数据库系统概论新技术篇

数据挖掘

李翠平

中国人民大学信息学院

概览

- 什么是数据挖掘？
- 数据挖掘的过程
- 数据挖掘在什么样的数据上进行？
- 是否所有挖掘出来的模式都是有用的？
- 数据挖掘的功能
- 数据挖掘的特点
- 数据挖掘的分类
- 数据挖掘技术构架
- 其它

什么是数据挖掘

数据挖掘是从大量数据中提取知识的过程

数据

知识

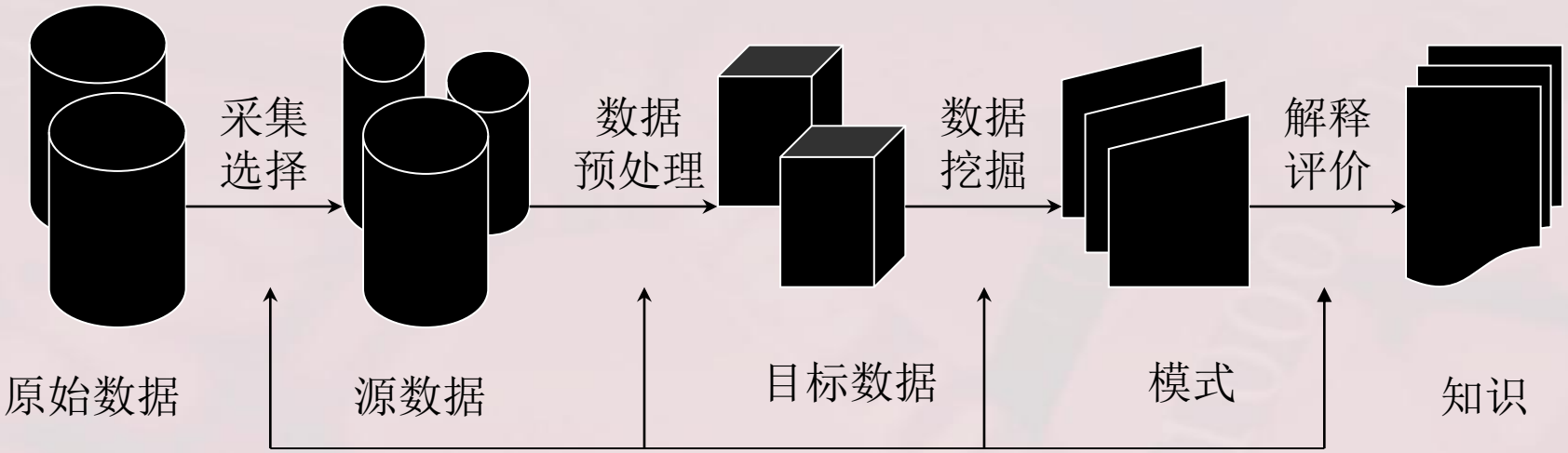
数据挖掘是分析数据的科学与艺术、是数据科学的核心部分

啤酒与尿布

有一次，美国沃尔玛公司的分店经理发现：一段时期以来，每逢周末店内啤酒和尿布的销量都会同比攀升。

分析发现：原来这些人习惯晚上边看球赛、边喝啤酒，对于要照顾的孩子，为了图省事就用一次性尿布。于是沃尔玛决定：把这两种商品集中摆在一起。

数据挖掘的过程



KDD（Knowledge Discovery from Database）

可看作数据挖掘的代名词

数据挖掘在什么样的数据上进行？

- 关系数据库数据、数据仓库数据、事务交易数据
- 高级数据库或应用所产生的数据
  - 对象-关系数据库数据、时态数据库数据、时间序列数据
  - 空间数据库数据和时空数据库数据
  - 文本数据库和多媒体数据库
  - 图数据库、网络数据
  - 数据流
  - …

是否所有挖掘出来的模式都是有用的？

- 数据挖掘可能产生成千上万的模式，并不都是有用的
  - 模式有用需要满足以下条件：容易被人所理解，在新的数据上仍然有效，是潜在的有用的新颖的，或者能够验证人们所关心的某种假设的

- 是否有用如何度量？
  - 客观评价：基于模式的统计或者结构进行分析，如支持度、可信度等
  - 主观评价：基于用户对数据的感觉进行分析，如新颖性、可执行性等

数据挖掘的功能

- 根据数据分析者的目标，可以将数据挖掘任务分为：
  - 模型挖掘（模型是对整个数据集的全局性的描述或总结）
    - 描述建模（聚类）
    - 预测建模（分类）
  - 模式挖掘（模式是局部的，它仅对一小部分数据做出描述）
    - 频繁模式（关联规则）
    - 异常模式（异常点挖掘）

数据挖掘的特点

- 第一，数据挖掘的数据源必须是真实的
  - 数据挖掘所处理的数据通常是已经存在的真实数据，如超市业务数据，而不是为了进行数据分析而专门收集的数据
  - 因此，数据收集本身不属于数据挖掘所关注的焦点，这是数据挖掘区别于大多数统计任务的特征之一

数据挖掘的特点

- 第二，数据挖掘所处理的数据必须是海量的
  - 如果数据集很小的话，采用单纯的统计分析方法就可以了
  - 但是，当数据集很大时，会面临许多新的问题，诸如，数据的有效存储、快速访问、合理表示等

数据挖掘的特点

- 第三，查询一般是决策制定者（用户）提出的随机查询
  - 查询要求灵活，往往不能形成精确的查询要求，要靠数据挖掘技术来寻找可能的查询结果

数据挖掘的特点

- 第四，挖掘出来的知识一般是不能预知的，数据挖掘发现的是潜在的、新颖的知识
  - 这些知识在特定环境下是可以接受、可以理解、可以运用的，但不是放之四海皆准的

数据挖掘的分类

- 根据挖掘的数据库类型分类
  - 数据库系统本身可以根据不同的标准分类，例如,按照数据模型或处理的数据所涉及的应用类型分类。每一类可能需要不同的数据挖掘技术。例如，根据数据模型分类，可以有关系的、面向对象的、对象-关系的、或数据仓库的数据挖掘
  - 如果根据所处理的数据的特定类型分类，有空间的、时间序列的、文本的、多媒体、或Web数据等数据挖掘

数据挖掘的分类

- 根据挖掘的知识类型分类
  - 例如特征分析、关联分析、分类分析、聚类分析、异常点分析、趋势和演化分析、偏差分析、类似性分析等
  - 此外，数据挖掘也可以根据所挖掘的知识的粒度或抽象级别进行区分，包括泛化知识（在高抽象层），原始层知识（在原始数据层），或多层知识（考虑若干抽象层）

数据挖掘的分类

- 根据所用的技术分类
  - 这些技术可以根据用户交互程度（例如，自动系统、交互探查系统、查询驱动系统）
  - 或所用的数据分析方法（例如，面向数据库或数据仓库的技术、机器学习、统计、可视化、模式识别、神经网络等等）
  - 复杂的数据挖掘通常采用多种数据挖掘技术，或采用有效的、集成的技术，以综合若干不同方法的优点

数据挖掘的分类

- 根据数据挖掘的应用领域分类
  - 例如，可能有些数据挖掘方法特别适合财政、电讯，有些数据挖掘方法特别适合DNA、股票市场等。
  - 不同的应用有适合该应用不同的数据挖掘方法。而通用的、全面的数据挖掘可能并不适合特定领域的挖掘任务。

数据挖掘技术构架

- Association rules discovery
- Sequential Pattern Discovery
- Cluster analysis
- Outlier Detection
- Classifier Building
- Data Cube/Data Warehouse Construction
- Visualization …

Applications

- Customer Relationship Management (CRM)
- Web pages Searches and Analysis
- Network Security
- Geographical Data Analysis
- Genomic Database …

Techniques

- Database Technology:
  - Indexing, Compression, Data Structure
- AI/ Machine Learning
- Statistics
- Information Theory
- Theoretical CS :
  - Approximate, Random, Online Algorithms
- Mathematical Programming
- Computational Geometry …

Principles

数据挖掘技术构架

- Association rules discovery
- Sequential Pattern Discovery
- Cluster analysis
- Outlier Detection
- Classifier Building
- Data Cube/Data Warehouse Construction
- Visualization …

Techniques

Principles

数据挖掘技术构架

Principles

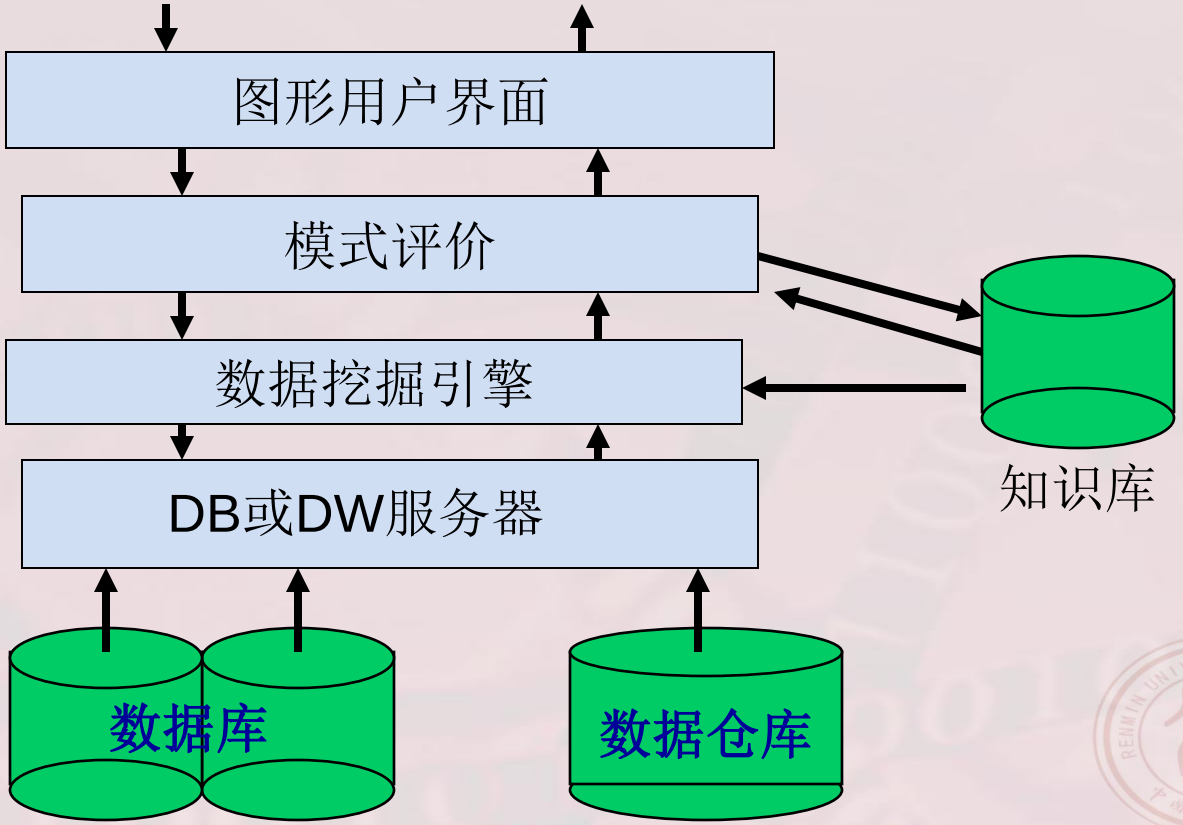
- Database Technology:
  - Indexing, Compression, Data Structure
- AI/ Machine Learning
- Statistics
- Information Theory
- Theoretical CS :
  - Approximate, Random, Online Algorithms
- Mathematical Programming
- Computational Geometry …

数据挖掘技术构架

Applications

- Customer Relationship Management (CRM)
- Web pages Searches and Analysis
- Network Security
- Geographical Data Analysis
- Genomic Database …

典型数据挖掘系统架构



数据挖掘简要发展史

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - 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)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.

数据挖掘主流会议和期刊

- **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

谢 谢！