國立雲林科技大學

資訊管理所

資料探勘專案作業四

比較關聯規則模型於交易商品關聯性之研究

指導教授：許中川教授

學生：工業工程與管理學系 碩一

M11121019 劉驊維

M11121027 洪源廷

M11121036 杜佳容

M11121047 黃惠妤

摘要

在電商高速發展的時代,每個店家都希望可以提供消費者更好的購物體驗，而最直接瞭解客戶的方式就是追蹤其在店家內的消費紀錄，透過歷史消費數據能夠使商家掌握每一位顧客的需求與喜好，商家也可以透過瞭解顧客的消費習慣，調整不同的銷售模式。本文透過比較 Apriori 演算法及 FP-Growth 演算法，並比較分析所花費時間，研究結果顯示得出的關聯分析結果為一致，但 Apriori 演算法所使用的時間會較短，因此選擇 Apriori 演算法會較佳，但局限於處理過後的商品類型資料。

一、緒論

1.1 動機

現今市場上的商品五花八門，商家除了想提供完善的服務體驗給消費者，更想極大化自己的銷售利潤，不管是實體店家的物品陳列還是電子商務的推薦商品，都是零售商重視的議題，因此本文透過關聯分析，分析商品的銷售紀錄，並將其應用至「購物籃分析」，此外，比較兩種不同演算法，分別為 Apriori 演算法及 FP-Growth 演算法進而找出最適合的方式。

1.2 目的

為了使顧客可以擁有更好的購物體驗，因此本文針對其購買的商品類型資料進行關聯分析，目標是在大量的購物記錄清單中，找出不同商品類型與商品類型間間可能存在的關係。本研究分別使用 Apriori 演算法及 FP-Growth 演算法，並比較兩者所花費的時間，並利用支持度與信賴度衡量商品類型項目之間的關聯性，找出績效較高的演算法進行其實務運用。

二、資料集

本研究資料集名稱為交易資料集，包含 7 個欄位總共 157396 筆資料，記錄了交易的相關資訊，資料屬性如表 2.1；資料部分內容表 2.2、表 2.3。

表 2.1 交易資料集資料屬性

屬性名稱	屬性說明	屬性類別
ITEM_ID	商品編號	Nominal
PRODUCT_TYPE	商品類型	Nominal
CUST_ID	客戶編號	Nominal
TRX_DATE	出貨日期	Interval Scale
INVOICE_NO	發票編號	Nominal
QUANTITY	數量	Ratio Scale

表 2.2 資料部分內容

ITEM_ID	ITEM_NO	PRODUCT_TYPE
3217532	M25P40-VMN6TPB	MEMORY_EMBEDED
3326781	AU80610006237AASLBX9	CPU / MPU
740487	MMBD2837LT1G	DISCRETE

表 2.3 資料部分內容

CUST_ID	TRX_DATE	INVOICE_NO	QUANTITY
3218	2016/7/26	CX47348203	2500
2470	2016/7/11	CX47346522	50
16135	2016/7/27	CX47348534	3000

三、方法

3.1 實作說明

本研究使用 Apriori 演算法、FP-Growth 演算法，研究流程如圖 3.1.1 所示。首先在 Excel 上進行資料刪除和訂單重複，接下來在 Python 進行資料轉換，再將轉換的資料投入演算法模型後進行績效評估和輸出結果。

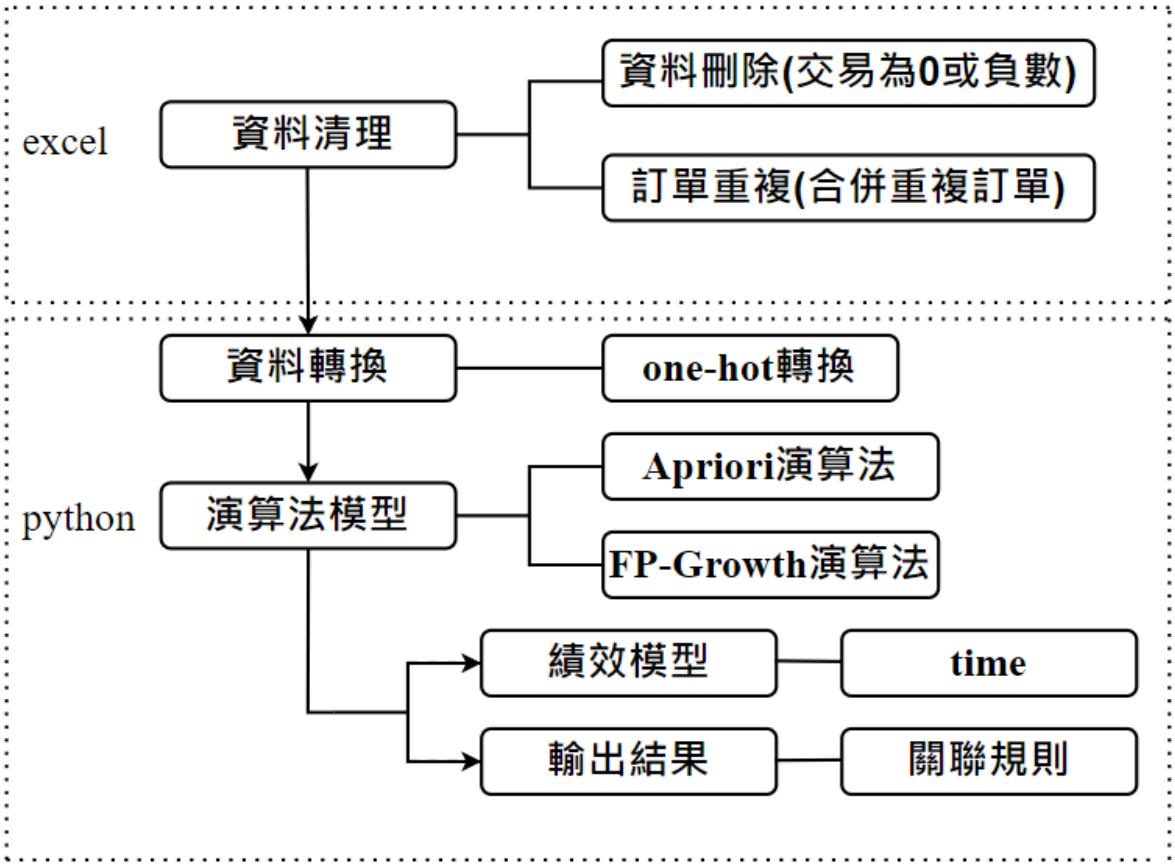

圖 3.1.1 研究流程

3.2 操作說明

3.2.1 安裝套件:

必須在環境裡安裝套件

1. Pandas
2. Numpy
3. Mlxtend
4. Time

3.2.2 程式結果:

Excel 和模型皆必須與程式在同一個資料夾。運行 Fpgrowth 算法的結果如圖 3.2.2.1 所示，運行 Apriori 演算法的結果如圖 3.2.2.2 所示，比較 2 個演算法的運行時間如圖 3.2.2.3 所示，程式結束後如圖 3.2.2.4 所示若想要找尋該商品類型組合的推薦商品類型可查看上表商品類型組合進行查詢。

```
fpgrowth                                antecedents                consequents  \
111  (OPTICAL AND SENSOR, LINEAR IC, LOGIC IC)                   (DISCRETE)
136               (PEMCO, LINEAR IC, LOGIC IC)                   (DISCRETE)
104             (OPTICAL AND SENSOR, LOGIC IC)                   (DISCRETE)
127                        (PEMCO, LOGIC IC)                   (DISCRETE)
36             (DISCRETE, LOGIC IC, CPU / MPU)                  (LINEAR IC)
..                                        ...                         ...
132                              (LINEAR IC)            (PEMCO, LOGIC IC)
150                               (DISCRETE)    (CHIPSET / ASP, LINEAR IC)
151                              (LINEAR IC)     (DISCRETE, CHIPSET / ASP)
143                               (DISCRETE)  (PEMCO, LINEAR IC, LOGIC IC)
144                              (LINEAR IC)   (PEMCO, DISCRETE, LOGIC IC)

     antecedent support  consequent support   support  confidence      lift  \
111            0.001534            0.257905  0.001397    0.910714  3.531207
136            0.001123            0.257905  0.001014    0.902439  3.499120
104            0.002438            0.257905  0.002192    0.898876  3.485307
127            0.001726            0.257905  0.001425    0.825397  3.200397
36             0.002082            0.260398  0.001699    0.815789  3.132858
..                  ...                 ...       ...         ...       ...
132            0.260398            0.001726  0.001123    0.004314  2.499228
150            0.257905            0.002740  0.001069    0.004143  1.512187
151            0.260398            0.001480  0.001069    0.004104  2.773534
143            0.257905            0.001123  0.001014    0.003931  3.499120
144            0.260398            0.001425  0.001014    0.003893  2.732505
```

圖 3.2.2.1 FP-Growth 演算法

```
                                antecedents                consequents  \
119  (OPTICAL AND SENSOR, LINEAR IC, LOGIC IC)                   (DISCRETE)
146               (PEMCO, LINEAR IC, LOGIC IC)                   (DISCRETE)
58              (OPTICAL AND SENSOR, LOGIC IC)                   (DISCRETE)
67                         (PEMCO, LOGIC IC)                   (DISCRETE)
90             (DISCRETE, LOGIC IC, CPU / MPU)                  (LINEAR IC)
..                                        ...                         ...
86                               (LINEAR IC)            (PEMCO, LOGIC IC)
10                                (DISCRETE)    (CHIPSET / ASP, LINEAR IC)
11                               (LINEAR IC)     (DISCRETE, CHIPSET / ASP)
153                               (DISCRETE)  (PEMCO, LINEAR IC, LOGIC IC)
154                              (LINEAR IC)   (PEMCO, DISCRETE, LOGIC IC)

     antecedent support  consequent support   support  confidence      lift  \
119            0.001534            0.257905  0.001397    0.910714  3.531207
146            0.001123            0.257905  0.001014    0.902439  3.499120
58             0.002438            0.257905  0.002192    0.898876  3.485307
67             0.001726            0.257905  0.001425    0.825397  3.200397
90             0.002082            0.260398  0.001699    0.815789  3.132858
..                  ...                 ...       ...         ...       ...
86             0.260398            0.001726  0.001123    0.004314  2.499228
10             0.257905            0.002740  0.001069    0.004143  1.512187
11             0.260398            0.001480  0.001069    0.004104  2.773534
153            0.257905            0.001123  0.001014    0.003931  3.499120
154            0.260398            0.001425  0.001014    0.003893  2.732505
```

圖 3.2.2.2 Apriori 演算法

```
      fpg  aprioi
0.067009   0.031
```

圖 3.2.2.3 比較運算時間

```
商品清單：
117    (LINEAR IC, LOGIC IC, OPTICAL AND SENSOR)
144                 (LINEAR IC, LOGIC IC, PEMCO)
57               (LOGIC IC, OPTICAL AND SENSOR)
66                             (PEMCO, LOGIC IC)
91              (CPU / MPU, LOGIC IC, DISCRETE)
                        ...
87                                  (LINEAR IC)
11                                   (DISCRETE)
10                                  (LINEAR IC)
155                                  (DISCRETE)
153                                 (LINEAR IC)
Name: antecedents, Length: 156, dtype: object

請輸入上表商品類型組合：
```

圖 3.2.2.4 查尋推薦商品類型

四、實驗

4.1 前置處理

- 篩選資料：本研究選擇“PRODUCT_TYPE”，以商品類型為主，找尋此商品的關聯規則。
- 資料清理：
    1. 因有退貨或銷毀商品的情形，導致交易數量為 0 與負數，進而刪除這類筆數。
    2. 因交易訂單含有重複的狀況，所以將重複的訂單合併一塊。
- 資料轉換：運用 TransactionEncoder 模組，將資料轉換成 one-hot 形式。

4.2 實驗設計

資料前處理過後，先經由統計交易次數較高的商品類型作為重點研究，而再提出兩種關聯規則的模型，分別為 Apriori Algorithm 與 Fp-growth Algorithm，設定最小支持度的參數為 1%、0.5%以及 0.1%，並設計三種實驗評估，個別為 Apriori 的關聯法則、Fp-growth 的關聯法則以及兩者訓練時間比較，進而查看得出商品類型之間支持度與信心水準的結果是否相同以及訓練模型時間的績效指標，評估兩種各個在資料集的關聯規則結果，最後，選取該資料集較適用的關聯規則模型。

4.3 實驗結果

由表 4.3.1 表格為交易次數前 10 的商品類型，進而得知 LINEAR IC 的類型交易次數最高達到 9504 次，而從圖(4.3.1)可知在 LOGIC IC 交易次數有明顯的下降趨勢，與前一名已相差 4977 次交易次數。

表 4.3.1 交易次數前 10 的商品類型

Index	items	incident_count
0	LINEAR IC	9504
1	DISCRETE	9413
2	LOGIC IC	4436
3	PEMCO	4361
4	OTHERS	4139
5	CPU / MPU	4042
6	MEMORY_EMBEDED	2435
7	CHIPSET / ASP	1610
8	OPTICAL AND SENSOR	1455
9	MEMORY_SYSTEM	775

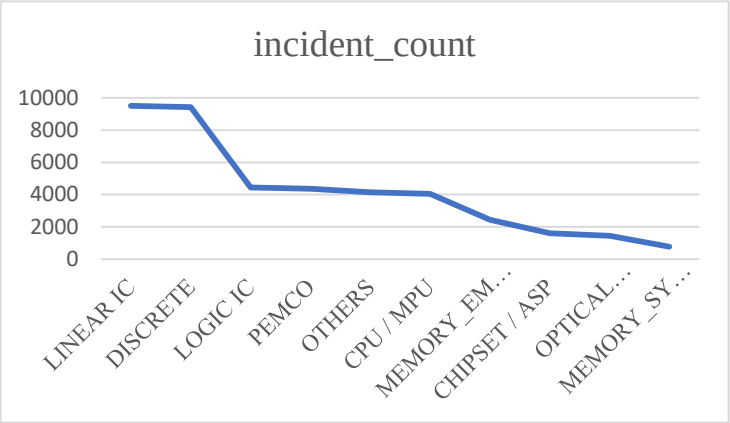

圖 4.3.1

透過上述結果來看得知 LINEAR IC 與 DISCRETE 為交易次數較重要的商品類型，所以此研究將著重在輸入兩種商品類型的關聯規則。

4.3.1 實驗一_ Apriori Algorithm

表 4.3.1.1 為 Apriori Algorithm 在最小支持度=0.01，所整理出的關聯規則結果，由此發現當購買 DISCRETE 與 LOGIC IC 所推薦的商品信心水準高達 55%，支持度 1%；另外，支持度最高為單一交易 LINEAR IC 所推薦 LOGIC IC，信心水準達 17%，支持度達到 4%。

表 4.3.1.1 Apriori_support=0.01

商品類型	推薦商品類型	支持度	信心水準
({'DISCRETE', 'LOGIC IC'})	({'LINEAR IC'})	1%	55%
({'DISCRETE', 'LINEAR IC'})	({'LOGIC IC'})	1%	42%
({'LINEAR IC', 'LOGIC IC'})	({'DISCRETE'})	1%	34%
({'LINEAR IC'})	({'LOGIC IC'})	4%	17%
({'DISCRETE'})	({'LINEAR IC', 'LOGIC IC'})	1%	6%
({'LINEAR IC'})	({'DISCRETE', 'LOGIC IC'})	1%	6%

表 4.3.1.2 為 Apriori Algorithm 在最小支持度=0.005，所整理出的關聯規則結果，結果發現與最小支持度=0.01 的關聯結果相同，由此可知，調整至 0.5%時，並無增加商品的關聯規則。

表 4.3.1.2 Apriori_support=0.005

商品類型	推薦商品類型	支持度	信心水準
({'DISCRETE', 'LOGIC IC'})	({'LINEAR IC'})	1%	55%
({'DISCRETE', 'LINEAR IC'})	({'LOGIC IC'})	1%	42%
({'LINEAR IC', 'LOGIC IC'})	({'DISCRETE'})	1%	34%
({'LINEAR IC'})	({'LOGIC IC'})	4%	17%
({'DISCRETE'})	({'LINEAR IC', 'LOGIC IC'})	1%	6%
({'LINEAR IC'})	({'DISCRETE', 'LOGIC IC'})	1%	6%

表 4.3.1.3 為 Apriori Algorithm 在最小支持度=0.001，所整理出的關聯規則結果，因整理的數據眾多，所以只挑選信心水準較高的與支持度較高的資料，由此發現當購買 LINEAR IC、LOGIC IC 與 OPTICAL AND SENSOR 所推薦的 DISCRETE 商品類型信心水準高達 91%，支持度 0.1%；另外，支持度最高為單一交易 LINEAR IC 所推薦 LOGIC IC，信心水準達 17%，支持度達到 4%。

表 4.3.1.3 Apriori_support=0.001

商品類型	推薦商品類型	支持度	信心水準
({'LINEAR IC', 'LOGIC IC', 'OPTICAL AND SENSOR'})	({'DISCRETE'})	0.1%	91%
({'LINEAR IC', 'LOGIC IC', 'PEMCO'})	({'DISCRETE'})	0.1%	90%
({'DISCRETE', 'CPU / MPU', 'LOGIC IC'})	({'LINEAR IC'})	0.2%	82%
({'DISCRETE', 'LOGIC IC', 'OTHERS'})	({'LINEAR IC'})	0.2%	80%
({'LINEAR IC', 'LOGIC IC', 'OTHERS'})	({'DISCRETE'})	0.2%	78%
({'LINEAR IC'})	({'LOGIC IC'})	4.3%	17%
({'DISCRETE', 'LOGIC IC'})	({'LINEAR IC'})	1.5%	55%

由上述 Apriori 的實驗，可查看當支持度調整至 0.1%時所歸納的關聯規則結果數量增加許多，而其中，單一交易 LINEAR IC 所推薦 LOGIC IC 在此三種實驗支持度皆為最高。

4.3.2 實驗二_ Fp-growth Algorithm

表 4.3.2.1 為 Fp-growth Algorithm 在最小支持度=0.01，所整理出的關聯規則結果，由此對照發現與 Apriori 所找出的關聯規則結果一致。

表 4.3.2.1 Fp-growth _support=0.01

商品類型	推薦商品類型	支持度	信心水準
({'DISCRETE', 'LOGIC IC'})	({'LINEAR IC'})	1%	55%
({'DISCRETE', 'LINEAR IC'})	({'LOGIC IC'})	1%	42%
({'LINEAR IC', 'LOGIC IC'})	({'DISCRETE'})	1%	34%
({'LINEAR IC'})	({'LOGIC IC'})	4%	17%
({'DISCRETE'})	({'LINEAR IC', 'LOGIC IC'})	1%	6%
({'LINEAR IC'})	({'DISCRETE', 'LOGIC IC'})	1%	6%

表 4.3.2.2 為 Fp-growth Algorithm 在最小支持度=0.005，所整理出的關聯規則結果，由此對照發現與 Apriori 所找出的關聯規則結果一致。

表 4.3.2.2 Fp-growth _support=0.005

商品類型	推薦商品類型	支持度	信心水準
({'DISCRETE', 'LOGIC IC'})	({'LINEAR IC'})	1%	55%
({'DISCRETE', 'LINEAR IC'})	({'LOGIC IC'})	1%	42%
({'LINEAR IC', 'LOGIC IC'})	({'DISCRETE'})	1%	34%
({'LINEAR IC'})	({'LOGIC IC'})	4%	17%
({'DISCRETE'})	({'LINEAR IC', 'LOGIC IC'})	1%	6%
({'LINEAR IC'})	({'DISCRETE', 'LOGIC IC'})	1%	6%

表 4.3.2.3 為 Fp-growth Algorithm 在最小支持度=0.001，所整理出的關聯規則結果，由此對照發現與 Apriori 所找出的關聯規則結果一致。

表 4.3.2.3 Fp-growth _support=0.001

商品類型	推薦商品類型	支持度	信心水準
({'LINEAR IC', 'LOGIC IC', 'OPTICAL AND SENSOR'})	({'DISCRETE'})	0.1%	91%
({'LINEAR IC', 'LOGIC IC', 'PEMCO'})	({'DISCRETE'})	0.1%	90%
({'DISCRETE', 'CPU / MPU', 'LOGIC IC'})	({'LINEAR IC'})	0.2%	82%
({'DISCRETE', 'LOGIC IC', 'OTHERS'})	({'LINEAR IC'})	0.2%	80%
({'LINEAR IC', 'LOGIC IC', 'OTHERS'})	({'DISCRETE'})	0.2%	78%
({'LINEAR IC'})	({'LOGIC IC'})	4.3%	17%
({'DISCRETE', 'LOGIC IC'})	({'LINEAR IC'})	1.5%	55%

4.3.3 實驗三_ 花費時間

表 4.3.3.1 為 Apriori 與 Fp-growth 在尋找關聯規則所花費的時間，可發現雖然秒數皆花費不到 1 秒，但 Fp-growth 花費時間比 Apriori 花費時間較高，而在過往的文獻中表示，理論上 Fp-growth 花費時間應比 Apriori 較低，所以本研究會造成 Apriori 花費較少的結果，評估原因為使用商品類型較少，而 Fp-growth 在資料種類多的架構下才會有表現突出的情況，以至於此研究結果使 Apriori 花費時間較少。

表 4.3.3.1 花費時間

	Time(s)	
min_support	Apriori	Fp-growth
1%	0.0319	0.2145
0.5%	0.0319	0.1655
0.1%	0.0589	0.1875

綜括實驗一、實驗二以及實驗三，可發現 Fp-growth 與 Apriori 所得出的關聯規則皆一致，而 Apriori 所花費時間較少，所以在結果相同下，花費時間比較少的情況下，選擇 Apriori 為較佳的決定，但僅侷限此研究所資料處理過後的資料，若資料處理不一致，則無法得出相同結果。

五、結論

本研究使用 Apriori 與 Fp-growth 針對交易資料集進行比較分析，不管是利用 Apriori 還是 Fp-growth 在支持度 0.01 以及 0.005 時，結果皆顯示 DISCRETE 與 LOGIC IC 所推薦商品類別為 LINEAR IC 且信心水準最高，而支持度= 0.005 時信心水準最高的皆為 LINEAR IC、LOGIC IC 推薦商品類別是 OPTICAL AND SENSORDISCRETE，以上結果可提供商家依照這些具有較高關聯的類別陳列商品，以利於提供消費者更佳的購物體驗且提高銷售利潤。

參考文獻

1. *https://artsdatascience.wordpress.com/2019/12/10/python-%E5%AF%A6%E6%88%B0%E7%AF%87%EF%BC%9Aapriori-algorithm/*
2. *https://medium.com/@tinahuang_4101/associating-rule-%E9%97%9C%E8%81%AF%E5%BC%8F%E6%B3%95%E5%89%87%E6%87%89%E7%94%A8-apriori-fp-growth-3ab46deeeb77*
3. *https://www.twblogs.net/a/5b7fbf712b717767c6b167d6*
4. *https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/*