國立雲林科技大學資訊管理系

資料探勘

National Yunlin University of Science and Technology

Department of Information Management

專案作業四

以 Apriori 演算法及 FP-Growth 演算法進行交易資料關聯規則分析

M11123048 陳韋彤

M11123060 毛俐蘋

M11123019 張芸婷

指導老師：許中川

Advisor：Chung-Chain Hsu，Ph.D.

中華民國 111 年 12 月

December 2022

摘要

本研究透過 Apriori 演算法及 FP-Growth 演算法進行資料探勘，對交易資料進行關聯規則分析，目的是在一個資料集當中，找出不同項與項之間可能存在的關係。本研究利用兩種演算法，FP-growth 演算法與 Apriori 演算法兩者進行比較。不同演算法下的花費時間比較中，FP-growth 演算法所需的計算時間較 Apriori 演算法短，此外，當兩個演算法之最小支持度越大，其計算出的關聯規則皆越少，同時計算時間也相對較快；支持度越小時，推薦出的商品數量相對較多。

關鍵字：關聯規則分析、資料探勘、交易資料

一、 緒論

1.1 動機

關聯分析能夠在大型資料庫中尋找變數之間的關係，主要透過「支持度」(Support) 與「信賴度」(Confidence) 對資料間頻繁出現的項目集進行篩選，判斷其中的關聯。本研究選用交易資料集進行關聯分析，交易資料集提供了商品編號、商品型號、產品類別、顧客編號、交易日期、交易編號和數量的數據資料，利用商品編號和交易編號之數據尋找關聯規則，進行分析。

1.2 目的

本次研究使用 Apriori 演算法及 FP-Growth 演算法進行交易資料關聯規則分析，透過關聯規則，找出相關的推薦產品，並比較兩者所花費的時間差異。於交易資料集中，利用商品編號和交易編號項目集中作為判斷依據，了解不同參數值設定與推薦產品數量多寡之關係，並比較兩種演算法所花費的時間。

二、 方法

本研究實驗流程如圖 1 所示。

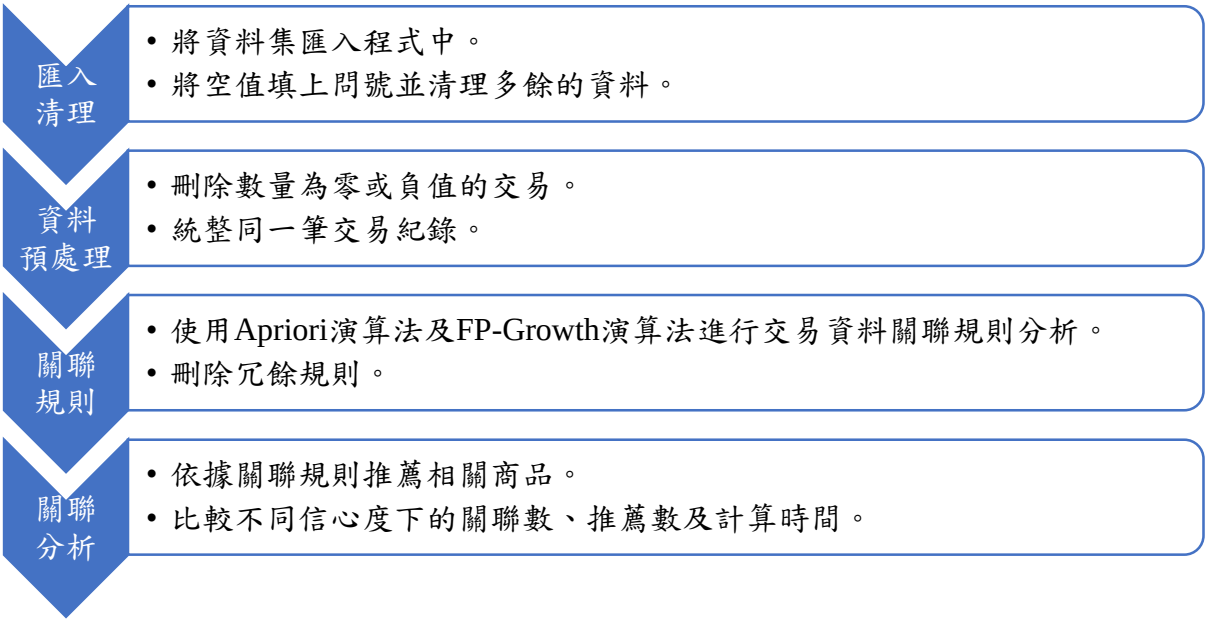

圖 1 實驗方法流程圖

三、 實驗

3.1 資料集

交易資料集包含了商品編號、商品型號、產品類別、顧客編號、交易日期、交易編號和數量的類別。共有 7 種屬性，詳細屬性資訊如表 1 所示。

表 1

交易資料集屬性資料

屬性名稱	詳細資料描述
ITEM_ID	商品編號。
ITEM_NO	商品型號。
PRODUCT_TYPE	產品類別。
CUST_ID	顧客的代表編號。
TRX_DATE	交易訂單成立之日期。
INVOICE_NO	交易編號，相同即為同一筆交易紀錄。
QUANTITY	指交易訂單中的交易數量。

3.2 前置處理

交易資料集中以「?」代替缺失的資料，並將空值刪除。QUANTITY (數量) 為零或為負值，代表該筆交易已退貨或註銷，將之刪除。在 INVOICE_NO (交易編號) 中，相同著編號代表同一筆交易紀錄，將交易數量加總，合併為一筆資料。根據 INVOICE_NO (交易編號) 和 ITEM_NO (商品型號)，列出每筆交易紀錄中，分別購買了哪些商品，有購買以「1」代表，無購買則以「0」代表。

3.3 實驗設計

本研究以 Apriori 演算法及 FP-growth 演算法，對交易資料關聯規則進行分析。

Apriori 演算法在電腦科學以及資料探勘領域中，Apriori 演算法是「關聯規則學習」或是「關聯分析」 (Associative Analysis) 的經典演算法之一，目的是在一個資料集當中，找出不同項與項之間可能存在的關係。

FP-growth 演算法不同於 Apriori 演算法生成頻繁項集再檢查是否頻繁，不斷掃描事物集。而是使用一種稱為頻繁模式樹 (FP-Tree，PF 代表頻繁模式，Frequent Pattern) 選單緊湊資料結構組織資料，並直接從該結構中提取頻繁資料集，不需要產生候選集。每個事物被對映到 FP-tree 的一條路徑上，不同的事物會有相同的路徑，因此重疊的越多，壓縮效果越好。

3.4 實驗結果

3.4.1 交易資料集

交易資料集在 Apriori 演算法下，透過調整最小支持度 (min_support) 計算關聯規則。本研究將分別研究 0.001、0.0008、0.005 三種不同支持度下，規則數量和執行時間之變化，另以商品編號「70509」作為搜尋前因，觀察推薦的商品編號及數量之變化，詳細如表 2 所示。

表 2

Apriori 演算法下各項數據比較表

最小支持度	0.001	0.005	0.0008
數量	142	0	582
時間	0:00:05.292492	0:00:01.703079	0:00:43.848932
輸入商品編號	70509	70509	70509
推薦產品數量	11	0	16
商品編號	1707 70448 70449 87118 88444 88623 88676 135492 135493 135789 3336149	沒有推薦的商品	1696 1707 26001 70126 70128 70274 70448 70449 87120 87945 88675 135494 135495 135505 4709595 15021864

交易資料集在 FP-growth 演算法下，透過調整最小支持度 (min_support) 計算關聯規則。本研究將分別研究 0.001、0.0008、0.005 三種不同支持度下，規則數量和執行時間之變化，另以商品編號「70509」作為搜尋前因，觀察推薦的商品編號及數量之變化，詳細如表 2 所示。

表 3

FP-growth 演算法下各項數據比較表

支持度	0.001	0.005	0.0008
規則數量	142	0	582
執行時間	0:00:02.309820	0:00:01.903905	0:00:02.067515
輸入商品編號	70509	70509	70509
推薦產品數量	8	0	18
推薦商品編號	87118 70449 70448 88623 88676 135492 135789 1707	沒有推薦商品	1696 87945 4709595 26001 70449 70448 135495 70128 70126 87120 15021864 135789 1707 135505 88622 70274 135494 88675

四、 結論

根據結果所示，FP-growth 演算法與 Apriori 演算法兩者相比，以計算時間來看，FP-growth 演算法所需的計算時間較 Apriori 演算法短一些；關聯規則的數量則相同；推薦的商品數量中，Apriori 演算法在支持度較低的情況下，推薦的商品數量相較 FP-growth 演算法多一些，支持度較高的情況下，FP-growth 演算法推薦的商品數量較 Apriori 演算法多。兩個演算法中，支持度越大，計算出的關聯規則越少，同時計算時間也相對較快；支持度越小時，推薦出的商品數量相對較多。

參考文獻

(2018)關聯規則演算法及 FP-growth 的使用和原始碼解析。(通俗易懂)關聯規則演算法及 FP-growth 的使用和原始碼解析 - IT 閱讀 (itread01.com)

(2018) 關聯分析演算法：FP-Growth。https://www.gushiciku.cn/pl/2CPC/zh-tw

(2019) *Apriori Algorithm — Know How to Find Frequent Itemsets* Sayantini Deb https://medium.com/edureka/apriori-algorithm-d7cc648d4f1e

(2018) *apriori: Frequent itemsets via the Apriori algorithm* https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/

(2019). *Estimation of obesity levels based on eating habits and physical condition Data Set*. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+#

(2019). *Iris Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/iris