序列型樣探勘與軟體操作

國立宜蘭大學資訊工程系
吳政瑋 助理教授
wucw@niu.edu.tw

本堂教學重點

- 傳統頻繁項目集探勘缺點
- 序列型樣探勘目的
- 序列型樣探勘應用
- 序列型樣探勘技術
- 序列型樣探勘軟體操作

傳統頻繁項目集探勘缺點 (Cont. 1/2)

- 頻繁項目集探勘(Frequent Itemset Mining；簡稱FIM)常用於**購物籃分析(Market Basket Analysis)**的應用中。

- 在 FIM 的架構中
  - 僅能找出大部份顧客經常**同時一起購買**的商品組合
  - 無法找出大部份顧客**循序購買**商品的行為

傳統頻繁項目集探勘缺點 (Cont. 2/2)

- 隨著**條碼技術(Bar-code Technology)**的進步，零售組織能夠收集大量的銷售資料及交易資料。
- 此類資料中的記錄通常包含**交易日期(Transaction Date)**以及顧客所**購買的商品(Bought Items)**。
- 當顧客使用信用卡或會員卡購物時，或需填寫訂單時，此類資料通常還會包含**顧客的 ID (Customer ID)**。

A Customer-Sequence Database & Sequential Patterns

Transaction Time	Customer Id	Items Bought
June 10 '93	2	10, 20
June 12 '93	5	90
June 15 '93	2	30
June 20 '93	2	40, 60, 70
June 25 '93	4	30
June 25 '93	3	30, 50, 70
June 25 '93	1	30
June 30 '93	1	90
June 30 '93	4	40, 70
July 25 '93	4	90

Customer Id	TransactionTime	Items Bought
1	June 25 '93	30
1	June 30 '93	90
2	June 10 '93	10, 20
2	June 15 '93	30
2	June 20 '93	40, 60, 70
3	June 25 '93	30, 50, 70
4	June 25 '93	30
4	June 30 '93	40, 70
4	July 25 '93	90
5	June 12 '93	90

Customer Id	Customer Sequence
1	⟨ (30) (90) ⟩
2	⟨ (10 20) (30) (40 60 70) ⟩
3	⟨ (30 50 70) ⟩
4	⟨ (30) (40 70) (90) ⟩
5	⟨ (90) ⟩

Sequential Patterns with support > 25%
⟨ (30) (90) ⟩
⟨ (30) (40 70) ⟩

序列型樣探勘的應用

- **購物籃分析**
  - ex: <(營養補品), (無香味乳液), (無香味肥皂), (大包裝棉花球)>
- **商品出租**
  - ex: <(星際大戰), (帝國大反擊), (絕地大反攻)>
- **路徑分析**
  - ex: <(弘志路), (女中路), (神農路), (宜蘭大學)>
- **3C用品銷售**
  - ex: <(電腦), (列表機, 列印紙), (墨水匣)>

序列型樣探勘相關定義 (Cont. 1/3)

- **Item**
  - ex: (40), (70)
- **Itemset**
  - ex: (40 70)
- **Sequence**
  - ex: <(30) (40 70) (90)>
- **Sub-sequence**
  - ex: <(30)(90)> is a sub-sequence of <(30)(40 70)(90)>
- **Super-sequence**
  - ex: <(30) (40 70) (90)> is a super-sequence of <(30)(90)>
- **Support count of a sequence**
  - ex: <(30)(90)> : 2
- **Support of a sequence**
  - ex: <(30)(90)> : 2/5 = 40%

- **Sequence Database**

Customer Id	Customer Sequence
1	⟨(30) (90)⟩
2	⟨(10 20) (30) (40 60 70)⟩
3	⟨(30 50 70)⟩
4	⟨(30) (40 70) (90)⟩
5	⟨(90)⟩

序列型樣探勘相關定義 (Cont. 2/3)

- **Sequential Pattern (SP)**
  - A sequence having a support no less than *min_sup*
  - ex: *min_sup* = 25%, <(30)(90)>, <(30)(40 70)> are sequential patterns
- **Frequent Sequence**
  - A sequential pattern is also called ***frequent sequence***
- **Infrequent Sequence**
  - A sequence is called ***infrequent sequence*** if its support is less than *min_sup*
- **Problem Statement**
  - Given a *min_sup*, a sequence database D, the problem to be solved is to find from *D* all the sequences having a support no less than *min_sup*.

序列型樣探勘的主要挑戰

- 項目的爆炸性組合問題
  - 例如: 有三個項目a, b, c，則可以組出的 itemset 有 {a}, {b}, {c}, {ab}, {ac}, {bc}, {abc}
  - 可以組出的 sequence 有⟨a⟩, ⟨b⟩, ⟨c⟩, ⟨aa⟩, ⟨ab⟩, ⟨ac⟩, ⟨ba⟩, ⟨bb⟩, ⟨bc⟩, ⟨ca⟩, ⟨cb⟩, ⟨cc⟩, ⟨abc⟩, ⟨abcc⟩, ⟨abcca⟩…
- 計算成本更高
- 搜尋空間更大
- 空間用量更多

以暴力法探勘 Sequential Patterns

- **A Brute Force Approach for SP Mining**
  - 列舉出資料庫中所有 Sequence，並掃描資料庫計算每個 Sequence 的支持數。

- **暴力法的主要缺點**
  - **空間問題**：若將所有 Sequence 儲存於 Memory 中，則需耗費大量空間。
  - **時間問題**：搜尋 Sequence 的位置、比對 Sequence 耗費大量執行時間。

常見的序列型樣探勘演算法

- 常見的序列型樣探勘演算法包括：
  - AprioriAll
  - GSP
  - FreeSpan
  - SPADE
  - PrefixSpan

GSP 演算法簡介

- GSP 演算法沿用 Apriori 架構，以廣度優先方式找出每個長度的 Sequential Patterns。
- **Step 1**：找出 FS_1，再利用 FS_1 組出長度為 2 的 Candidate Sequences (CS_2)
- **Step 2**：掃描原始資料庫一次，計算 $CS_k (k \geq 2)$的支持數，找出 FS_k。
- **Step 3**：以 FS_k 組出 CS_{k+1}。
- **Step 4**：$k = k+1$，重複執行 Step 2 & 3 直到沒有任何 Candidate Sequences 可再被產生出來。

舉例說明：GSP 演算法

Sequence Database	
SID	**Sequence**
S_1	⟨(a)(b)(a)(c)⟩
S_2	⟨(a)(a)(c)(b)⟩
S_3	⟨(b)(c)⟩
S_4	⟨(c)(b)(e)(b)⟩

Seq	SC
⟨(a)⟩	2
⟨(b)⟩	4
⟨(c)⟩	4
⟨(d)⟩	1

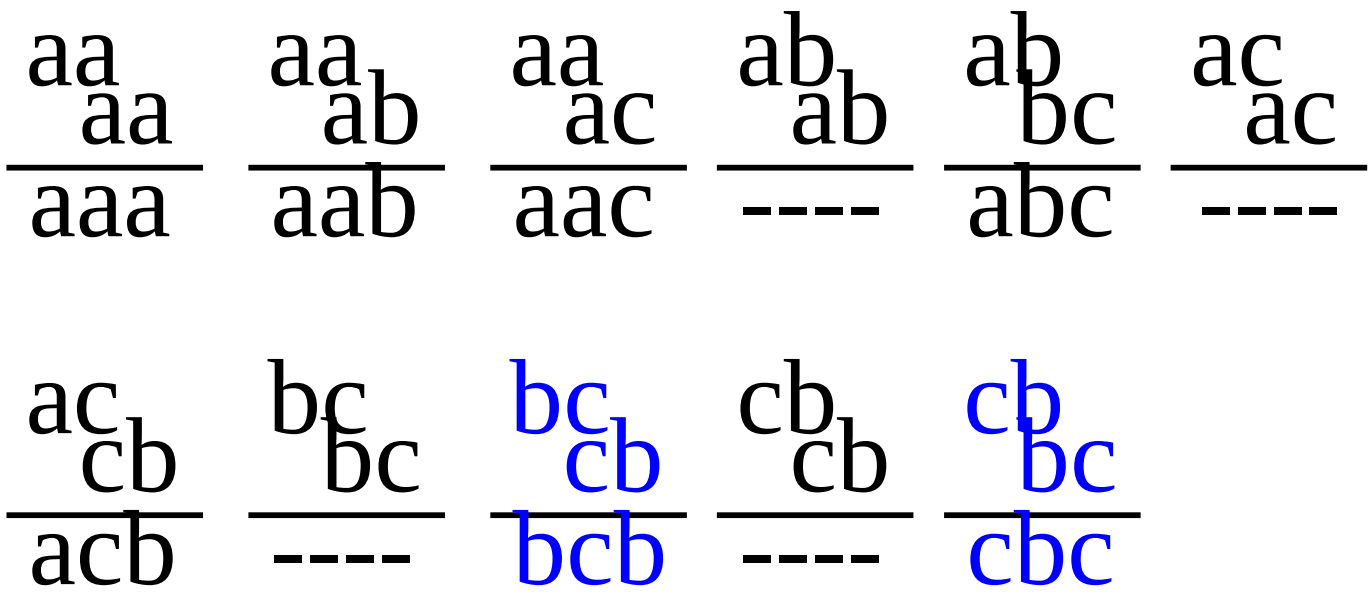
Seq	SC
⟨(a)(a)⟩	2
⟨(a)(b)⟩	2
⟨(a)(c)⟩	2
⟨(b)(a)⟩	1
⟨(b)(b)⟩	1
⟨(b)(c)⟩	2
⟨(c)(a)⟩	0
⟨(c)(b)⟩	2
⟨(c)(c)⟩	0

Seq	SC
⟨(a)(a)(a)⟩	0
⟨(a)(a)(b)⟩	1
⟨(a)(a)(c)⟩	2
⟨(a)(b)(c)⟩	1
⟨(a)(c)(b)⟩	1
⟨(b)(c)(b)⟩	0
⟨(c)(b)(c)⟩	0

Candidate Sequence 產生方式

- 產生規則：x 去頭、y 去尾、中間一樣，就可以組出 Candidate Sequence

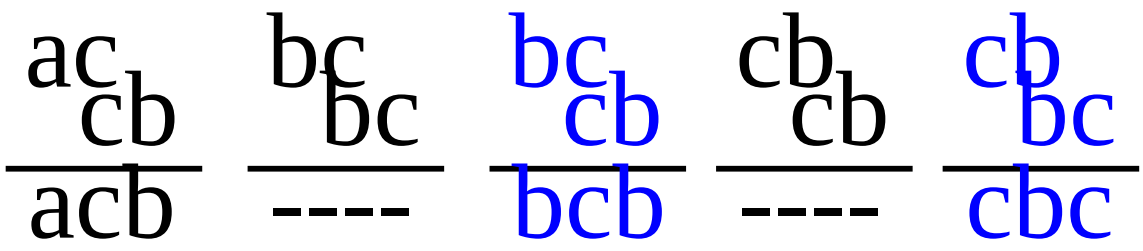
Seq	SC
⟨(a)(a)⟩	2
⟨(a)(b)⟩	2
⟨(a)(c)⟩	2
⟨(b)(a)⟩	1
⟨(b)(b)⟩	1
⟨(b)(c)⟩	2
⟨(c)(a)⟩	0
⟨(c)(b)⟩	2
⟨(c)(c)⟩	0



Subsequence Checking

- 每當一個 Candidate (k+1)-Sequence Z 被產生出來時，演算法將執行 Subsequence Checking，檢查 Z 長度減 1 的子序列是否皆為 Frequent Sequence。
- 若 Z 存在一個長度減 1 的子序列不為 Frequent Sequence，則 Z 不可能為 Frequent Sequence。

Seq	SC
⟨(a)(a)⟩	2
⟨(a)(b)⟩	2
⟨(a)(c)⟩	2
⟨(b)(c)⟩	2
⟨(c)(b)⟩	2



Downward Closure Property in SPM

- 在傳統 SPM 的架構中，Sequence 遵循**向下封閉性(Downward Closure Property)**
  - 若 S 為 Frequent Sequence，則它所有的 Sub-sequence 皆為 Frequent Sequence
  - 若 S 為 Infrequent Sequence，則它所有的 Super-sequence 皆為 Infrequent Sequence

Sequence Database	
SID	**Sequence**
S_1	⟨(a)(b)(a)(c)⟩
S_2	⟨(a)(a)(c)(b)⟩
S_3	⟨(b)(c)⟩
S_4	⟨(c)(b)(e)(b)⟩

Seq	SC
⟨(a)⟩	2
⟨(b)⟩	4
⟨(c)⟩	4
⟨(d)⟩	1

Applications

- **Basket Analysis**
  - i.e., Mining sequential patterns from sequence of transactions
- **Web Mining**
  - i.e., Mining sequential patterns from web logs
- **Mobile Mining**
  - i.e., Mining sequential patterns from moving logs (trajectories)
- **Bioinformatics**
  - i.e., Mining sequential patterns form DNA sequences
- **Multimedia**
  - i.e., Mining repeated patterns form music data

Research Topics

- Constrained Sequential Pattern Mining
- Time-Interval Sequential Patterns
- Time-Gap Sequential Patterns
- Non-redundant Sequential Patterns
- Incremental Mining Sequential Patterns
- Mining Sequential Patterns over Data Streams
- High Utility Sequential Patterns
- Mining Sequential Patterns from Uncertain Data

結論

- 傳統頻繁項目集探勘缺點
- 序列型樣探勘目的
- 序列型樣探勘應用
- 序列型樣探勘技術
- 序列型樣探勘軟體操作

Reference

- [1] Agrawal R., Srikant R., Mining sequential patterns, Proceedings 1995 Int. Conf. Very Large Data Bases (VLDB'94), pp. 487-499, 1995.
- [2] Srikant R.,Agrawal R., Mining sequential pattern: Generalizations and performance improvements, Proceedings 5th Int. Conf. Extending Database Technology (EDBT'96), pp. 3-17, 1996.
- [3] Han J., Dong G.,Mortazavi-AslB., Chen Q.,DayalU., Hsu M.-C., Freespan: Frequent pattern-projected sequential pattern mining, Proceedings 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD'00), pp. 355-359, 2000.
- [4] Pei J., Han J., Mortazavi-Asl J., Pinto H., Chen Q., Dayal U., Hsu M., PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth , 17th International Conference on Data Engineering (ICDE), April 2001.
- [5] Zhao Q., BhowmickS. S., Sequential Pattern Mining: A Survey. Technical Report Center for Advanced Information Systems, School of Computer Engineering, Nanyang Technological University, Singapore, 2003.

教師資訊

- 姓名：吳政瑋 (小吳老師)
- 現職：宜大資工助理教授
- 學歷：成功大學資工博士
- 研究興趣：資料探勘、人工智慧、AIoT應用
- 通訊方式
  - 電子信箱：wucw@niu.edu.tw
  - 校內電話: (03)9317331
  - Line: silvemoonfox
  - Office: 格致大樓E405室
  - 數位學習園區
- 實驗室：AI與資料科學實驗室
  - https://sites.google.com/view/cwwuadslab/

意見交流

歡迎提供意見與指導!!

您的寶貴意見將使本系更進步!!