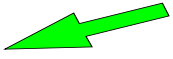


資料探勘： 概念與方法

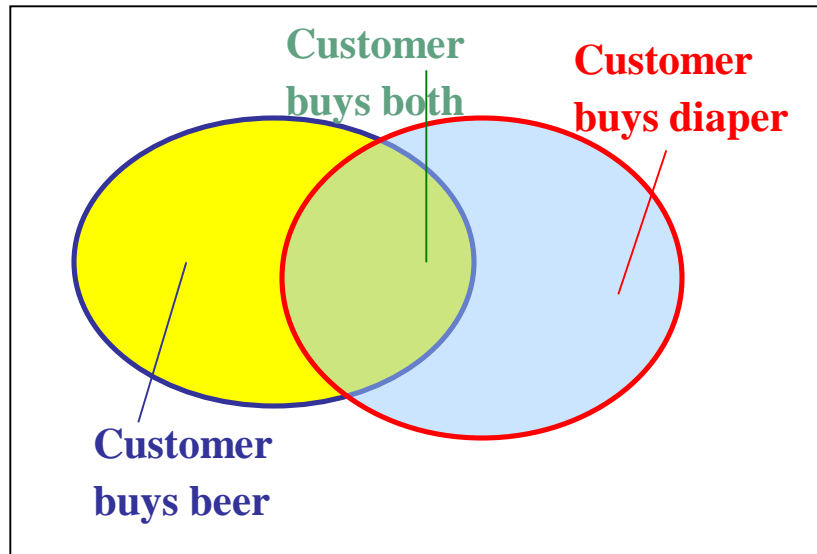
— 第五章 —

第五章：探勘頻繁樣式, 關聯與相互關係

- 基本概念與本章架構 
- 有效率並具度量頻繁項目探勘方法
- 探勘不同類型關聯規則
- 從關聯探勘到相互關係分析
- 限制式關聯探勘
- 總結

基本概念：頻繁樣式與關聯規則

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F



- 項目集 $X = \{x_1, \dots, x_k\}$
- 尋找所有具有最小支持度與信賴度規則 $X \rightarrow Y$
 - **支持度**, s , 一個交易包含 $X \cup Y$ 的機率
 - **信賴度**, c , 一個交易同時包含 X 與 Y 的條件機率

當 $sup_{min} = 50\%$, $conf_{min} = 50\%$
頻繁樣式: $\{A:3, B:3, D:4, E:3, AD:3\}$
關聯規則:


$A \rightarrow D$ (60%, 100%)

$D \rightarrow A$ (60%, 75%)

緊密與最大樣式

- 一個長樣式會包含龐大數目的子樣式, 例., $\{a_1, \dots, a_{100}\}$ 包含 $\binom{100}{1} + \binom{100}{2} + \dots + \binom{100}{100} = 2^{100} - 1 = 1.27 \times 10^{30}$ 子樣式!
- 解答: 探勘緊密與最大樣式
- 如果沒有任何一個項目集X的真母項目集 (proper super-itemset) Y與項目集X有相同的支持個數, 我們稱項目集X為緊密 (proposed by Pasquier, et al. @ ICDT'99)
- 當項目集為頻繁項目集, 而且沒有任何一個項目集的真母項目集為頻繁項目集, 我們稱項目集為最大頻繁項目集 (proposed by Bayardo @ SIGMOD'98)
- 緊密樣式為頻繁樣式的無損化壓縮
 - 降低樣式與規則數目

第五章：探勘頻繁樣式，關聯與相互關係

- 基本概念與路線圖
- 有效率並具度量頻繁項目探勘方法 
- 探勘不同類型關聯規則
- 從關聯探勘到相互關係分析
- 限制式關聯探勘
- 總結

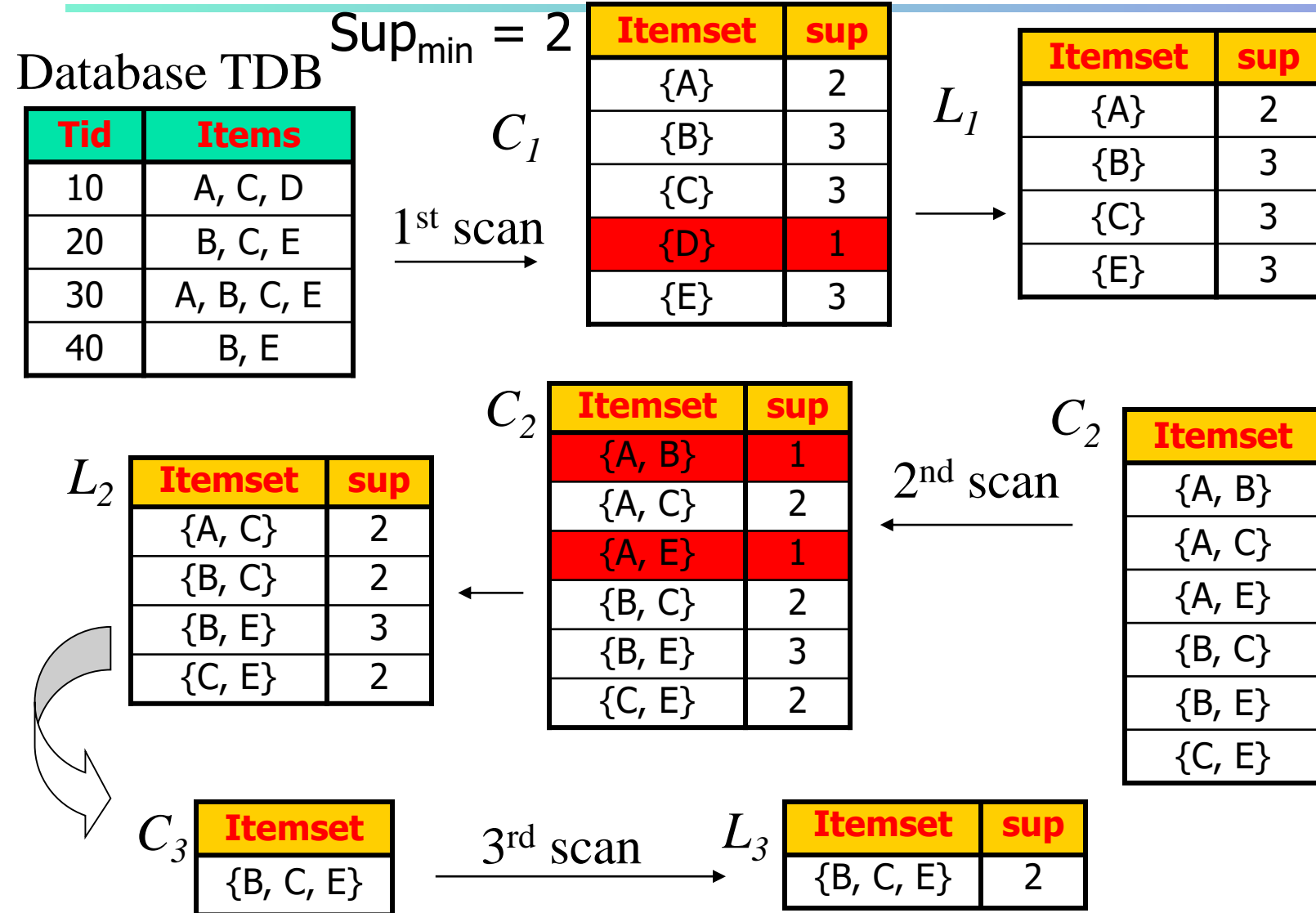
具量度頻繁項目集的探勘方法

- 頻繁樣式向下堆論性質
 - 一個頻繁項目集中非空集合的子項目集也是頻繁項目集
 - 如果 **{beer, diaper, nuts}** 為頻繁, **{beer, diaper}** 也是頻繁
 - 也就是說, 每個交易包含 {beer, diaper, nuts} 也包含 {beer, diaper}
- 具量度性探勘方法: 三個主要方法
 - Apriori (Agrawal & Srikant@VLDB'94)
 - 頻繁樣式成長 (FPgrowth—Han, Pei & Yin @SIGMOD'00)
 - 垂直資料格式 (Charm—Zaki & Hsiao @SDM'02)

Apriori: 一個產生並測試後選項目的方法

- Apriori 修剪原則: 任何一個不頻繁($k-1$)-項目集，它不會是任一個頻繁 k -項目集的子集合 (Agrawal & Srikant @VLDB'94, Mannila, et al. @ KDD' 94)
- 方法:
 - 一開始尋找頻繁1-項目集
 - 利用頻繁 k -項目集產生長度 $(k+1)$ 後選項目集
 - 將候選項目集與資料庫進行比較
 - 一直執行到沒有頻繁或候選項目集為止

The Apriori 範例



The Apriori 運算法則

運算法則：**Apriori**。根據候選產生並使用逐層的方式尋找頻繁項目集。

輸入：

- D 一個交易資料庫；
- min_sup 為最小支持度。

輸出： L 為 D 中頻繁項目集

方法：

- (1) $L_1 =$ 尋找 1-頻繁項目集
- (2) for ($k = 2; L_{k-1} \neq \phi; k++$) {
- (3) $C_k = \text{apriori_gen}(L_{k-1})$;
- (4) **for** 每個交易 $t \in D$ { // 檢視 D 為了計算個數
- (5) $C_t = \text{subset}(C_k, t)$; // 找出 t 的子集合並且它為候選集
- (6) 每個候選 $c \in C_t$
- (7) $c.\text{count}++$;
- (8) }
- (9) $L_k = \{c \in C_k \mid c.\text{count} \geq min_sup\}$
- (10) }
- (11) 傳回 $L = \bigcup_k L_k$;

The Apriori 運算法則(Cont.)

程序 apriori_gen (L_{k-1} : 頻繁 $(k-1)$ -項目集)

- ```

(1) for 每個項目集 $I_1 \in L_{k-1}$
(2) for 每個項目集 $I_2 \in L_{k-1}$
(3) if $(I_1[1] = I_2[1]) \wedge (I_1[2] = I_2[2]) \wedge \dots \wedge (I_1[k-2] = I_2[k-2]) \wedge (I_1[k-1] < I_2[k-1])$
 then {
(4) $c = I_1 \bowtie I_2$; // 結合步驟，產生候選
(5) if has_infrequent_subset(c, L_{k-1}) then
(6) delete c ; // 刪除步驟：刪除沒有結果的候選
(7) else 將 c 加入 C_k ;
(8) }
(9) 傳回 C_k

```

程序 **procedure** **has\_infrequent\_subset** ( $c$ :  $k$ -候選項目集;  $L_{k-1}$ :  $(k-1)$ -頻繁項目集) // 使用先前知識

- ```

(1)  for  $c$  的每個  $(k-1)$  子集合  $s$ 
(2)      if  $s \in L_{k-1}$  then
(3)          傳回 TRUE
(4)  傳回 FALSE

```

不需產生候選項目集來尋找頻繁項目集

- 使用區域頻繁樣式, 從短樣式來產生長樣式
 - "abc" 為頻繁樣式
 - 找出所有包含 "abc" 交易: DB|abc
 - "d" 為 DB|abc 區域頻繁樣式 → abcd 為頻繁樣式

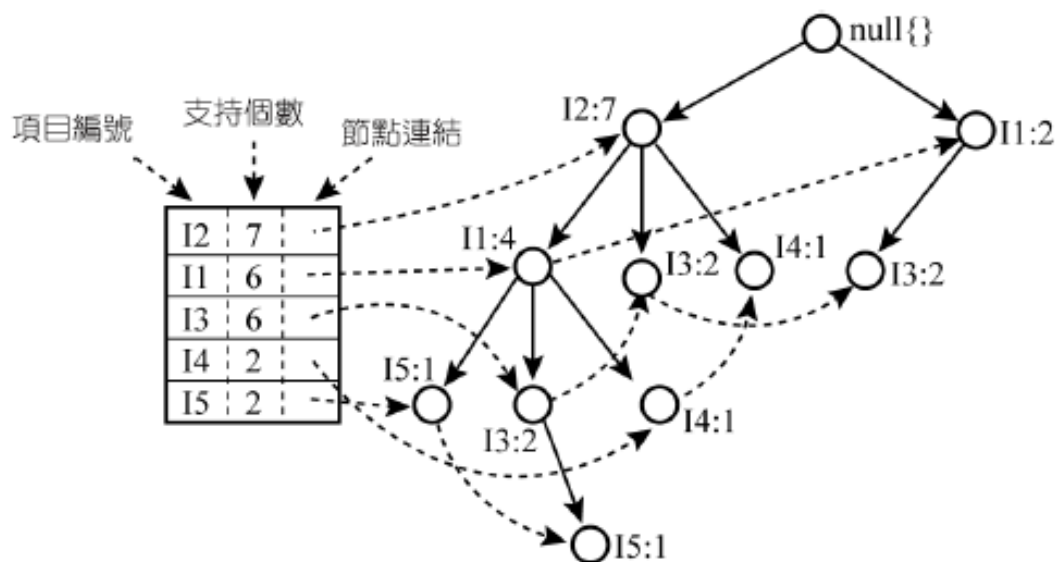
從交易資料庫建立 FP-樹

表 5.2 利用條件基礎樣式探勘 FP 樹

項目	條件基礎樣式	條件 FP 樹	產生頻繁樣式
I5	$\{\{I2, I1:1\}, \{I2, I1, I3:1\}\}$	$\langle I2:2, I1:2 \rangle$	$\{I2, I5:2\} \{I1, I5:2\}$ $\{I2, I1, I5:2\}$
I4	$\{\{I2, I1:1\}, \{I2:1\}\}$	$\langle I2:2 \rangle$	$\{I2, I4:2\}$
I3	$\{\{I2, I1:2\}, \{I2:2\}, \{I1:2\}\}$	$\langle I2:4, I1:2 \rangle, \langle I1:2 \rangle$	$\{I2, I3:4\}, \{I1, I3:4\},$ $\{I2, I1, I3:2\}$
I1	$\{\{I2:4\}\}$	$\langle I2:4 \rangle$	$\{I2, I1:4\}$

最小支持度 = 2


1. 尋找頻繁1-項目集 (單一項目樣式)
2. 對頻繁樣式依照支持個數遞減排序, f-list
3. 建立 FP-樹



使用垂直資料格式尋找頻繁項目集

- 如果資料格式記錄項目所在交易 ($\{\text{項目} : \text{交易編號}\}$)，則此種資料格式被稱為垂直資料格式
- 透過對每一對的頻繁單獨項目的交易編號進行交集得之
 - $t(X) = t(Y)$: X 與 Y 都同時出現
 - $t(X) \subset t(Y)$: 有交易 X 同時有交易 Y
- 使用差集合(Diffset)加速探勘
 - 僅記錄不同交易編號
 - $t(X) = \{T_1, T_2, T_3\}$, $t(XY) = \{T_1, T_3\}$
 - $\text{Diffset}(XY, X) = \{T_2\}$
- Eclat/MaxEclat (Zaki et al. @KDD'97), VIPER(P. Shenoy et al.@SIGMOD'00), CHARM (Zaki & Hsiao@SDM'02)

第五章：探勘頻繁樣式，關聯與相互關係

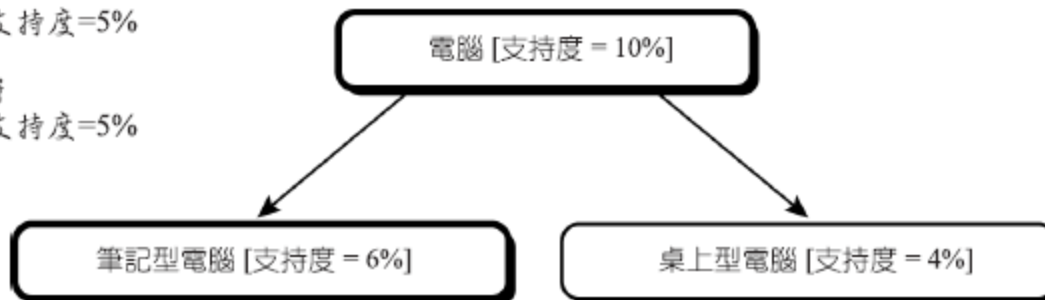
- 基本概念與路線圖
- 有效率並具度量頻繁項目探勘方法
- 探勘不同類型關聯規則 
- 從關聯探勘到相互關係分析
- 限制式關聯探勘
- 總結

探勘多層次關聯規則

- 項目經常構成階層
- 彈性支持設定
 - 在較低層的項目會具較低支持
- 共享多維度探勘 (Agrawal & Srikant@VLB'95, Han & Fu@VLDB'95)

單一支持度

第一層
最小支持度=5%
第二層
最小支持度=5%



較低支持度

第一層
 $\text{min_sup} = 5\%$

第二層
 $\text{min_sup} = 3\%$

多層次關聯：過濾多餘

- 多層次關聯規則探勘的副作用，就是在層與層之間產生許多關聯規則，這些是因為概念架構祖先關係所造成。
- Example
 - 購買 (X, “筆記型電腦”) \Rightarrow 購買 (X, “HP印表機”)
[support = 8%, confidence = 70%]
 - 購買 (X, “IBM筆記型電腦”) \Rightarrow 購買 (X, “HP印表機”) [support = 2%, confidence = 72%]
- 第一個規則為第二規則祖先
- 當一個規則的支持度相當接近規則本身祖先規則的期望值時，則該規則是多餘的。

探勘多維度關聯

- 單一維度關聯規則:

購買 (X, “數位相機”) \Rightarrow 購買 (X, “HP印表機”)

- 多維度關聯規則: ≥ 2 維度或敘述

- 維度間關聯規則 (有重複敘述)

年齡 (X, “20...29”) \wedge 職業 (X, “學生”) \Rightarrow 購買 (X, “筆記型電腦”)

- 混合維度關聯規則 (重複敘述)

年齡 (X, “20...29”) \wedge 購買 (X, “筆記型電腦”) \Rightarrow 購買 (X, “HP印表機”)

- 類別屬性: 類別屬性的值有限，而且在值與值之間並無特定大小順序存在

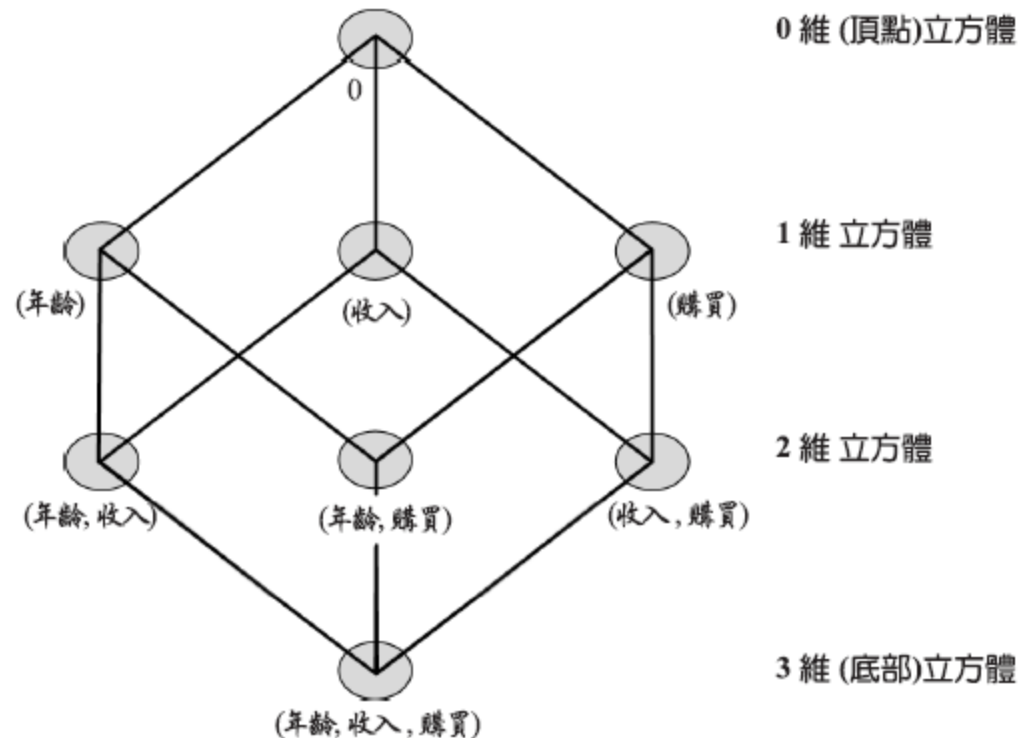
- 數值屬性: 數值屬性的值有特定順序—離散化, 群組與梯度方法

探勘數值關聯

- 將數值屬性轉變為類別資料方法
- 1. 根據預先定義概念階層進行各訂離散化(資料方塊方法)
- 2. 根據資料分佈進行動態離散化 (數量規則, 例., Agrawal & Srikant@SIGMOD96)
- 3. 群組: 距離式關聯規則 (例., Yang & Miller@SIGMOD97)
 - 單維度群組而後進行關聯

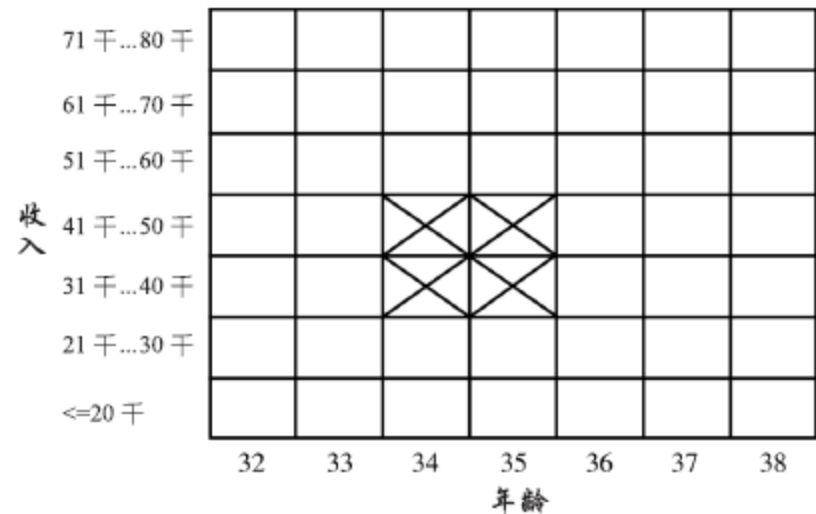
固定分割數值屬性

- 利用事先定義概念架構或是其他分割方法進行分割.
- 將數值屬性轉變為類別資料.
- 資料方塊非常適合探勘.
- n -維度立方體的節點對應至相關敘述集.
- 經由資料方塊探勘會比較快.

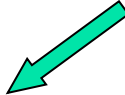


數值關聯規則

- 由Lent, Swami and Widom ICDE'97 提出
- 動態分割數值屬性的方法
 - 尋找最大信賴度關聯規則或最短關聯規則
- 2-維數值式關聯規則： $A_{quan1} \wedge A_{quan2} \Rightarrow A_{cat}$
- 將滿足類別屬性成對的數值屬性對應至2-維方格，接下來搜尋這些方格並進行群組產生關聯規則
- 範例
年齡 (X, “30...39”) \wedge 收入 (X, “42千...48千”) \Rightarrow 購買 (X, “高解析度電視”)



第五章：探勘頻繁樣式, 關聯與相互關係

- 基本概念與路線圖
- 有效率並具度量頻繁項目探勘方法
- 探勘不同類型關聯規則
- 從關聯探勘到相互關係分析 
- 限制式關聯探勘
- 總結

有趣指標：相互關係（增益, Lift）

- 購買 (X, “電腦遊戲”) \Rightarrow 購買 (X, “影帶”) [支持度40%, 信賴度66%] 是誤導的
 - 因為購買影帶機率為75%，遠高於66%
- 增益 (lift) 是一個相互關係的簡單指標
- 增益值 = $P(\{\text{電腦遊戲}, \text{影帶}\}) / (P(\{\text{電腦遊戲}\}) \times P(\{\text{影帶}\})) = 0.40 / (0.60 \times 0.75) = 0.89$

$$lift = \frac{P(A \cup B)}{P(A)P(B)}$$

表5.7 2 × 2 競爭表說明購買遊戲與影帶的交易

	遊戲	影帶	$\Sigma_{\text{列}}$
遊戲	4,000	3,500	7,500
影帶	2,000	500	2,500
$\Sigma_{\text{欄}}$	6,000	4,000	10,000

lift 與 χ^2 是否為好的相互關係指標?

- "購買 核桃 \Rightarrow 購買牛奶 [1%, 80%]" 是誤導
 - 如果 85% 的客戶都購買牛奶
- 支持度與信賴度無法有效代表相互關係
- 許多有趣指標? (Tan, Kumar, Sritastava @KDD'02)

$$lift = \frac{P(A \cup B)}{P(A)P(B)}$$

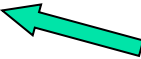
	牛奶	牛奶	$\Sigma_{列}$
咖啡	mc	\overline{mc}	c
咖啡	$m\overline{c}$	$\overline{m\overline{c}}$	\overline{c}
$\Sigma_{欄}$	m	\overline{m}	Σ

$$all_conf = \frac{\sup(X)}{\max_item_sup(X)}$$

$$cosine(A, B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}}$$

資料集	mc	\overline{mc}	$m\overline{c}$	$\overline{m\overline{c}}$	$all_conf.$	$cosine$	$lift$	χ^2
A_1	1,000	100	100	100,000	0.91	0.91	83.64	83,452.6
A_2	1,000	100	100	10,000	0.91	0.91	9.26	9,055.7
A_3	1,000	100	100	1,000	0.91	0.91	1.82	1,472.7
A_4	1,000	100	100	0	0.91	0.91	0.99	9.9
B_1	1,000	1,000	1,000	1,000	0.50	0.50	1.00	0.0
C_1	100	1,000	1,000	100,000	0.09	0.09	8.44	670.0
C_2	1,000	100	10,000	100,000	0.09	0.29	9.18	8,172.8
C_3	1	1	100	10,000	0.01	0.07	50.0	48.5

第五章：探勘頻繁樣式，關聯與相互關係

- 基本概念與路線圖
- 有效率並具度量頻繁項目探勘方法
- 探勘不同類型關聯規則
- 從關聯探勘到相互關係分析
- 限制式關聯探勘 
- 總結

限制式關聯探勘

- 知識類型限制:
 - 如關聯或相互關係.
- 資料限制 — 使用類似 SQL 查詢
- 維度/層限制
 - 指定所需資料集的維度或是概念架構的某些層
- 規則限制
 - 小銷售 ($\text{price} < \$10$) 引發 大銷售 ($\text{sum} > \200)
- 有趣限制
 - 有效規則: 最小支持度 $\geq 3\%$, 最小信賴度 $\geq 60\%$

限制推移的反一致性

TDB (min_sup=2)

- 反一致性
 - 任何一個項目集只要它不滿足條件，那包含它所有的超集合也不會滿足條件
 - $sum(S.Price) \leq v$ 具反一致性
 - $sum(S.Price) \geq v$ 不具反一致性
- 例. C: $range(S.profit) \leq 15$ 具反一致性
 - 當項目及 *ab* 違反 C
 - 所有 *ab* 超集合也違反 C

TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Item	Profit
a	40
b	0
c	-20
d	10
e	-30
f	30
g	20
h	-10

限制推移的一致性

TDB (min_sup=2)

■ 一致性

- 任何一個項目集只要它滿足條件，那包含它所有的超集合也會滿足條件
- $sum(S.Price) \geq v$ 具一致性
- $min(S.Price) \leq v$ 具一致性

■ 例. C: $range(S.profit) \geq 15$

- 項目集 *ab* 滿足 C
- 所有 *ab* 超集合也會滿足C

TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Item	Profit
a	40
b	0
c	-20
d	10
e	-30
f	30
g	20
h	-10

簡明

- 簡明:
 - 假設 A_I 滿足簡明限制 C 的項目集, 則任何滿足 C 的集合 S 一定根據 A_I , 也就是說, S 包含 A_I 的子集合
 - 想法: 不需檢視交易資料庫, 項目集 S 是否滿足限制 C 可由項目選擇來決定
 - $\min(S.Price) \leq v$ 具簡明
 - $\sum(S.Price) \geq v$ 不具簡明
- 最佳化: 如果 C 具簡明, C 為可推移事先計算

轉換嚴格限制

- 藉由適當排序將嚴格限制轉換為至反一致或一致
- 例 C: $\text{avg}(S.\text{profit}) \geq 25$
 - 將項目值遞減排序
 - $\langle a, f, g, d, b, h, c, e \rangle$
 - 如果項目 afb 違反 C
 - $afbh, afb^*$ 也違反 C
 - 變成反一致!

TDB (min_sup=2)


TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Item	Profit
a	40
b	0
c	-20
d	10
e	-30
f	30
g	20
h	-10

限制式探勘

限制	反一致	一致	簡明
$v \in S$	no	yes	yes
$S \supseteq V$	no	yes	yes
$S \subseteq V$	yes	no	yes
$\min(S) \leq v$	no	yes	yes
$\min(S) \geq v$	yes	no	yes
$\max(S) \leq v$	yes	no	yes
$\max(S) \geq v$	no	yes	yes
$\text{count}(S) \leq v$	yes	no	weakly
$\text{count}(S) \geq v$	no	yes	weakly
$\text{sum}(S) \leq v (\forall a \in S, a \geq 0)$	yes	no	no
$\text{sum}(S) \geq v (\forall a \in S, a \geq 0)$	no	yes	no
$\text{range}(S) \leq v$	yes	no	no
$\text{range}(S) \geq v$	no	yes	no
$\text{avg}(S) \theta v, \theta \in \{\leq, \geq\}$	convertible	convertible	no
$\text{support}(S) \geq \xi$	yes	no	no
$\text{support}(S) \leq \xi$	no	yes	no
$\text{all_confidence}(S) \geq \xi$	yes	no	no
$\text{all_confidence}(S) \leq \xi$	no	yes	no

第五章：探勘頻繁樣式，關聯與相互關係

- 基本概念與路線圖
- 有效率並具度量頻繁項目探勘方法
- 探勘不同類型關聯規則
- 從關聯探勘到相互關係分析
- 限制式關聯探勘
- 總結 

頻繁樣式探勘：總結

- 頻繁樣式探勘—資料探勘中一項重要工作
- 具量度性頻繁樣式探勘方法
 - Apriori (產生候選並測試)
 - 映射式 (FPgrowth, CLOSET+, ...)
 - 垂直格式方法 (CHARM, ...)
- 探勘不同規則與樣式
- 限制式探勘