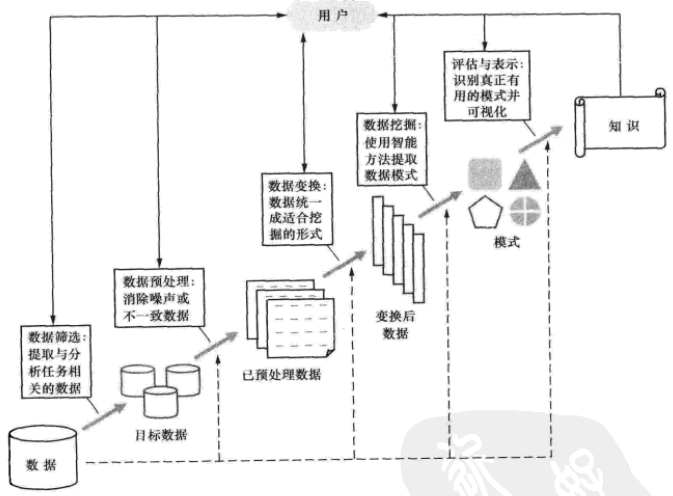
## 数据挖掘



主要任务：预测（分类与预测、演化分析）和描述特性（关联、聚类、利群点分析）

### 数据预处理

数据准备、选取、预处理（完整性、一致性、噪声、冗余、补充缺失数据）、变换（通过投影或利用数据库的其他操作减少数据量）

### 数据挖掘

确定目标、选择算法、挖掘（提取知识并以一定方式表示）

### 知识评估与表示

1. 模式评估：是否有冗余或不符合要求
2. 知识表示：使用可视化表示

## 数据挖掘算法

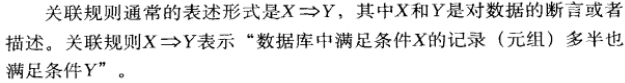
### 关联分析

支持度：满足一定条件的百分比

置信度：最终结果里满足支持度的百分比

频繁项集：大于给定支持度的集合

若大于给定的置信度，则得到一个关联规则：



### 分类和预测

找出描述和区分不同数据类或概念的模型或函数，划分的类已知。

分类规则、判定树、数学公式、神经网络

### 聚类分析

划分的类未知，将多个对象划分成簇。

划分方法（k-means）、层次方法（合并或分裂）、基于密度（临近区域的密度）、基于网格（对象量化为具有规则形状的单元格）、基于模型（数据是根据潜在的概率分布生成，为每个聚类生成模型，如统计学方法和神经网络）

### 利群点分析

Outlier。有些视为噪声而丢弃，但是小概率事件仍有意义。

基于统计、距离和偏移的方法。

### 演化分析

根据变化规律和趋势，对其建模，为相关决策提供参考。

时间相关的分析方法：

1. 趋势分析：n阶的变化平均值等
2. 相似搜索：找出与给定序列最接近的数据序列
3. 序列模式挖掘：挖掘相对时间或其他维属性出现频率高的模式
4. 周期分析：关联规则的改进。