第一章概述

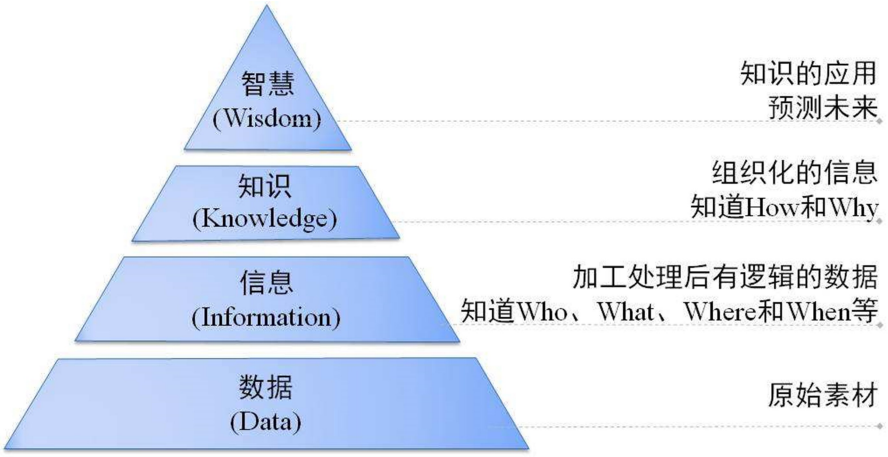
Wednesday, September 15, 2021 6:23 PM

1. 背景

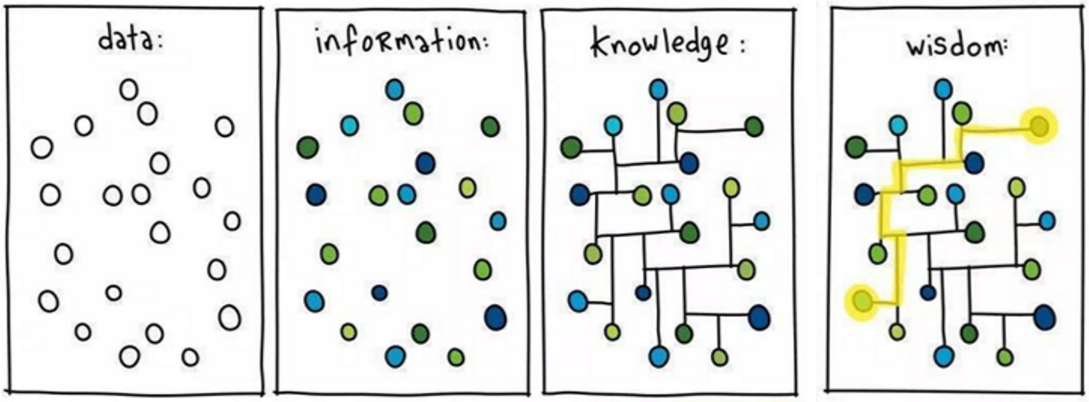
- 全球信息量以惊人的速度急剧增长
- 数据库系统无法发现数据中存在的关系和规则
- 为了充分利用现有信息资源，从海量数据中找出隐藏的知识，数据挖掘技术应运而生并显示出强大的生命力

2. 基本概念

- 数据
- 信息
- 知识
- 智慧



- 数据：指对客观事件进行记录并可以鉴别的符号，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合
- 信息：信息是具有时效性的，有一定含义的，有逻辑的、经过加工处理的、对决策有价值的数据流
- 知识：人们实践经验的结晶且为新的实践所证实的；是关于事物运动的状态和状态变化的规律；是对信息加工提炼所获得的抽象化产物
- 智慧：是人类基于已有的知识，针对物质世界运动过程中产生的问题根据获得的信息进行分析、对比、演绎找出解决方案的能力



- 数据经过处理和加工，变成了信息。
- 信息之间产生了联系，形成了知识。
- 通过现有知识，发现了一些知识之间的新关系，并且串联起来，形成了智慧。

数据挖掘的定义：

技术角度：数据挖掘（Data Mining)是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程

3. 数据挖掘发展历史

详细内容见第一章PPT。

最初，数据挖掘是作为KDD（知识发现knowledge discovery in database）中利用算法处理数据的一个步骤，其后逐渐演变成KDD的同义词。现在，人们往往不加区别地使用两者。KDD常常被称为数据挖掘（Data Mining）。

4. 主要功能

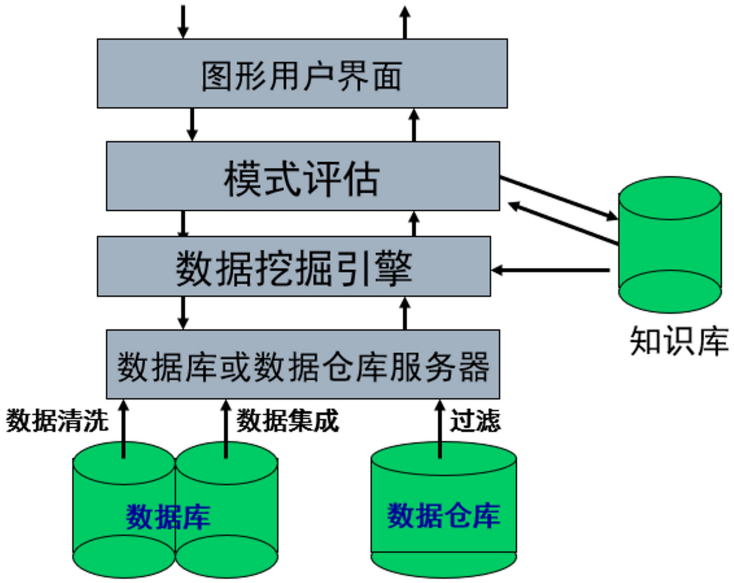
- 关联分析
- 分类
- 回归
- 聚类分析
- 离散点分析
- 时间序列分析

5. 知识发现的过程

a. 数据清洗（消除噪音或不一致的数据）
b. 数据集成（多种数据源组合到一起）
c. 数据选择（从数据库中提取与分析任务相关的数据）
d. 数据变换（变换或统一成合适挖掘的形式）
e. 数据挖掘（使用智能方法提取数据模式）
f. 模式评估（根据某种兴趣程度度量识别提供知识的真正有趣的模式）

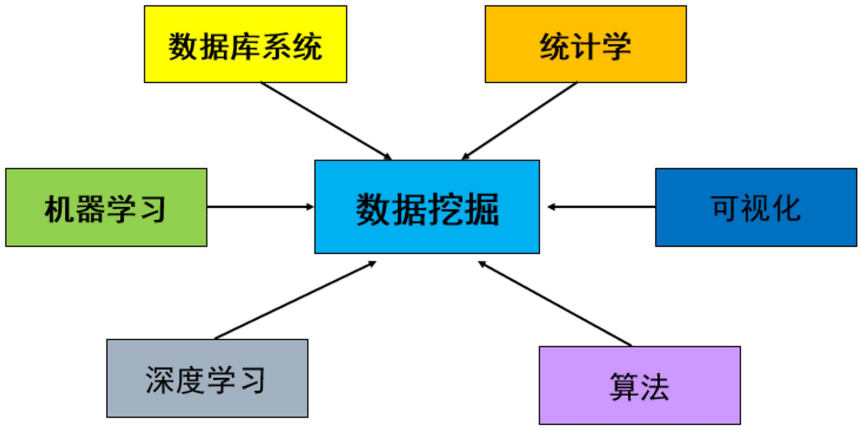
g. 知识表示（向用户提供挖掘的知识）

典型数据挖掘系统的体系结构



6. 数据挖掘与其他学科的关系

- 数据挖掘作为一门新兴的交叉学科，涉及数据库系统、统计学、机器学习、可视化、信息检索和高性能计算等诸多领域
- 此外，还与神经网络、模式识别、空间数据分析、图像处理、信号处理、概率论、图论和归纳逻辑等等领域关系密切



7. 数据挖掘的应用

- 数据库营销（Database Marketing）
- 客户群体划分（Customer Segmentation & Classification）
- 背景分析（Profile Analysis）
- 交叉销售（Cross-selling）等市场分析行为
- 客户流失分析(Churn Analysis)
- 客户信用评分(Credit Scoring)
- 欺诈甄别(Fraud Detection)
- 网站的数据挖掘（Web site data mining）
- 生物信息或基因的数据挖掘

- 文本挖掘（Textual mining）
- 多媒体挖掘

8. 未来趋势

- 发现语言的形式化描述
- 寻求数据挖掘过程中的可视化方法
- 研究在网络环境下的数据挖掘技术
- 加强对各种非结构化数据的挖掘
- 知识的维护更新
- 隐私保护