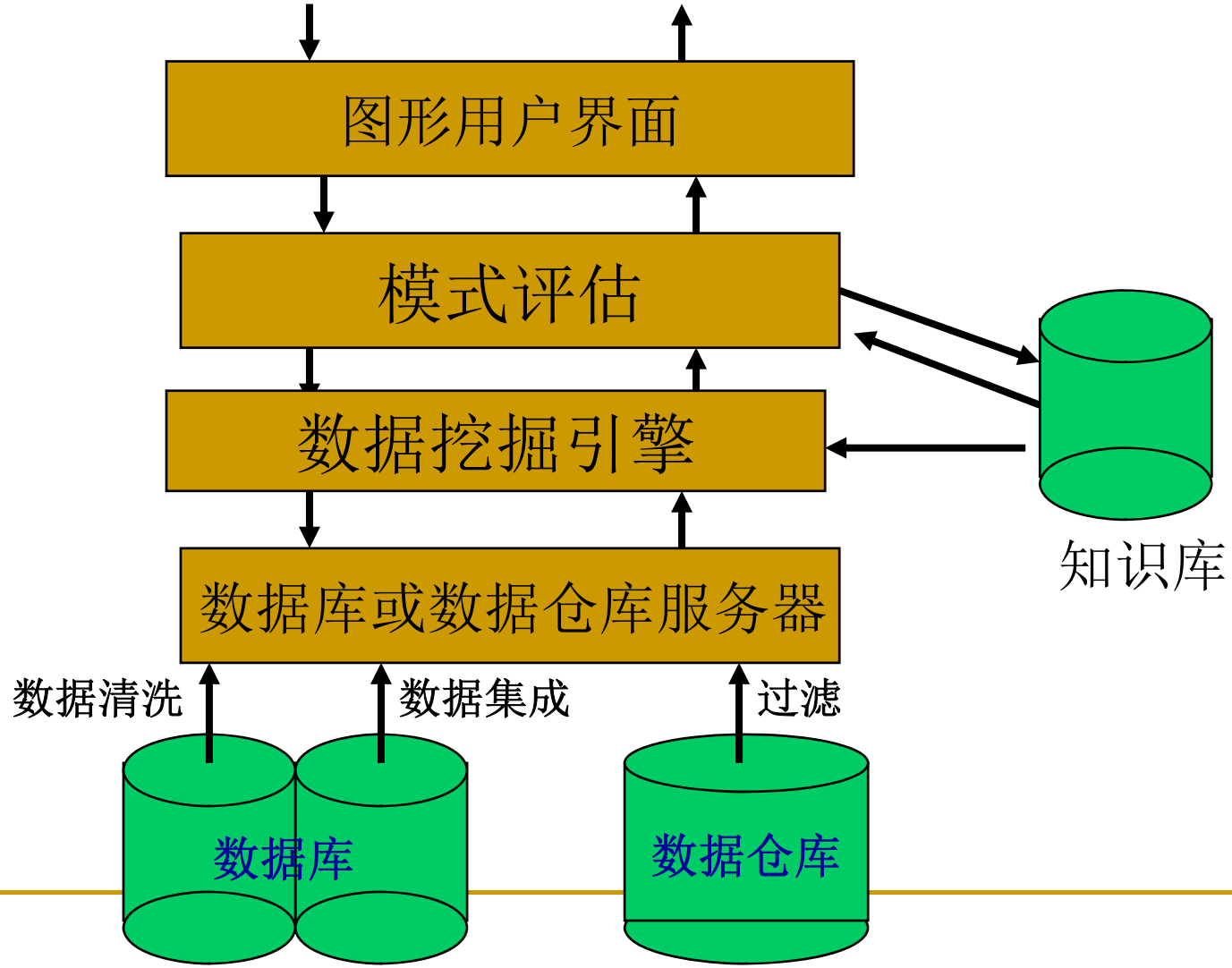
知识挖掘的步骤

- 了解应用领域
  - 了解相关的知识和应用的目标
- 创建目标数据集: 选择数据
- 数据清理和预处理: (这个可能要占全过程60％的工作量)
- 数据缩减和变换
  - 找到有用的特征，维数缩减/变量缩减，不变量的表示。
- 选择数据挖掘的功能
  - 数据总结, 分类模型数据挖掘, 回归分析, 关联规则挖掘, 聚类分析等.
- 选择挖掘算法
- 数据挖掘: 寻找感兴趣的模式
- 模式评估和知识表示
  - 可视化，转换，消除冗余模式等等
- 运用发现的知识

数据挖掘和商业智能


支持商业决策的
潜能不断增长
最终用户
决策支持
商业分析家
数据表示
可视化技术
数据分析家
数据挖掘
信息发现
数据探索
统计分析，查询和报告
数据仓库/数据市场
在线分析处理（OLAP），多维分析(MDA)
DBA
数据源
论文, 文件, 信息提供商, 数据库系统, 联机事务处理系统(OLTP)

典型数据挖掘系统的体系结构



在何种数据上进行数据挖掘

- 关系数据库
- 数据仓库
- 事务数据库
- 高级数据库系统和信息库
  - 空间数据库
  - 时间数据库和时间序列数据库
  - 流数据
  - 多媒体数据库
  - 面向对象数据库和对象-关系数据库
  - 异种数据库和历史(legacy)数据库
  - 文本数据库和万维网(WWW)

空间数据库

- 空间数据库是指在关系型数据库（DBMS-Database Management System）内部对地理信息进行物理存储。空间数据库中存储的海量数据包括对象的空间拓扑特征、非空间属性特征以及对象在时间上的状态变化。
- 常见的空间数据库数据类型
  - 地理信息系统(GIS)
  - 遥感图像数据
  - 医学图像数据
- 数据挖掘技术的应用：通过空间分类和空间趋势分析，引入机器学习算法，对有用模式进行智能检索

时间数据库和时序数据库

- 时间数据库和时序数据库都存放与时间有关的数据。时间数据库通常存放包含时间相关属性的数据。时序数据库存放随时间变化的值序列。
- 对时间数据库和时序数据库的数据挖掘,可以通过研究事物发生发展的过程，有助于揭示事物发展的本质规律，可以发现数据对象的演变特征或对象变化趋势。

流数据

- 与传统的数据库技术中的静态数据不同，流数据是连续的、有序的、变化的、快速的、大量的数据输入的数据。
- 主要应用场合
  - 网络监控
  - 网页点击流
  - 股票市场
  - 流媒体…等等
- 与传统数据库技术相比，流数据在存储、查询、访问、实时性的要求等方面都有很大区别。

多媒体数据库

- 多媒体数据库实现用计算机管理庞大复杂的多媒体数据，主要包括图形(graphics)、图象(image)、声音(audio)、视频(video)等等，现代数据库技术一般将这些多媒体数据以二进制大对象的形式进行存储。
- 对于多媒体数据库的数据挖掘，需要将存储和检索技术相结合。目前的主要方法包括构造多媒体数据立方体、多媒体数据库的多特征提取和基于相似性的模式匹配。

面向对象数据库和对象-关系数据库

- 面向对象数据库是面向对象技术和数据库技术结合的产物，该技术对数据以对象的形式进行存储，并在这个基础上实现了传统数据库的功能，包括持久性、并发控制、可恢复性、一致性和查询数据库的能力等。
- 对象－关系数据库基于对象－关系模型构造，该模型通过处理复杂对象的丰富数据类型和对象定位等功能，扩充关系模型。
- 面向对象数据库和对象－关系数据库中的数据挖掘会涉及一些新的技术，比如处理复杂对象结构、复杂数据类型、类和子类层次结构、构造继承以及方法和过程等等。

异构数据库和历史(legacy)数据库

- 历史数据库是一系列的异构数据库系统的集合，包括不同种类的数据库系统，像关系数据库、网络数据库、文件系统等等。
- 有效利用历史数据库的关键在于实现不同数据库之间的数据信息资源、硬件设备资源和人力资源的合并和共享。
- 对于异构数据库系统，实现数据共享应当达到两点：一是实现数据库转换；二是实现数据的透明访问。
  - WEB SERVICE技术的出现有利于历史数据库数据的重新利用。

文本数据库和万维网(WWW)

- 文本数据库存储的是对对象的文字性描述。
- 文本数据库的分类
  - 无结构类型（大部分的文本资料和网页）
  - 半结构类型（XML数据）
  - 结构类型（图书馆数据）
- 万维网(WWW)可以被看成最大的文本数据库
- 数据挖掘内容
  - 内容检索
  - WEB访问模式检索

数据挖掘的主要功能

——可以挖掘哪些模式？

- 一般功能
  - 描述性的数据挖掘
  - 预测性的数据挖掘
- 通常，用户并不知道在数据中能挖掘出什么东西，对此我们会在数据挖掘中应用一些常用的数据挖掘功能，挖掘出一些常用的模式，包括：
  - 概念/类描述: 特性化和区分
  - 关联分析
  - 分类和预测
  - 聚类分析
  - 孤立点分析
  - 趋势和演变分析

概念/类描述: 特性化和区分

- 概念描述：为数据的特征化和比较产生描述（当所描述的概念所指的是一类对象时，也称为类描述）
  - 特征化：提供给定数据集的简洁汇总。
    - 例：对AllElectronic公司的“大客户”（年消费额$1000以上）的特征化描述：40－50岁，有固定职业，信誉良好，等等
  - 区分：提供两个或多个数据集的比较描述。
    - 例：

Status	Birth_country	Age_range	Gpa	Count
Graduate	Canada	25-30	Good	90
Undergraduate	Canada	25-30	Good	210

关联分析

- 关联规则挖掘：
  - 从事务数据库，关系数据库和其他信息存储中的大量数据的项集之间发现有趣的、频繁出现的模式、关联和相关性。
  - 广泛的用于购物篮或事务数据分析。
- 例：

$$age(X,"30...39") \wedge income(X,"42k...48k") \Rightarrow buys(X,"computer")$$

$$[\sup port = 20\%, confidence = 70\%]$$

分类和预测

- 根据训练数据集和类标号属性，构建模型来分类现有数据，并用来分类新数据（分类），用来预测类型标志未知的对象类（预测）。
  - 比如：按气候将国家分类，按汽油消耗定额将汽车分类
  - 导出模型的表示: 决策树、分类规则、神经网络
  - 可以用来预报某些未知的或丢失的数字值
- 例：
  - IF age = “<=30” AND student = “no” THEN buys_computer = “no”
  - IF age = “<=30” AND student = “yes” THEN buys_computer = “yes”
  - IF age = “31…40” THEN buys_computer = “yes”
  - IF age = “>40” AND credit_rating = “excellent” THEN buys_computer = “yes”
  - IF age = “>40” AND credit_rating = “fair” THEN buys_computer = “no”

聚类分析

- 聚类分析：
  - 将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程。
  - 最大化类内的相似性和最小化类间的相似性
- 例：对WEB日志的数据进行聚类，以发现相同的用户访问模式

孤立点分析

- 孤立点分析
  - 孤立点:一些与数据的一般行为或模型不一致的孤立数据
  - 通常孤立点被作为“噪音”或异常被丢弃，但在欺骗检测中却可以通过对罕见事件进行孤立点分析而得到结论。
- 应用
  - 信用卡欺诈检测
  - 移动电话欺诈检测
  - 客户划分
  - 医疗分析（异常）

趋势和演变分析

- 描述行为随时间变化的对象的发展规律或趋势（时序数据库）
  - 趋势和偏差：回归分析
  - 序列模式匹配：周期性分析
  - 基于类似性的分析

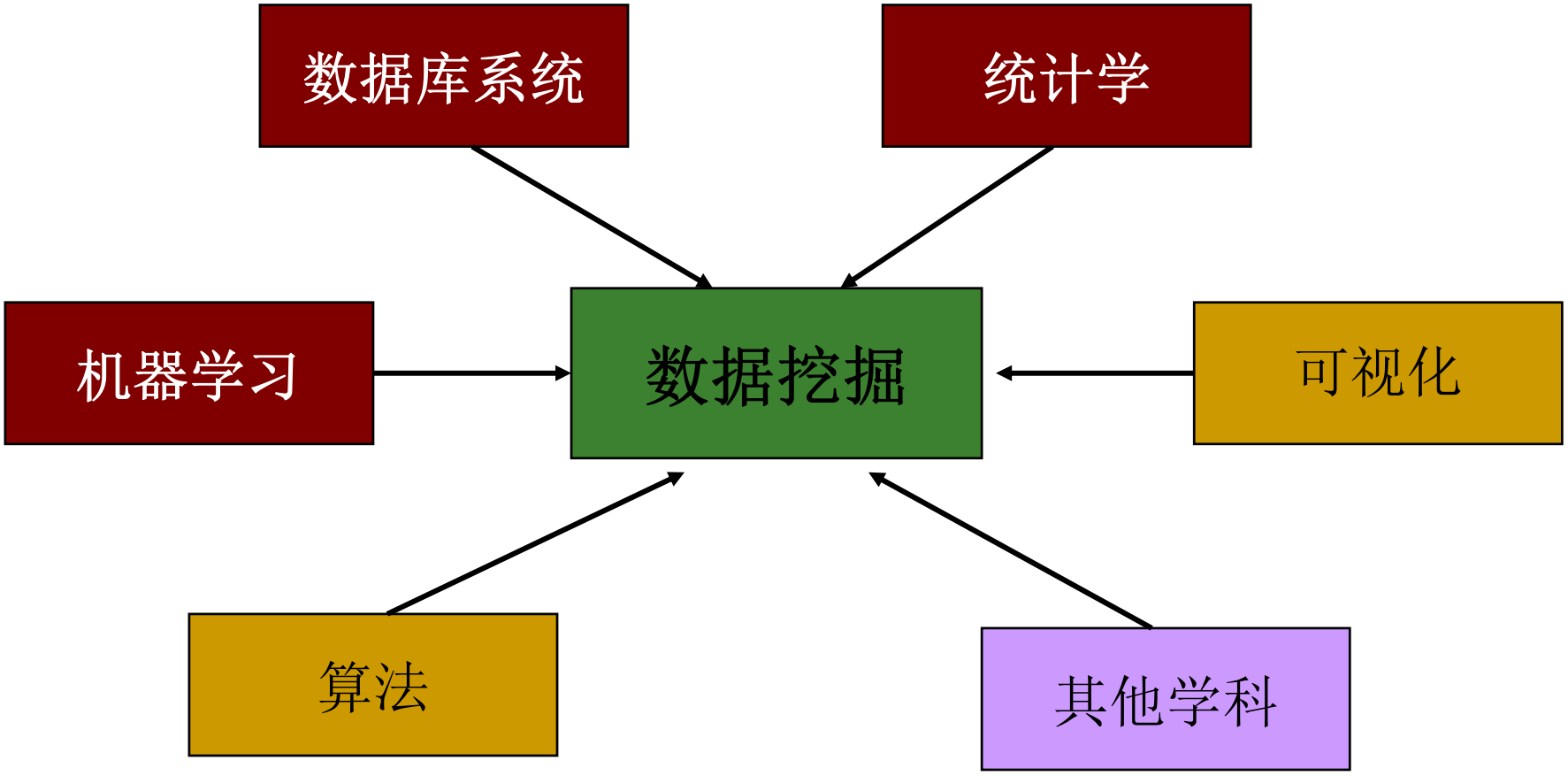
所有模式都是有趣的吗？

- 数据挖掘可能产生数以千计的模式或规则，但并不是所有的模式或规则都是令人感兴趣的。
- 模式兴趣度的度量
  - 一个模式是有趣的，如果(1) 它易于被人理解；（2）在某种程度上，对于新的或测试数据是有效的；（3）具有潜在效用；（4）新颖的；（5）符合用户确信的某种假设
- 模式兴趣度的客观和主观度量
  - 客观度量: 基于所发现模式的结构和关于它们的统计， 比如：支持度、置信度等等
  - 主观度量: 基于用户对数据的判断。比如：出乎意料的、新颖的、可行动的等等

能够产生所有有趣模式并且仅产生有趣模式吗？

- 找出所有有趣的模式: 数据挖掘算法的完全性问题
  - 数据挖掘系统能够产生所有有趣的模式吗?
  - 试探搜索 vs.穷举搜索
  - 关联 vs. 分类 vs. 聚类
- 只搜索有趣的模式: 数据挖掘算法的最优化问题
  - 数据挖掘系统可以仅仅发现有趣的模式吗？
  - 方法
    - 首先生成所有模式然后过滤那些无趣的.
    - 仅仅生成有趣的模式—挖掘查询优化

数据挖掘:多个学科的融合



数据挖掘系统的分类 (1)

- 数据挖掘的多学科融合的特性，决定了数据挖掘的研究将产生种类繁多的数据挖掘系统。
- 根据所挖掘的数据库分类
  - 关系数据库，事务数据库，流式数据，面向对象数据库，对象关系数据库，数据仓库，空间数据库，时序数据库，文本数据库，多媒体数据库，异构数据库，历史数据库，WWW

数据挖掘系统的分类 (2)

- 根据挖掘的知识类型
  - 特征分析, 区分, 关联分析, 分类聚类, 孤立点分析/演变分析, 偏差分析等等.
  - 多种方法的集成和多层机挖掘
- 根据挖掘所用的技术
  - 面向数据库的挖掘、数据仓库 、OLAP(在线分析处理)、机器学习、统计学、可视化等等.
- 根据挖掘所用的应用
  - 金融,电信,银行, 欺诈分析, DNA分析,股票市场, Web挖掘等等.

数据挖掘应用——市场分析和管理

- 数据从那里来?
  - 信用卡交易, 会员卡, 商家的优惠卷, 消费者投诉电话, 公众生活方式研究
- 目标市场
  - 构建一系列的“客户群模型”，这些顾客具有相同特征: 兴趣爱好, 收入水平, 消费习惯,等等
  - 确定顾客的购买模式
- 交叉市场分析
  - 货物销售之间的相互联系和相关性，以及基于这种联系上的预测

数据挖掘应用——欺诈行为检测和异常模式的发现

- 方法: 对欺骗行为进行聚类和建模，并进行孤立点分析
- 应用: 卫生保健、零售业、信用卡服务、电信等
  - 汽车保险: 相撞事件的分析
  - 洗钱: 发现可疑的货币交易行为
  - 医疗保险
    - 职业病人, 医生以及相关数据分析
    - 不必要的或相关的测试
  - 电信: 电话呼叫欺骗行为
    - 电话呼叫模型: 呼叫目的地，持续时间，日或周呼叫次数. 分析该模型发现与期待标准的偏差
  - 零售产业
    - 分析师估计有38%的零售额下降是由于雇员的不诚实行为造成的
  - 反恐怖主义

其他应用

- 体育竞赛
  - 美国NBA的29个球队中，有25个球队使用了IBM分析机构的数据挖掘工具，通过分析每个对手的数据（盖帽、助攻、犯规等数据）来获得比赛时的对抗优势。
- 天文学
  - JPL实验室和Palomar天文台就曾经在数据挖掘工具的帮助下发现了22颗新的恒星
- 网上冲浪
  - 通过将数据挖掘算法应用于网络访问日志，从有市场相关的网页中发现消费者的偏爱和行为，分析网络行销的有效性，改善网络站点组织。这就是新兴的WEB挖掘研究

小结

- **数据挖掘**是从大量数据中发现有趣模式，这些数据可以存放在数据库、数据仓库或其它信息存储中。这是一个年青的跨学科领域，源于诸如数据库系统、数据仓库、统计、机器学习、数据可视频化、信息提取和高性能计算。
- **知识发现**过程包括数据清理、数据集成、数据变换、数据挖掘、模式评估和知识表示。
- 数据模式可以从不同类型的**数据库**挖掘；如关系数据库，数据仓库，事务的、对象-关系的和面向对象的数据库。有趣的数据模式也可以从其它类型的**信息存储**中提取，包括空间的、时间相关的、文本的、多媒体的和遗产数据库，以及万维网。
- **数据仓库**是一种数据的长期存储，这些数据来自多数据源，是有组织的，以便支持管理决策。这些数据在一种一致的模式下存放，并且通常是汇总的。数据仓库提供一些数据分析能力，称作**OLAP（联机分析处理）**。
- **数据挖掘功能**包括发现概念/类描述、关联、分类、预测、聚类、趋势分析、偏差分析和类似性分析。特征和区分是数据汇总的形式。
- 模式提供**知识**，如果它易于被人理解、在某种程度上对于测试数据是有效的、潜在有用的、新颖的，或者它验证了用户关注的某种预感。**模式兴趣度**度量，无论是客观的还是主观的，都可以用来指导发现过程。
- **数据挖掘系统**可以根据所挖掘的数据库类型、所挖掘的知识类型、或所使用的技术加以分类。

作业

- 以某一种高级数据库系统 (面向对象数据库、空间数据库、文本数据库、多媒体数据库、万维网等)为背景,结合具体的应用领域,查找相关数据挖掘方法的研究现状和发展趋势进行归纳,写一篇综述性的报告,字数在3000字左右。