

國立雲林科技大學

資訊管理系

資料探勘

作業四

關聯分析實作

D10630002 陳冠臻

M11023006 謝媛衣

M11123044 沈俊良

M11123061 王成綱

指導教授:許中川

2 0 2 2 年 1 月 2 日

## 摘要

本研究採用 trade 資料集測試 support、confidence 及 lift，使用 MS SQL Server 進行資料清洗、正規化資料，透過 Python 訓練演算法(Apriori 及 FP-Growth)、迴圈找出 trade 資料集之最佳支持度及關連性，並畫出各關聯圖及折線圖，並比較 trade 資料集在各演算法間哪項最佳，最後資料集支持度及關連性最高之屬性為 DISCRETE。

關鍵字: MS SQL Server、Apriori、FP-Growth、support、confidence、lift

## 一、緒論

### 1.1 動機

面對已佔有龐大市場的大型電商，中小型電商不必妄自菲薄，掌握「個人化行銷」：了解客戶需求、透過服務創造差異化，一樣能贏得顧客的心。大型電商如亞馬遜，由於規模大、商品數量多，無法細緻地提供顧客購買建議及相關服務，因此，細緻度成為中小型電商切入的關鍵。據美國行銷平台 SmarterHQ 調查，在亞馬遜購物的用戶，57%都是為了購買特定物品而來、其中 63%更事先決定好要購買的商品才來購物。BI Intelligence 的市場研究員 Daniel Keyes 解釋，這代表多數人是因亞馬遜規模大、商品數量多而來購物，並非因為亞馬遜是最好「逛」或「發掘新商品」的購物網站。若是其他中小型電商能注重顧客體驗及需求，提供完整的服務，便能在此處創造差異化，替品牌創造客戶及支持者。「個人化行銷」與多數人熟悉的「客製化行銷」不同，更注重體驗。「客製化行銷」將消費者視為一個「群體」，由廠商先分析消費者的喜好、依結果設計產品功能和服務後，再讓客戶選擇最適合自己的項目；「個人化行銷」則以「個人」的角度分析消費者，廠商會依據消費者個人過往的資料及消費習慣分析，主動提出客戶會最感興趣、或最符合客戶需求的內容。因此本研究希望能提升國內中小型電商的銷售，因此著眼於行銷方式上，首先會先與傳統零售購買時定時定點是的行銷策略做區隔，同時也不與大型電商提供成千上萬件眼花撩亂的商品方式相同，反而是發展出一套選擇的機制，能貼近消費者本身行銷方式，而這也是本次研究的首要目標。

## 1.2 目的

傳統大量行銷策略逐漸轉向顧客導向的行銷方向，與顧客進行更直接、互動關係更頻繁方式進行溝通。而多樣化與個人化的溝通方式，可使個別的消費者發展長期互惠的關係，來達到客戶滿意度的提升和創造更多顧客價值，而非僅發展成買與賣的單項式交易關係。個人化行銷的關鍵是「優良的體驗」，以下為個人化消費體驗的方式：一、以產品為基礎的個人化行銷分析消費者的行為模式，不但不用費時去蒐集每個用戶較獨立的個人資料，更有過去經驗作為佐證。倘若消費者不因這次看到下方的推薦欄位而將其商品放進購物車中掏錢購買，至少對商品來說也有多一次的曝光機會，當下一次消費者再進到電商網站時，說不定便會形成轉換。二、以消費者為基礎的個人化行銷，以消費者為基礎的個人化行銷中，可細分消費者為首次進入電商網站的潛在消費者與過去曾造訪過甚至是有購買紀錄的既有消費者。三、即時性個人化行銷，善加利用即時性的資訊個人化行銷，可以提供消費者耳目一新的新鮮感，例如：某品牌服飾官網會依據造訪者所在的地點提供天氣資訊，並且同時推薦適合該天氣所穿著的服飾。因此我們可以根據上述個人化行銷原則，發展出體驗的消費方式，因此本研究希望可以推出一種個人化的優惠券，可根據消費者的消費習慣與購買過的物品紀錄，來形塑他的未來可能的消費樣態，給予他消費建議清單，並可以搭配即時特價的服務，刺激當下購買的衝動，更細緻且不讓人有衝動消費後的罪惡感，提供消費者有良好的消費經驗和體驗。

## 二、真實資料集

### 2.1 資料集

表 1

*Trade Dataset Data Set* 資料屬性一覽

欄位名稱	欄位意義	型態
ITEM_ID	商品編號	Integer
ITEM_NO	商品型號	Char
PRODUCT_TYPE	商品名稱	Char
CUST_ID	銷售 ID	Integer
TRX_DATE	銷售日期	Date
INVOICE_NO	發票編號	Integer
QUANTITY	銷售數量	Integer
樣本數	屬性個數	
157397	7	

## 三、方法

### 3.1 實作說明

將 Trade Dataset Data Set 使用 Microsoft SQL Server Management Studio 匯入至 Microsoft SQL Server 資料表中，透過 T-SQL 語法進行正規化處理，並匯出成 csv 檔。

使用 Python 作為開發程式語言，透過 Apriori、FP-Growth 演算法套件，對資料集進行關聯分析，並計算支持度與執行時間分析各演算法之最佳參數值，再依個別演算法之支持度及信心度比較關聯分析結果。

### 3.2 操作說明

可自行輸入 Apriori 關聯演算法程式之支持度、信心度及輸入商品類行為何，找出輸入之商品類型所最推薦商品，並記錄各組關聯分析之執行時間、信心度及支持度。

可自行輸入 FP-Growth 關聯演算法程式之支持度、信心度及輸入商品類行為何，找出輸入之商品類型所最推薦商品，並記錄各組關聯分析之執行時間、信心度及支持度。

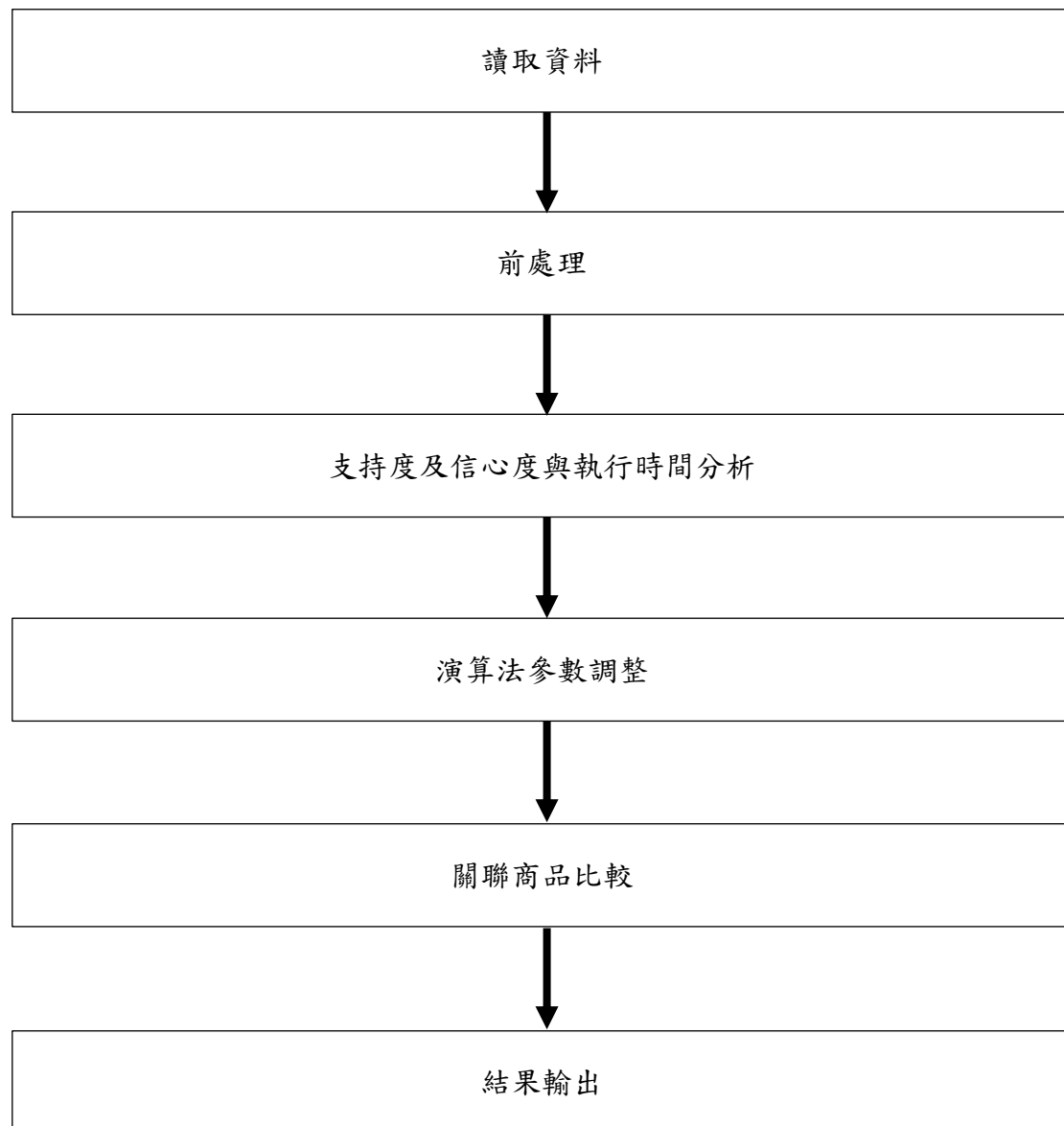


圖 1 資料集關聯分析流程圖

## 四、實驗

### 4.1 前置處理

對於 Trade Dataset Data Set 之名目尺度特徵屬性進行剔除零或負值，以及將交易資料集中相同 INVOICE\_NO 表示為同一筆交易紀錄。

ITEM ID	ITEM NO	PRODUCT TYPE	CUST ID	TRX DATE	INVOICE NO	QUANTITY
3217532	M25P40-VMN6TPB	MEMORY EMBEDDED	3218	2016/7/26	CX47348203	2500
3326781	AU80610006237AASLBX9	CPU / MPU	2470	2016/7/11	CX47346522	50
740487	MMBD2837LT1G	DISCRETE	16135	2016/7/27	CX47348534	3000
3434776	IHLPI616ABER2R2M11	PEMCO	999999999	2016/7/29	A20160700174	0
70072	MMBT3906LT1G	DISCRETE	2356	2016/7/6	CX47346184	12000
3204503	PCA9555DWR	LOGIC IC	2506	2016/7/21	A10085337	0
3420352	TMP103AYFFR	LINEAR IC	10228	2016/7/25	CX47347899	3000
3311565	OV6922-V09N	OPTICAL AND SENSOR	38381	2016/7/6	CX47346191	1152
140887	SN74AHC1G126DCKR	LOGIC IC	999999999	2016/7/31		5119
3216410	SI2303CDS-T1-GE3	DISCRETE	27495	2016/7/11	CX47346636	3000
123164	TLZ24B-GS08	DISCRETE	30377	2016/7/28	CX47348680	2500
14380729	CC2530F256RHAR	LINEAR IC	999999999	2016/7/31		5119
14831058	QCA9550-AT4A	CHIPSET / ASP	16686	2016/7/1	CX47345496	651
148734	SN74LVC1G04DBVR	LOGIC IC	2506	2016/7/28	CX47348656	3000
88067	SN74AHC32PWR	LOGIC IC	38430	2016/7/22	CX47347745	50000
3315806	UCC2813PWTR-5	LINEAR IC	3736	2016/7/26	CX47348078	2000
2901183	CAT24C08WLTG3	MEMORY SYSTEM	16135	2016/7/13	CX47346899	9000
14677766	FH8065501516762SR1S9	CPU / MPU	2506	2016/7/15	CX47347172	14
15145043	RF2946-000	PEMCO	2454	2016/7/14	CX47347047	4000
3434915	LM75BIMM-3/NOPB	LINEAR IC	53917	2016/7/6	CX47346157	3000
2442248	ESD5B5.0ST1G	DISCRETE	2494	2016/7/4	CX47345910	15000
3326173	BAV998.215	DISCRETE	16135	2016/7/6	CX47346143	9000
135739	1PS79SB30.115	DISCRETE	35211	2016/7/28	CX47348630	12000
15041950	VJ0805Y102KXABP31	PEMCO	56711	2016/7/1	CX47345591	10000
3252971	CM1213A-02SR	DISCRETE	2470	2016/7/11	CX47346492	3000
149830	SN74LVC1G17DBVR	LOGIC IC	2549	2016/7/26	CX47348394	3000
15006948	CM8063401286503SR1AJ	CPU / MPU	3447	2016/7/1	CX47345430	1
8929651	WG1210T.SLJXT	CHIPSET / ASP	9552	2016/7/1	CX47345554	1350
2219803	NTLJD3115PT1G	DISCRETE	2767244	2016/7/4	CX47345741	3000

圖 2 Trade Dataset Data Set 原始資料

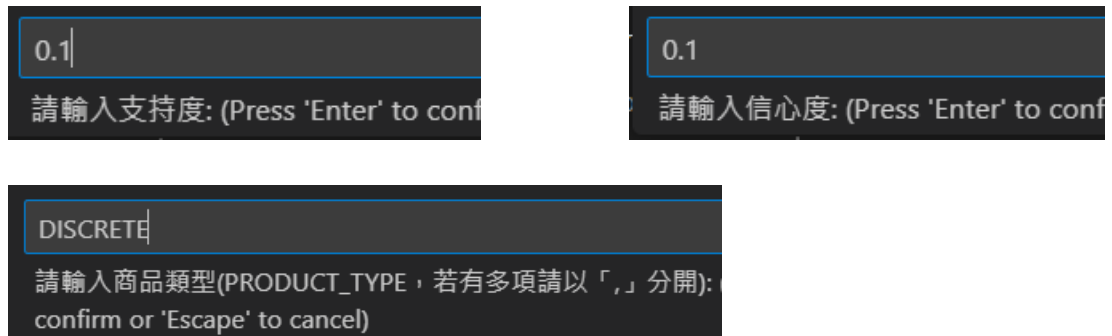
LINEAR IC	LOGIC IC	MEMORY EMBEDDED			
CPU / MPU	DISCRETE	LINEAR IC	PEMCO		
CHIPSET / ASP	CPU / MPU	LINEAR IC			
DISCRETE	LINEAR IC	OPTICAL AND SENSOR			
CHIPSET / ASP	DISCRETE	LINEAR IC	MEMORY SYSTEM		
CPU / MPU	LOGIC IC	MEMORY EMBEDDED			
OTHERS	PEMCO				
DISCRETE	MEMORY SYSTEM				
PEMCO					
LINEAR IC	MEMORY EMBEDDED	OTHERS			
DISCRETE	LINEAR IC	LOGIC IC	MEMORY EMBEDDED	MEMORY SYSTEM	
CHIPSET / ASP	DISCRETE	LINEAR IC	LOGIC IC	OPTICAL AND SENSOR	OTHERS
CPU / MPU	DISCRETE				
CHIPSET / ASP	LOGIC IC	MEMORY EMBEDDED			
DISCRETE	LOGIC IC	OTHERS			
LOGIC IC	OPTICAL AND SENSOR				
DISCRETE	LINEAR IC	LOGIC IC	MEMORY EMBEDDED		
CHIPSET / ASP	MEMORY SYSTEM				
DISCRETE	LINEAR IC	MEMORY EMBEDDED	PEMCO		
CPU / MPU	LINEAR IC	LOGIC IC	MEMORY EMBEDDED		
LINEAR IC	MEMORY SYSTEM				
LOGIC IC	OPTICAL AND SENSOR	OTHERS			
CPU / MPU	MEMORY SYSTEM				
OPTICAL AND SENSOR					
MEMORY SYSTEM					
CPU / MPU	LINEAR IC	OPTICAL AND SENSOR			
CHIPSET / ASP	DISCRETE	LINEAR IC	OTHERS		
DISCRETE	LOGIC IC	OPTICAL AND SENSOR	OTHERS		
CHIPSET / ASP	CPU / MPU	DISCRETE			
OTHERS					

圖 3 Trade Dataset Data Set 正規化後資料



## 4.2 實驗設計

在 Apriori 演算法下，自行輸入支持度、信心度及商品，則顯示當支持度為 0.1、信心度為 0.1 及輸入商品類型為 DISCRETE 時顯示推薦商品為 OPTICAL AND SENSOR、LINEAR IC 及 LOGIC IC。



0.1  
請輸入支持度: (Press 'Enter' to confirm)

0.1  
請輸入信心度: (Press 'Enter' to confirm)

DISCRETE  
請輸入商品類型(PRODUCT\_TYPE, 若有多項請以「,」分開):  
confirm or 'Escape' to cancel)

圖 4 Trade Data Set 之輸入支持度、信心度及商品類型圖



```
您輸入的支持度為:0.1  
您輸入的信心度為:0.1  
您輸入的商品類型為:DISCRETE  
=====  
為您推薦以下商品類型:  
['OPTICAL AND SENSOR', 'LINEAR IC', 'LOGIC IC']
```

圖 5 Trade Data Set 之輸入支持度、信心度及商品類型圖結果圖

針對 Support、Confidence 及 Lift 進行交叉分析及比較各支持度及信心度在每個值之間分佈

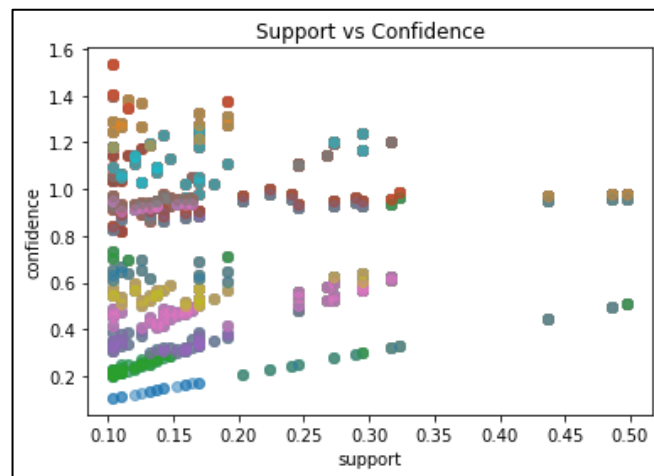


圖 6 Trade Data Set 之 Support vs Confidence 之比較

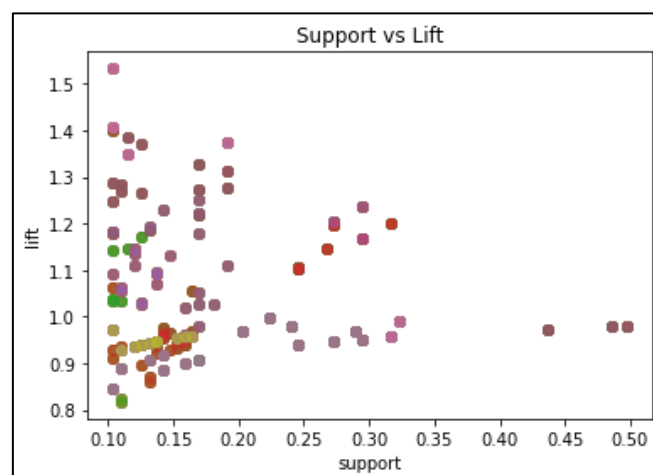


圖 7 Trade Data Set 之 Support vs Lift 之比較

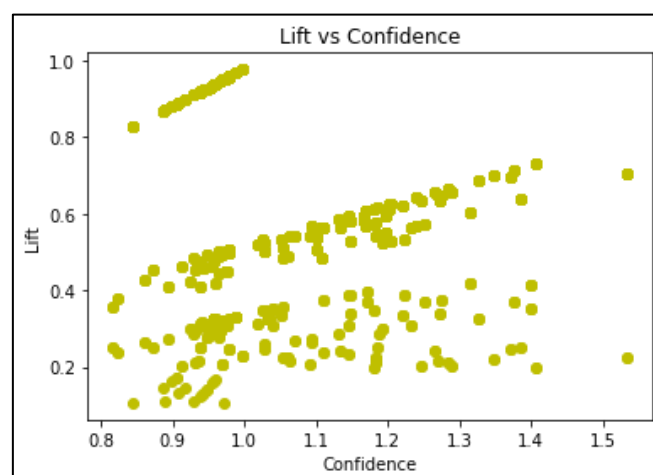


圖 8 Trade Data Set 之 Lift vs Confidence 之比較

在 FP-Growth 演算法下，自行輸入支持度及商品，則顯示當支持度為 0.1、輸入商品類型為 DISCRETE 時顯示推薦商品為 LINEAR IC、CPU / MPU、PEMCO、OPTICAL AND SENSOR、CHIPSET / ASP、MEMORY\_SYSTEM、LOGIC IC、MEMORY\_EMBEDDED、OTHERS。



圖 9 Trade Data Set FP-Growth 之輸入支持度及商品類型圖

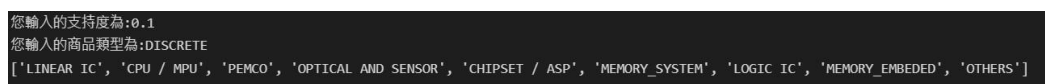


圖 10 Trade Data Set FP-Growth 之輸入支持度及商品類型圖

在 FP-Growth 演算法下，算出最有好的前五名的商品以及利用 treemap 畫出矩形式樹狀結構圖，找出最好的商品

	items	incident_count
0	DISCRETE	95
1	LINEAR IC	93
2	LOGIC IC	84
3	MEMORY_EMBEDDED	62
4	OTHERS	61

圖 11 Trade Data Set 之商品 incident\_count 排名前五圖

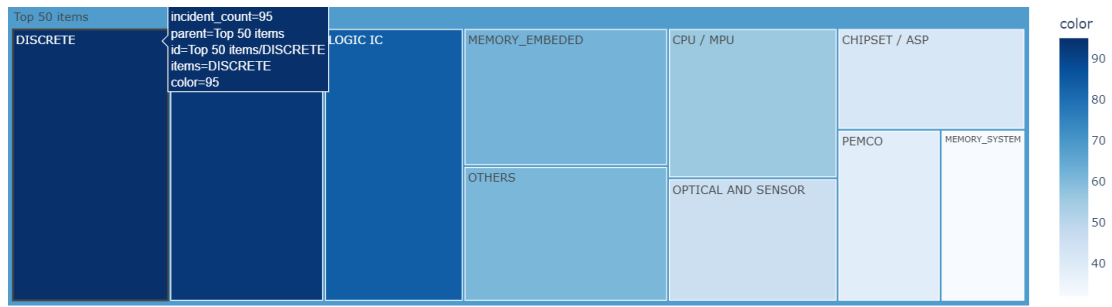


圖 12 Trade Data Set 之矩形式樹狀結構圖-為 incident\_count 排名第一

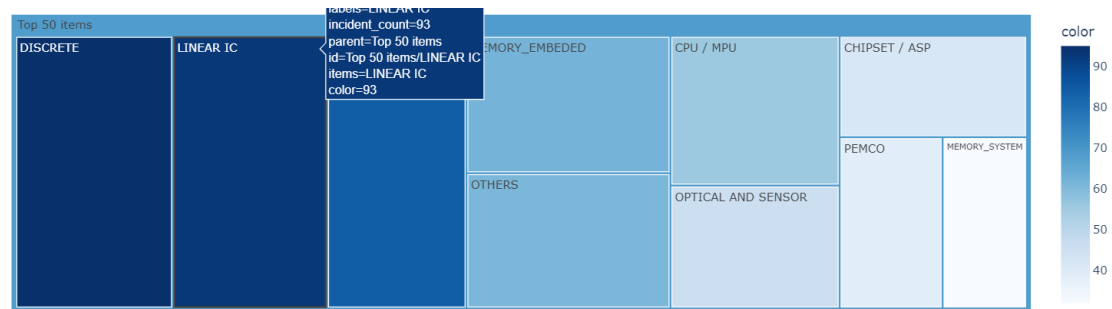


圖 13 Trade Data Set 之矩形式樹狀結構圖-為 incident\_count 排名第二

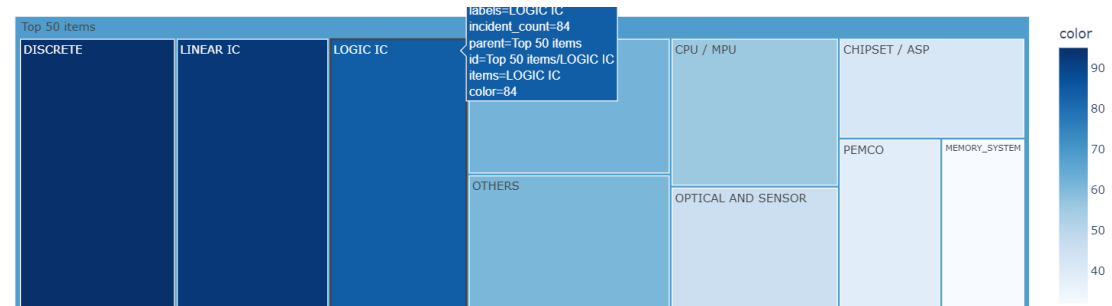


圖 14 Trade Data Set 之矩形式樹狀結構圖-為 incident\_count 排名第三

在 FP-Growth 演算法下，算出 Support、Confidence 及 Lift 之排名前十之數值

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(DISCRETE)	(LINEAR IC)	0.519126	0.508197	0.316940	0.610526	1.201358	0.053122	1.262738
1	(LINEAR IC)	(DISCRETE)	0.508197	0.519126	0.316940	0.623656	1.201358	0.053122	1.277752
2	(LOGIC IC)	(LINEAR IC)	0.459016	0.508197	0.267760	0.583333	1.147849	0.034489	1.180328
3	(LINEAR IC)	(LOGIC IC)	0.508197	0.459016	0.267760	0.526882	1.147849	0.034489	1.143443
4	(DISCRETE)	(LOGIC IC)	0.519126	0.459016	0.295082	0.568421	1.238346	0.056795	1.253499
5	(LOGIC IC)	(DISCRETE)	0.459016	0.519126	0.295082	0.642857	1.238346	0.056795	1.346448
6	(DISCRETE, LOGIC IC)	(LINEAR IC)	0.295082	0.508197	0.191257	0.648148	1.275388	0.041297	1.397757
7	(DISCRETE, LINEAR IC)	(LOGIC IC)	0.316940	0.459016	0.191257	0.603448	1.314655	0.045776	1.364220
8	(LOGIC IC, LINEAR IC)	(DISCRETE)	0.267760	0.519126	0.191257	0.714286	1.375940	0.052256	1.683060
9	(DISCRETE)	(LOGIC IC, LINEAR IC)	0.519126	0.267760	0.191257	0.368421	1.375940	0.052256	1.159381

圖 15 Trade Data Set FP-Growth 之 Support、Confidence 及 Lift 排名圖

## 五、結論

本研究使用 SQL Server 進行前置處理，後利用 python 進行各演算法 (Apriori、FP-Growth) 分析，計算資料集在各演算法間哪項最佳，結果得知 Trade 資料集在 Apriori 及 FP-Growth 演算法下 DISCRETE 商品是最好的結果，且支持度在 Apriori 下最高。

## 六、參考文獻

- chwang12341 (Jul 2020)。github。  
<https://github.com/chwang12341/Machine-Learning/blob/master/Apriori/%E9%97%9C%E8%81%AF%E8%A6%8F%E5%89%87%E5%AF%A6%E6%88%B0.ipynb>
- MAX (Nov 2018)。[關聯分析] Apriori 演算法介紹 (附 Python 程式碼)。  
[https://www.maxlist.xyz/2018/11/03/python\\_apriori/](https://www.maxlist.xyz/2018/11/03/python_apriori/)
- 阿新 (Aug 2019)。Python --深入淺出 Apriori 關聯分析演算法 (二)  
Apriori 關聯規則實戰。  
<https://www.796t.com/content/1566475622.html>
- Yeh James (Oct 2017)。[資料分析&機器學習] 第 2.4 講：資料前處理  
(Missing data, One-hot encoding, Feature Scaling)。  
<https://medium.com/jameslearningnote/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E7%AC%AC24%E8%AC%9B%E8%B3%87%E6%96%99%E5%89%8D%E8%99%95%E7%90%86-missing-data-one-hot-encoding-feature-scaling-3b70a7839b4a>
- Harsh (Sep 2019)。Association Analysis in Python。  
<https://medium.com/analytics-vidhya/association-analysis-in-python-2b955d0180c>
- Is 泰 (Aug 2018)。Python 機器學習 — 關聯規則 (Apriori、FP-growth)。  
<https://www.twblogs.net/a/5b7dd3152b717768385411e0>
- Sebastian Raschka (2022)。fpgrowth: Frequent itemsets via the FP-growth algorithm。  
[http://rasbt.github.io/mlxtend/user\\_guide/frequent\\_patterns/fpgrowth/](http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpgrowth/)
- Bashir Alam (Feb 2022)。Implementation of FP-growth algorithm using Python。  
<https://hands-on.cloud/implementation-of-fp-growth-algorithm-using-python/>