

博资考高级信息系统

要点

- 时代沿革中的信息系统
- 商务分析与使能
- 行为与实证分析
- 不确定性信息
- 领域展望

一、背景,

1. 大数据时代

① 数据特征 (4V)

volume 超规模, variety 富媒体 (多源异构) value (低价值密度)

velocity (流数据)

② 问题特征

2. 信息系统研究

二、范式转变

三、技术方法视角

四、关联模式

1. 支持度与置信度

$I = \{I_1, \dots, I_m\}$: 商品集合, $T = \{t_1, \dots, t_n\}$, $t_i \subseteq I$: 事务数据集

支持度 $D_{\text{supp}}(X) = \frac{|X|}{|T|}$,

$|X|$: T 中包含 X 的记录数, $|T|$: 总记录数

给定支持度阈值 α , $D_{\text{supp}}(X) \geq \alpha$ 则称 X 为频繁项集

关联规则 $X \Rightarrow Y$ ($X, Y \neq \emptyset$, $X, Y \subseteq I$, $X \cap Y = \emptyset$)

$$D_{\text{supp}}(X \Rightarrow Y) = D_{\text{supp}}(X \cup Y)$$

置信度 $D_{\text{conf}}(X \Rightarrow Y) = \frac{|X \cup Y|}{|X|} = \frac{D_{\text{supp}}(X \cup Y)}{D_{\text{supp}}(X)}$

如果 $D_{\text{supp}}(X \cup Y) \geq \alpha$, $D_{\text{conf}}(X \Rightarrow Y) \geq \beta$, $X \Rightarrow Y$ 是 α, β 合格.

2. 关联规则挖掘方法 (Apriori)

分成两大步:

① 生成频繁项集 (Frequent itemset generation)

② 生成关联规则 (Association rule generation)

① 生成频繁项集

核心 $D_{\text{supp}}(X) = \frac{|X|}{|T|} \geq \frac{|X \cup Y|}{|T|} = D_{\text{supp}}(X \cup Y)$

如果一个集合不是频繁项集, 则其所有的超集均不是

- Generate frequent itemset $L_k (k=1)$

- Repeat until no new frequent itemsets are identified.

① Candidate generation: generate C_{k+1} from L_k 连接

② Prune candidate itemsets C_{k+1} 剪枝

③ Count support for each itemset in C_{k+1} (遍历数据集)

④ Remove infrequent itemsets in C_{k+1} and get L_{k+1} .

L_2 频繁2项集 $L_2 = \{AB, AC, AD, BD, EF, DF\}$

生成候选3-项集 \Downarrow

$C_3 = \{ABC, ABD, ACD\}$

方法: 按字母顺序排列, 当前 $k-1$ 个项相同, 第 k 个不同时合并.
其余不考虑, 如: ABE, EDF, ADF 等

C_3 中由于 BC, CD $\notin L_2$, \therefore 去掉 ABC, ACD

$\Rightarrow C_3 = \{ABD\}$. 之后遍历数据集看 ABD 是否满足支持度条件, 进而得到 L_3

② 生成关联规则

核心: 置信度对于从同一个项集中生成的规则而言是单调的

$$Dconf(ABC \rightarrow D) \geq Dconf(AB \rightarrow CD) \geq Dconf(A \rightarrow BCD)$$