關聯規則探勘

國立宜蘭大學資訊工程系
吳政瑋 助理教授
silvemoonfox@hotmail.com

關聯規則探勘 (Association Rule Mining)

Data Mining 範例: 啤酒與尿布

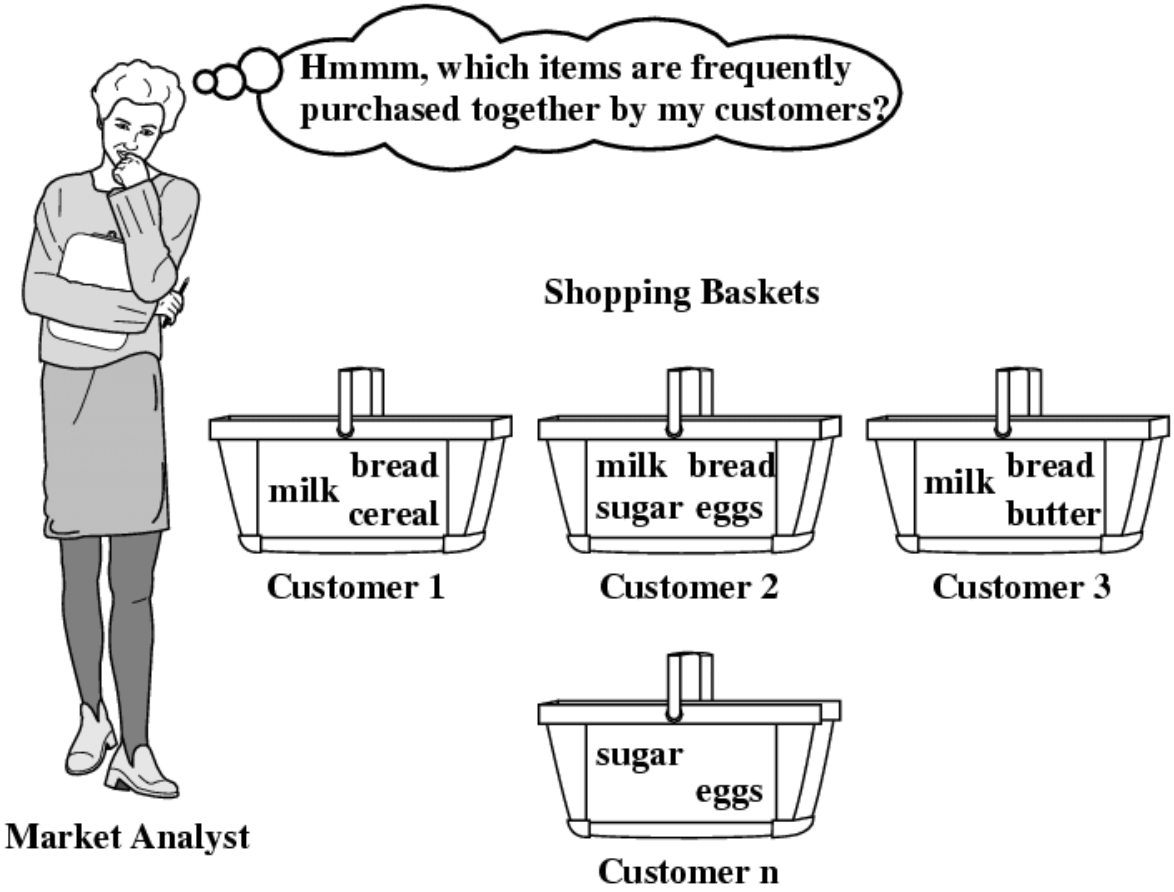
- https://www.youtube.com/watch?v=2W-yq0lGnrc

http://slicedbread.co.uk/blog/how-to-strike-gold-with-microsoft-data-mining-tools

沃瑪以關聯分析技術找出颶風與草莓餅乾之關聯性

- https://www.youtube.com/watch?v=rvij5kVttu0

關聯規則探勘
(又稱:購物籃分析)



TID	Transaction
T_1	{Milk, Bread, Cereal}
T_2	{Milk, Bread, Sugar, Eggs}
T_3	{Milk, Bread, Butter}
T_4	{Sugar, Eggs}

Milk → Bread [75%, 100%]

關聯規則探勘之相關定義 (Cont. 1/2)

- **Item (項目)**
  - e.g., A, B, C, D, E, F
- **Itemset (項目集)**
  - e.g., {ABC}
- **Contain (包含)**
  - e.g., {AC} 包含於 T_1, T_2
  - e.g., g({AC}) = {1, 2}
- **Subset (子集合)**
  - e.g., {ABC} 的子集合有{{A}, {B}, {C}, {AB}, {AC}, {BC}, {ABC}}
- **Superset (超集合)**
  - e.g., {ABC} 為{A}, {B}, {C}, {AB}, {AC}, {BC}, {ABC} 的超集合
  - e.g., {ABCD} 為 {ABC} 的超集合

Transaction Database	
TID	Transaction
1	ABC
2	AC
3	AD
4	BEF

關聯規則探勘之相關定義 (Cont. 1/3)

- **Support Count (支持數)**
  - e.g., $SC(\{AC\}) = 2$
- **Support of Itemset (支持度)**
  - e.g., $sup(\{AC\}) = SC(\{AC\})/|D| = 2/4 = 50\%$

Transaction Database	
TID	Transaction
1	ABC
2	AC
3	AD
4	BEF

- **Frequent Itemset (頻繁項目集)**
  - 令 δ 為**最小之持數門檻值 (*minimum support count threshold*)**。若一個項目集 X 的支持數不亞於 δ，則稱 X 為**頻繁項目集(*frequent itemset*)**；反之則稱 X 為**非頻繁項目集 (*infrequent itemset*)**。
  - e.g., $\delta = 2$，$SC(\{AC\}) \geq \delta$ ，{AC}為頻繁項目集

關聯規則探勘之相關定義 (Cont. 2/3)

- **Frequent Itemset (頻繁項目集)**
  - 令 θ 為**最小之持度門檻值 (*minimum support threshold*)**。若一個項目集 X 的支持度不亞於 θ，則稱 X 為**頻繁項目集(*frequent itemset*)**；反之則稱 X 為**非頻繁項目集 (*infrequent itemset*)**。
  - e.g., $\theta = 50\%$，$sup(\{AC\}) \geq \theta$，{AC}為頻繁項目集

Transaction Database	
TID	Transaction
1	ABC
2	AC
3	AD
4	BEF

相關商品組合探勘技術可切割成兩大主要步驟

輸入：
1. 交易資料庫
2. 使用者參數θ

↓

相關商品組合探勘演算法

↓

輸出：
1. 所有的相關商品組合

輸入：
1. 所有相關商品組合
2. 使用者參數δ

↓

關聯規則產生演算法

↓

輸出：
1. 所有的關聯規則

關聯規則探勘之相關定義 (Cont. 2/2)

- **Rule (規則)**
  - 一條規則的形式 $X \rightarrow Y$
  - e.g., {A}→{C}
- **Support of Rule (規則的支持度)**
  - $sup(X \rightarrow Y) = sup(XY)$
  - e.g., $sup(\{A\} \rightarrow \{C\}) = sup(\{AC\})$
- **Confidence of Rule (規則的信賴度)**
  - $conf(X \rightarrow Y) = SC(XY) / SC(X)$
  - e.g., $sup(\{A\} \rightarrow \{C\}) = SC(\{AC\}) / SC(\{A\}) = 2/3$
- **Association Rule (關聯規則)**
  - 若一條規則 $X \rightarrow Y$ 的**支持度**不亞於**使用者自訂的最小之持度門檻值θ**，且其**信賴度**不亞於使用者自訂的**最小信賴度門檻值 (Minimum Confidence Threhold) δ**，則稱此規則為**關聯規則**。
  - $sup(X \rightarrow Y) \geq \theta$ and $conf(X \rightarrow Y) \geq \delta$

Transaction Database	
TID	Transaction
1	ABC
2	AC
3	AD
4	BEF

關聯規則探勘技術之相關定義

- 規則的形式 為 $X \rightarrow Y$
  - 例如：{A}→{C}
  - X 和 Y 皆為商品組合，且 $X \cap Y = \varnothing$
  - X 稱為前項/左項
  - Y 稱為後項/右項

關聯規則探勘技術之相關定義

- 關聯規則
- 定義:

1. 若一條規則 $X \rightarrow Y$ 發生的次數不亞於使用者自訂的 最小之持度 θ；
2. 而且，它的信賴程度不亞於使用者自訂的最小信賴程度 δ；
3. 則稱$X \rightarrow Y$為相關商品組合規則。

對每一個 FI 產生規則
舉例說明

Frequent Itemset	Count
{A}	2
{B}	3
{C}	3
{E}	3

Frequent Itemset	Count
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

Frequent Itemset	Count
{B, C, E}	2

- 頻繁項目集{AC}可產生兩條規則，這些規則的支持度相同
  - {A} → {C}, *conf* = *SC*({AC})/*SC*({A}) = 2/2 = 100%
  - {C} → {A}, *conf* = *SC*({AC})/*SC*({C}) = 2/3 = 66.6%

對每一個 FI 產生規則
舉例說明

Frequent Itemset	Count
{A}	2
{B}	3
{C}	3
{E}	3

Frequent Itemset	Count
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

Frequent Itemset	Count
{B, C, E}	2

- {BCE}可產生$2^3 - 2$ 條規則，這些規則的支持度皆相同
  - {B} → {CE}, *conf* = *SC*({BCE})/*SC*({B}) = 2/3 = 66.6%
  - {C} → {BE}, *conf* = *SC*({BCE})/*SC*({C}) = 2/3 = 66.6%
  - {E} → {BC}, *conf* = *SC*({BCE})/*SC*({E}) = 2/3 = 66.6%
  - {CE} →{B} , *conf* = *SC*({BCE})/*SC*({CE}) = 2/2 = 100%
  - {BE} → {C}, *conf* = *SC*({BCE})/*SC*({BE}) = 2/3 = 66.6%
  - {BC} → {E}, *conf* = *SC*({BCE})/*SC*({BC}) = 2/3 = 66.6%

$\theta = 2$

編號	交易
10	ACD
20	BCE
30	ABCE
40	BE

1st scan

Candidate Itemset	count
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

Frequent Itemset	Count
{A}	2
{B}	3
{C}	3
{E}	3

Candidate Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

Candidate Itemset	count
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

Frequent Itemset	Count
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

Candidate Itemset
{B, C, E}

3rd scan

Frequent Itemset	Count
{B, C, E}	2