关联分析

- 通过发现顾客放入购物篮中的不同商品之间的联系，分析顾客的购买习惯
  - 哪些物品经常被顾客购买？
  - 同一次购买中，哪些商品经常会被一起购买？
  - 一般用户的购买过程中是否存在一定的购买时间序列？
- 具体应用：购物篮分析、交叉销售、分类设计等
  - 商品货架设计：更加适合客户的购物路径
    - 两种策略：
      - 商品放近， 增加销量
      - 商品放远， 增加其他商品的销量

关联规则挖掘：简单的说，就是发现大量数据中项集之间有趣的关联

在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构。

关联规则挖掘

- 给定一系列记录，找到其中隐含的令人感兴趣的联系，用以根据记录中某些项的出现来预测其他项的产生

商场购物篮事务

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

关联规则的例子

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

关联意味着同时出现，而并非因果关系

关联规则挖掘

- 关联规则
  - 形如X->Y的蕴含表达式，其中X和Y是不相交的项集
  - 如：{牛奶, 尿布} → {啤酒}
- 规则评估度量
  - 支持度(s)：规则可以用于给定数据集的频繁程度

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

s(牛奶, 尿布->啤酒)=0.4

  - 置信度(c)：Y在包含X的事务中出现的频繁程度

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

c(牛奶, 尿布->啤酒)=0.67

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

5

milk 牛奶
diaper 尿布
beer 啤酒

总的事务数 n = 5

{(牛奶，尿布) ---> 啤酒} 项集计数 为2
项集:(牛奶，尿布) 项集计数 为 3

支持度：

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

项集计数除于总的事务数。

{(牛奶，尿布) ---> 啤酒} 的 支持度 ： 2/5 = 0.4

置信度：

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

X U Y 的项集计数 除以 X 的项集计数

{(牛奶，尿布) ---> 啤酒} 的 置信度 ： 2/3 = 0.67

注意置信度是顺序排列，前后顺序颠倒，结果不同；

应用