第一章　引 论

张静
（华东理工大学计算机系）
（Jingzhang@ecust.edu.cn）

主要内容

- 为什么要进行数据挖掘?
- 什么是数据挖掘?
- 对何种数据进行数据挖掘?
- 可以挖掘什么类型的模式?
- 所有模式都是有趣的吗?
- 数据挖掘系统的分类
- 数据挖掘的主要问题

需要是发明之母（Necessity is the mother of invention）

- 数据爆炸问题（from terabytes to petabytes）
  - 自动数据收集工具和成熟的数据库技术导致了海量数据被存放在数据库、数据仓库和其他信息库中等待分析。
- 我们数据丰富，但信息贫乏！（We are drowning in data, but starving for knowledge!）
- 解决方案：数据仓库和数据挖掘
  - 联机事务处理（**Online transaction processing, OLTP**）
  - 数据仓库和联机分析处理（**Online Analytical Processing, OLAP**）
  - 从大型数据库中挖掘有趣的知识（规则、模式等）

Evolution of Sciences

Before 1600, **empirical science**

- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

数据库技术的变革

- 20世纪60年代
  - 数据搜集，数据库的创建，IMS以及网络DBMS
- 20世纪70年代
  - 关系数据模型，关系型DBMS的实现
- 20世纪80年代
  - RDBMS, 高级数据模型 (扩充关系, 面向对象, 对象-关系,演绎, 等等)
  - 面向应用的DBMS (空间的, 科学的，基于知识的等等)
- 20世纪90年
  - 数据挖掘，数据仓库，多媒体数据库，Web数据库
- 2000至今
  - 流数据管理和挖掘
  - 数据挖掘和应用
  - Web技术和全球信息系统

主要内容

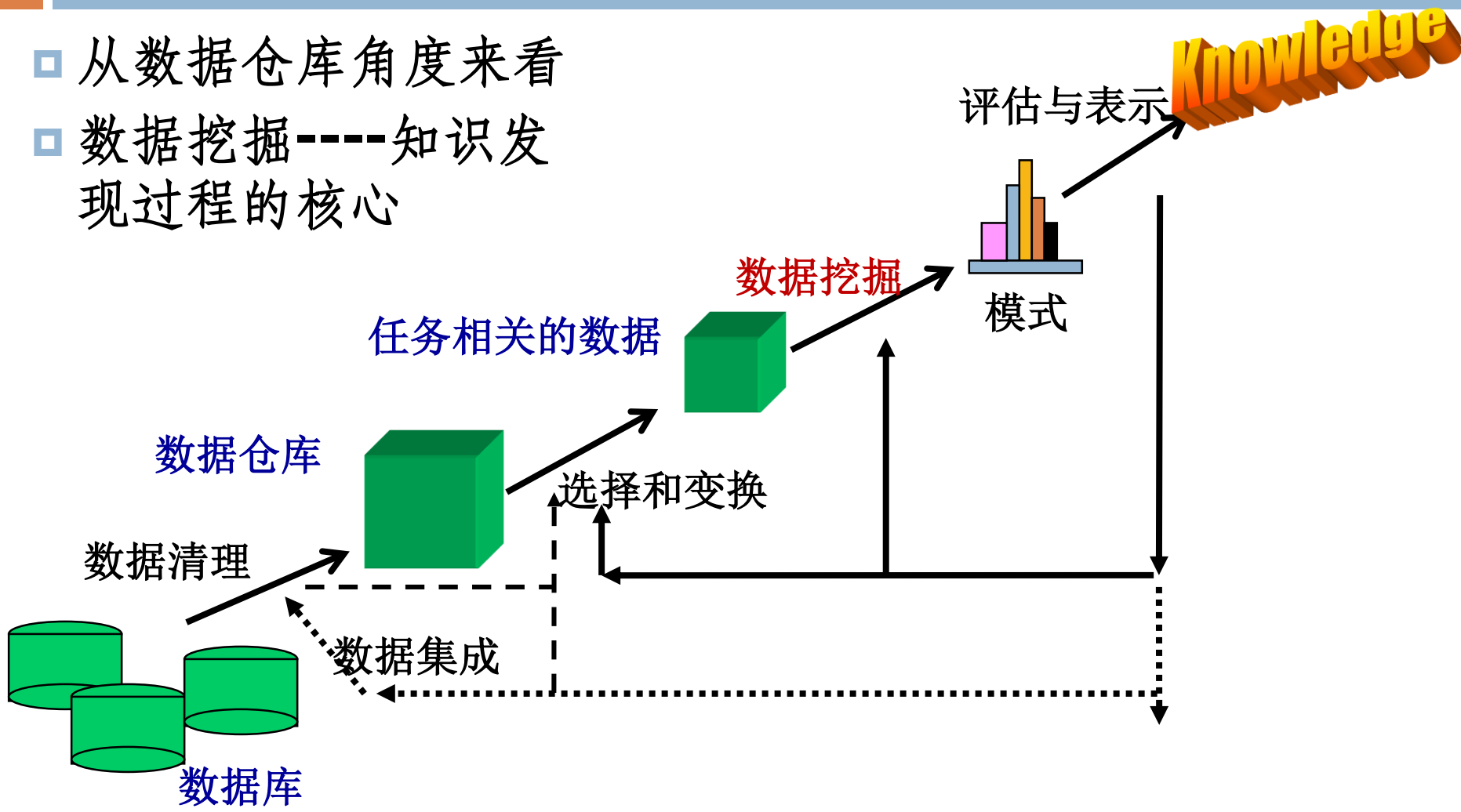
- 为什么要进行数据挖掘?
- 什么是数据挖掘?
- 对何种数据进行数据挖掘?
- 可以挖掘什么类型的模式?
- 所有模式都是有趣的吗?
- 数据挖掘系统的分类
- 数据挖掘的主要问题

什么是数据挖掘?

- 数据挖掘(从大量数据中提取或“挖掘” 知识的过程)
  - 从海量数据中抽取出有用的模式或者知识，这些模式或者知识应该是：非常识性的、隐藏的、当前未知的以及潜在有益的。
- 其他相近名字
  - 数据库中知识发现（**Knowledge discovery (mining) in databases (KDD)**）
  - 知识提取（**knowledge extraction**）
  - 数据/模式分析（**data/pattern analysis**）
  - 数据考古（**data archeology**）
  - 数据捕捞（ **data dredging**）
  - ……

数据挖掘：从数据库中挖掘知识之路

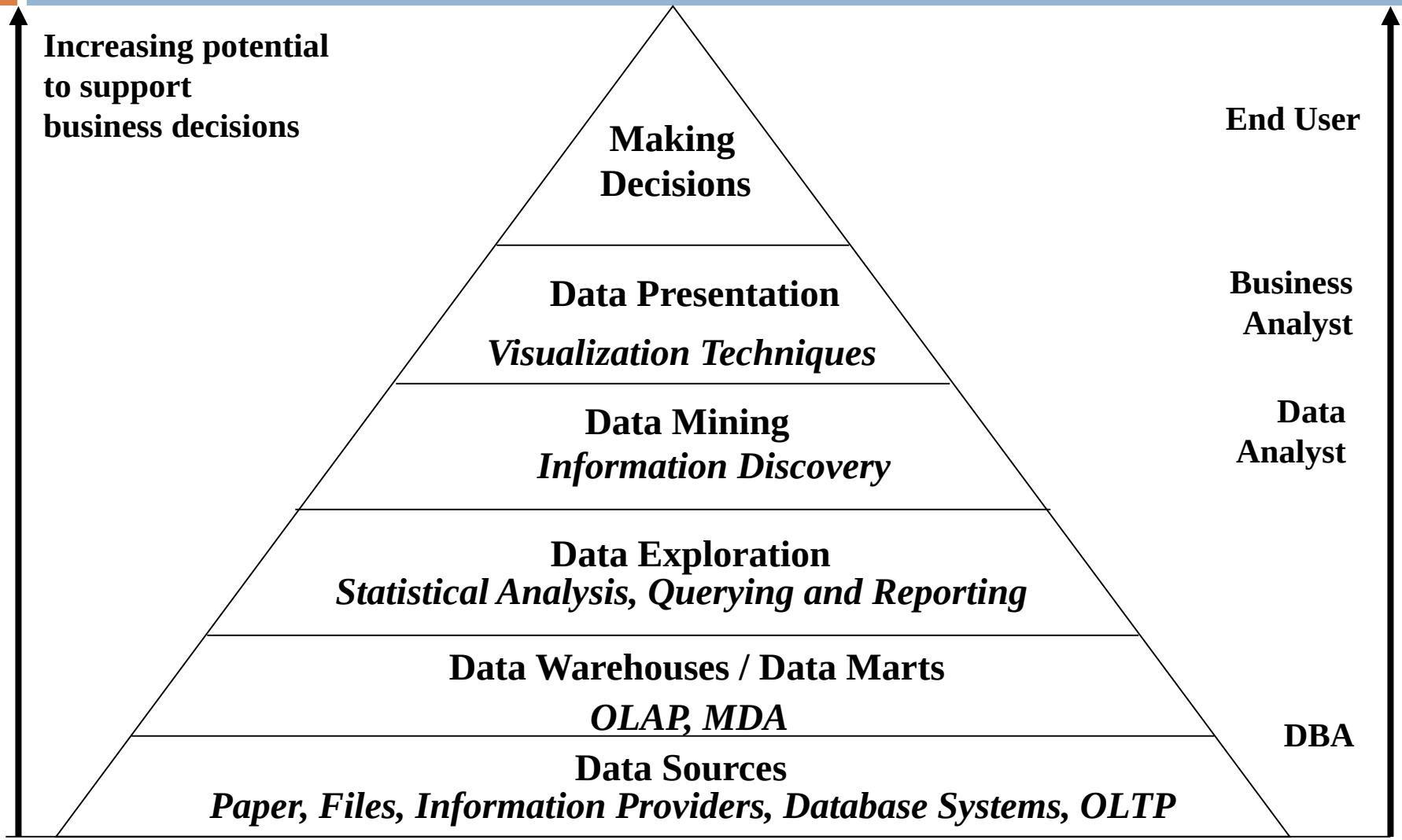
- 从数据仓库角度来看
- 数据挖掘----知识发现过程的核心



KDD的步骤

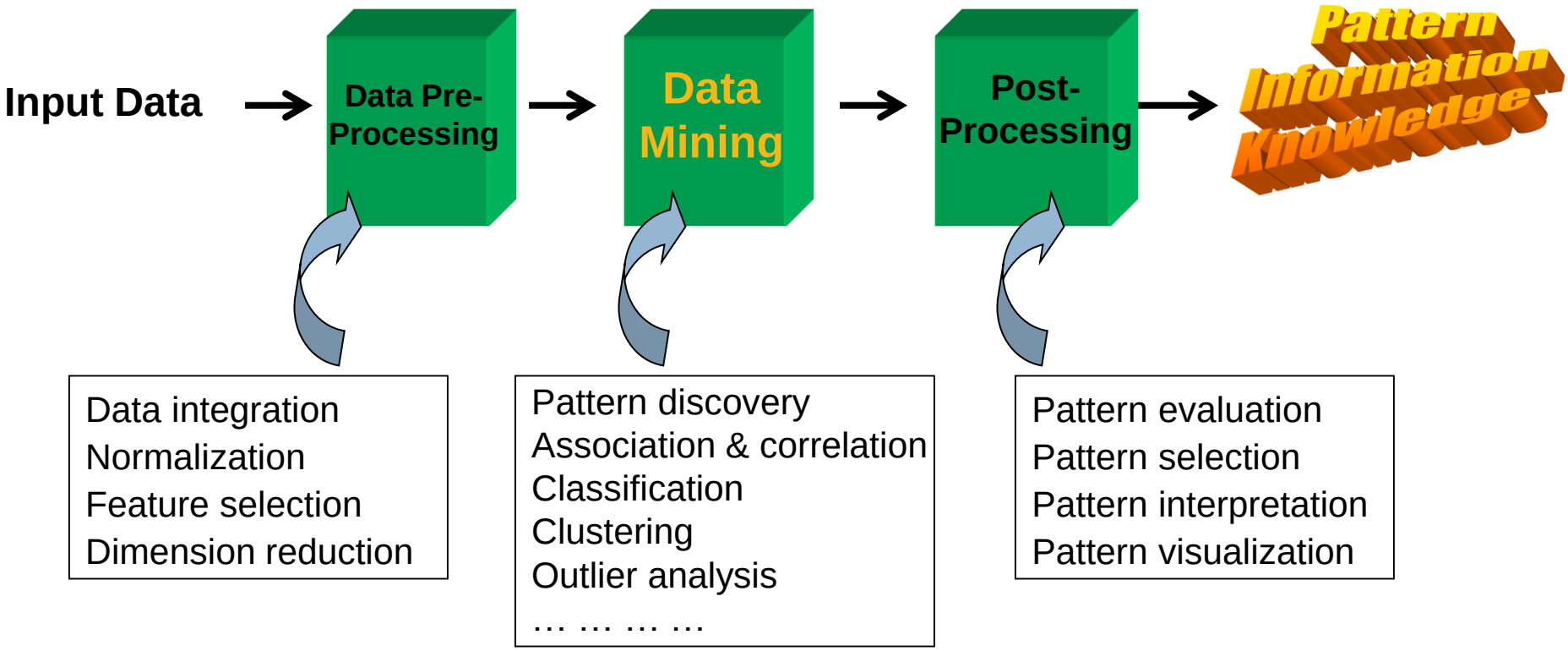
- 数据清理
  - 消除噪声或不一致数据
- 数据集成
  - 多种数据源可以组合在一起
- 数据选择
  - 从数据库中检索与分析任务相关的数据
- 数据变换
  - 数据变换或统一成适合挖掘的形式，如通过汇总或者聚集操作
- 数据挖掘
  - 基本步骤，使用智能方法提取数据模式
- 模式评估
  - 根据某种兴趣度度量，识别表示知识的真正有趣的模式
- 知识表示
  - 使用可视化和知识表示技术，向用户提供挖掘的知识

数据挖掘和商务智能


Increasing potential
to support
business decisions
Making
Decisions
End User
Data Presentation
Visualization Techniques
Business
Analyst
Data Mining
Information Discovery
Data
Analyst
Data Exploration
Statistical Analysis, Querying and Reporting
Data Warehouses / Data Marts
OLAP, MDA
DBA
Data Sources
Paper, Files, Information Providers, Database Systems, OLTP

数据库中知识发现（KDD）过程

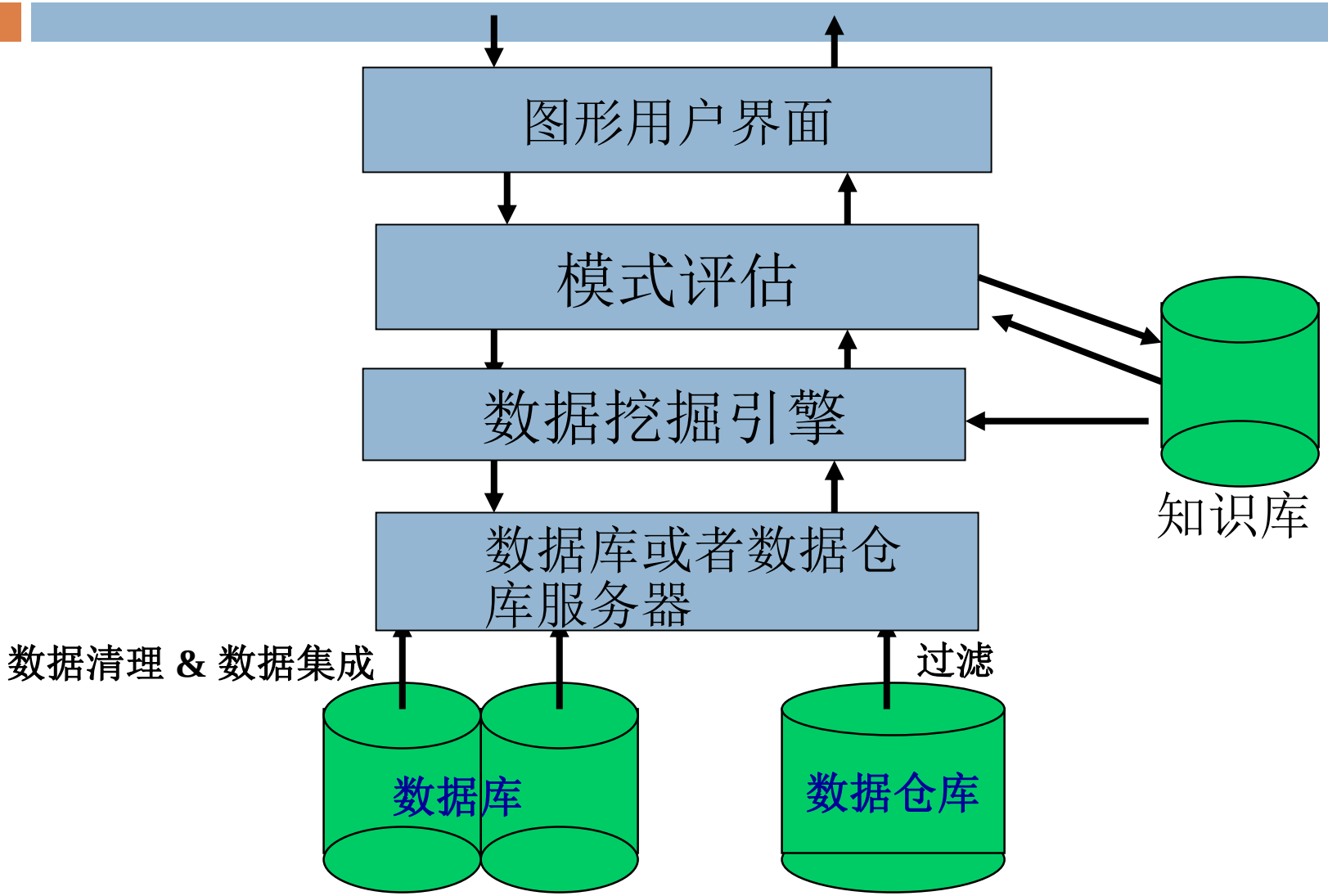
- 从机器学习和统计学的观点来看数据库中只是发现过程如下图所示



Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning
- Preprocessing of the data (including feature extraction and dimension reduction)
- Classification or/and clustering processes
- Post-processing for presentation

架构: 典型的数据挖掘系统



为什么要数据挖掘?—潜在应用

- 数据分析和决策支持
  - 市场分析和管理
    - 目标市场, 客户关系管理 (CRM), 超市购物篮分析, 市场划分
  - 风险分析和管理
    - 预测, 客户保持, 保险分析, 质量控制, 竞争分析
  - 欺诈检测和异常模式检测（outliers）
- 其他应用
  - 文本挖掘 (新闻组, 电子邮件, 文档) 和Web挖掘
  - 流数据挖掘
  - DNA和生物数据分析

市场分析和管理

- 数据从哪里来?
  - 信用卡事务日志, 贵宾卡, 打折激励, 客户投诉电话.....
- 目标市场
  - 发现具有相同“模式”的客户，他们有着相同的特性，如兴趣、收入、消费习惯等。
  - 发现客户购买模式的变化
- 交叉市场分析
  - 关联规则/所销售产品的相关性, & 基于关联规则的预测
- 客户轮廓
  - 何种类型客户会买何种产品？ (聚类或者分类)
- 客户需求分析
  - 确定不同客户所需要的最佳产品
  - 预测什么因素将会吸引新的客户们

公司分析&风险管理

- 金融规划和资产评估
  - 现金流分析和预测
  - 资产评估
  - 交叉和时间序列分析 (财务比率分析, 趋势分析等)
- 资源计划
  - 总结和比较资源和开销
- 竞争
  - 监测竞争对手和市场导向
  - 客户分组和各组客户的定价过程
  - 在一个激烈竞争市场上设置定价策略

欺诈检测和挖掘异常模式

- 方法: 聚类 & 针对欺诈的模型构造, 离群点分析
- 应用领域: 健康护理, 零售, 信用卡服务, 电信.
  - 汽车保险: 汽车碰撞
  - 洗钱: 不明转帐事务
  - 医疗保险
    - 职业病人； 医生, 和证明人
    - 非必要的的拍片检测
  - 电信: 电话欺诈
    - 电话模型: 电话的呼出方, 通话时长, 一个月的第几天或者星期几.
  - 零售业
    - 分析家估计38%的零售萎缩都是由于不诚实的雇员所致。
  - 反恐
    - 异常现象分析

其他应用

- 体育
  - IBM Advanced Scout analyzed NBA game statistics (盖冒，助攻和犯规)能够为一些队伍（例如纽约尼克斯队和迈阿密热队）获得竞争优势
- 航天
  - 美国喷气推进实验室（JPL）和巴洛马山天文台通过数据挖掘技术找到了22个恒星
- 因特网上网助手
  - 用户访问因特网会产生Web访问日志。IBM Surf-Aid利用数据挖掘算法结合这些日志能够发现客户的喜好和感兴趣的页面，从而分析网上浏览的效率，重新组织Web页面等

数据挖掘系统&工具

请参考 www.kdnuggets.com

- Oracle: Darwin
- IBM: Intelligence Miner
- SAS: Enterprise Miner
- Business Objects
- SPSS: Clementine
- Xchange: e-CRM
- Microsoft: SQL Server 2000
- … …

主要内容

- 为什么要进行数据挖掘?
- 什么是数据挖掘?
- 对何种数据进行数据挖掘?
- 可以挖掘什么类型的模式?
- 所有模式都是有趣的吗?
- 数据挖掘系统的分类
- 数据挖掘的主要问题

数据挖掘：基于何种数据?

- 关系数据库
- 数据仓库
- 事务数据
- 其他类型的数据
  - 空间数据（Spatial and temporal database）
  - 时间相关或序列数据（Time-series database ）
  - 流数据（Stream data）
  - 超文本和多媒体数据库（Multimedia database）
  - 文本数据库（Text databases）
  - 图和网状数据（ graphs, social networks and multi-linked data ）
  - 万维网（The World-Wide Web）

主要内容

- 为什么要进行数据挖掘?
- 什么是数据挖掘?
- 对何种数据进行数据挖掘?
- 可以挖掘什么类型的模式?
- 所有模式都是有趣的吗?
- 数据挖掘系统的分类
- 数据挖掘的主要问题

数据挖掘功能

- 概念/类描述: 数据特征化和数据区分
  - 数据特征化：目标类数据的一般特性或者特性的汇总
  - 数据区分：将目标类对象的一般特性与一个或多个对比类对象的一般特性比较
  - 例如：干旱地区和潮湿地区
- 频繁模式（或频繁项）
  - 在沃尔玛超市中哪些商品经常被一起购买?
- 关联和相关性(相关性和因果关系)
  - 一个典型的关联规则
    - 切片面包 → 奶酪 [0.5%, 75%] (support, confidence)
  - 强关联项是否一定是强相关的?

数据挖掘功能

- 用于预测分析的分类与回归
  - 分类构造模型，将所有数据划分为少数种类，用于将来的预测
    - 例如，根据气候对国家进行分类，或者根据排量对小车进行分类
  - 回归预测一些未知的或者丢失的数值
    - 例如：预测股票的涨幅
  - 分类预测类别（离散的、无序的）标号，回归建立连续值函数模型。
  - 分类的表示方式：决策树，分类规则，神经网络等
  - 回归的表示方式：逻辑回归等

数据挖掘功能

- 聚类分析（Cluster analysis）
  - 聚类分析数据对象而不考虑类标号
  - 类标签是未知的：将数据分组以找到新的类型，例如，将所有房子进行聚类，从而找到分布模式
  - 目标：最大化类内的相似性，最小化类间的相似性
- 离群点分析（Outlier analysis）
  - 离群点：一个数据对象，它并不遵循这类数据的通用行为
  - 噪声或错误？不是！对于欺诈检测、少见事件分析有好处

主要内容

- 为什么要进行数据挖掘?
- 什么是数据挖掘?
- 对何种数据进行数据挖掘?
- 可以挖掘什么类型的模式?
- 所有模式都是有趣的吗?
- 数据挖掘系统的分类
- 数据挖掘的主要问题

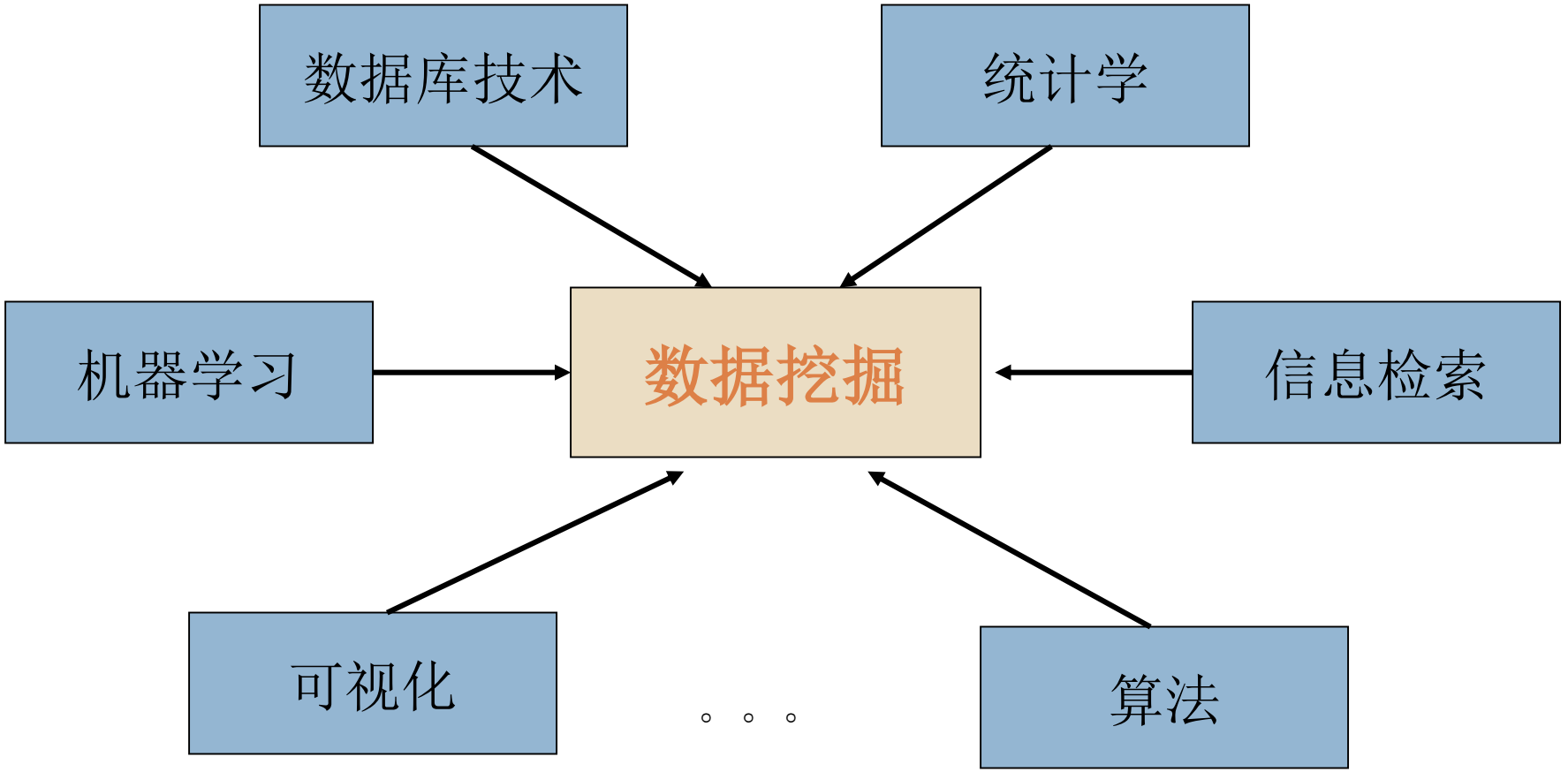
是否所有被“发现”的模式都是有趣的？（Evaluation of knowledge）

- 数据挖掘能够产生数以千计的模式，但是并非这些模式均为有趣的。
- 有趣的度量
  - 一个模式是有趣的（interesting），仅当它（1）易于被人理解；（2）在某种确信度上，对新的或者检验数据是有效的；（3）是潜在有用的；（4）是新颖的。
- 模式兴趣度的客观度量
  - 客观的（Objective）：基于一些统计结论和模式结构（statistics and structures of patterns），例如支持度（support），置信度（confidence）、准确率（precision）、覆盖率（coverage）等。
  - 主观的（Subjective）：基于用户的信念（user's belief），例如非预期性，新颖性，可行动的（actionable）等。

我们能否发现所有有趣的模式?

- 找到所有有趣的模式： 完全性（Completeness）
  - 一个数据挖掘系统能够找到所有有趣的模式吗?
  - 启发式（Heuristic） vs. 穷尽搜索（exhaustive search）
- 仅搜索有趣的模式：优化问题
  - 一个数据挖掘系统能够仅找到有趣的模式吗?
  - 方法
    - 首先生成所有的模式，然后将无趣的模式过滤掉。
    - 仅产生有趣的模式----挖掘查询优化问题

数据挖掘：多学科交叉



数据挖掘使用的主要技术

- 统计学
  - 统计学研究数据的收集、分析、解释和表示。
  - 统计学模型
- 机器学习
  - 考察计算机如何基于数据学习
  - 经典的机器学习问题
    - 监督学习：学习数据有类标记。
    - 无监督学习：学习数据无类标记。
    - 半监督学习：使用标记和未标记实例一起学习。
    - 主动学习：主动从用户获取知识来提高模型质量。

数据挖掘使用的主要技术

- 数据库系统与数据仓库
  - 数据库系统：创建、维护、使用数据库。
  - 数据仓库：集成多个数据源，用于分析决策。
- 信息检索
  - 对象
    - 文本、图像、视频、音频
    - 经典的文本信息检索模型
      - 布尔模型
      - 矢量模型
      - 概率模型

主要内容

- 为什么要进行数据挖掘?
- 什么是数据挖掘?
- 对何种数据进行数据挖掘?
- 可以挖掘什么类型的模式?
- 所有模式都是有趣的吗?
- 数据挖掘系统的分类
- 数据挖掘的主要问题

数据挖掘系统的分类

- 根据一般功能分类
  - 描述性数据挖掘
    - 刻画数据库中数据的一般特性
  - 预测性数据挖掘
    - 在当前数据上进行推断，以进行预测
- 根据不同方面分类
  - 根据挖掘的数据库类型分类
  - 根据挖掘的知识类型分类
  - 根据所用的技术分类
  - 根据应用分类

数据挖掘系统的分类

- 待挖掘的数据库（Database to be mined）
  - 关系数据库，数据仓库，事务日志，数据流，对象-关系数据库，空间数据，时间序列，文本，多媒体，异构数据，WWW
- 待挖掘的知识（Knowledge to be mined）
  - 关联规则，分类，聚类，趋势分析，离群点分析等。
- 使用的技术（Techniques utilized）
  - 数据库，数据仓库（OLAP），机器学习，统计，可视化表示等。
- 应用场景
  - 零售业，电信业，银行业，欺诈分析，生物数据挖掘，股市分析，Web挖掘等。

主要内容

- 为什么要进行数据挖掘?
- 什么是数据挖掘?
- 对何种数据进行数据挖掘?
- 可以挖掘什么类型的模式?
- 所有模式都是有趣的吗?
- 数据挖掘系统的分类
- 数据挖掘的主要问题

数据挖掘的主要问题

- 挖掘方法
  - 挖掘各种新的知识类型
  - 挖掘多维空间中的知识
  - 跨学科
  - 提升网络环境下的发现能力
  - 处理不确定性、噪声、或不完全数据
  - 模式评估和模式或约束指导的挖掘

数据挖掘的主要问题

- 用户界面
  - 交互挖掘
  - 结合背景知识
  - 特定的数据挖掘和数据挖掘语言
  - 数据挖掘结果的表示和可视化
- 有效性和可伸缩性
  - 数据挖掘算法的有效性和可伸缩性
  - 并行、分布式和增量挖掘算法

数据挖掘的主要问题

- 数据库类型的多样性
  - 处理复杂的数据类型
  - 挖掘动态的、网络的、全球的数据库
- 数据挖掘与社会
  - 数据挖掘的社会影响
  - 保护隐私的数据挖掘
  - 无形的数据挖掘

总之，数据挖掘是……

- 从海量数据中发现有益的模式
- 是数据库技术的自然进化，在很多应用中有着巨大的需求
- 一个完整的知识发现过程包含：数据清理、数据集成、数据选择、数据变换、数据挖掘、模式评估和知识表示
- 数据挖掘可以基于多种类型的数据库
- 数据挖掘功能：概念/类描述，关联规则，分类，预测，聚类，孤立点检测和趋势分析等
- 多维数据挖掘
- 数据挖掘存在很多挑战性课题

数据挖掘研究的简要历史

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

数据挖掘相关的会议和期刊

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

参考资料来源： DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

推荐的参考书目

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd ed., 2006
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005

思考题

- 数据仓库和数据库有哪些不同？有哪些相似之处？
- 区分和分类的差别是什么？特征化和聚类的差别是什么？分类和回归呢？对于每一对任务，它们有何相似之处？
- 与挖掘少量数据相比，挖掘海量数据的主要挑战是什么？
- 谈谈你对当前比较流行的大数据（big data）的认识？

大数据的特点

- Volume（大量）
  - 数据量巨大
  - 从TB级别，跃升到PB级别
- Variety（多样）
  - 数据类型繁多
  - 网络日志、视频、图片、地理位置信息等等
- Value（低价值密度）
  - 数据价值密度低
  - 以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒。
- Velocity（高速）
  - 获得数据的速度快
- Veraticy（真实性）
  - 数据的质量

结　束