**Data minding homework1**

資訊管理研究所 林家輝 R76061045

1. **資料蒐集過程以及資料處理**

* **IBM的過程:**

利用IBM Quest Data Generator輸入以下指定

* **KAGGLE的過程:**

因為大學在念MIS時就覺得可以找到尿布啤酒的關聯性，**並以此關聯提升賣場的銷售感受到這就是念資訊的力量**，所以找到Stanislav Semenov

提供的賣場資料(作者沒說明是哪一間，學生猜測可能有商業機密的問

題)，想自己嘗試找出賣場有用的關聯。

1. **程式碼撰寫過程**

* Brute-force algorithm

1. 輸入資料為: IBM Quest Data Generator所輸出的格式，採讀檔(txt)的**方式一次讀入一行**；一行內包含交易編號與商品編號兩個字串屬性。
2. 建立出一個**Hash array 統計各項目出現的次數**，以備之後要進行篩選，淘汰掉不符合support的項目，選出C1。
3. 將C1轉成物件，**new 出Hashset將frequent 1-item存入**，以便後續判斷是否為子集合。
4. 寫出**組合函數程式**並以總frequent 1-item個數設為總數量，建立FOR迴圈依次序將C1、C2、C3…找出並存入perfectset當中。並建立**次數對照表Covert array**，以備後續算confident所用。
5. 將perfectset**倆倆比對是否大集合有包括小集合**，若是則叫出Covert array並算出confident。
6. 輸出資料為: 印出滿足minimum support以及confidence的所有規則。

* **FP growth algorithm**

1. 輸入資料為: IBM Quest Data Generator所輸出的格式，採讀檔(txt)的方式**一次讀入一行**；一行內包含交易編號與商品編號兩個字串屬性。
2. 第一次掃描檔案。**找出frequent 1-item**並根據交易數量與minimum support計算threshold(support count)。之後將frequent 1-item**由頻率高至低排序**，並輔以threshold排除不含標準之項目。
3. 第二次掃描檔案。以一筆交易為單位，將該筆交易內的項目依frequent 1-item排序並刪除低頻者，隨後**建構Header table**與FP-tree。
4. 根據已**建置好的FP-tree尋找Conditional pattern base**，接著以類似1)、2)之步驟建置Conditional FP-tree，並尋找Frequent patterns。
5. **尋找每一個Frequent patterns的潛在規則**(candidate rules)，個別計算其confidence並篩除不滿足標準的規則。
6. 輸出資料為: 印出滿足minimum support以及confidence的所有規則
7. **WEKA資料前處理過程**

* IBM的過程:

將IBM產生的檔案另存成txt檔案，**再透過自己撰寫的java轉檔程式**

將txt檔案轉換成橫向txt檔案，在另存成csv檔案。最後透過IBM

Quest Data Converter的-b指令轉成WEKA可以吃得二維矩陣格式。

* **KAGGLE的過程**:

先將下載KAGGLE直式csv檔案透過自己撰寫的java轉檔程式，將資料轉

成橫向csv檔案，**再透過IBM Quest Data Converter的-b指令轉成WEKA**

**可以吃得二維矩陣格式**。

1. **實驗結果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Brute-force algorithm | FP growth algorithm | WEKA  (Apriori) | WEKA  (fp growth ) |
| 分析正確率 | 正確 | 資料多時會有BUG | 正確 | 正確 |
| 執行時間 | 92790m | 177ms | 不到1秒鐘 | 不到1秒鐘 |

1. **利用Kaggle資料得到的有用關聯**

* **列出最顯著的前面五項關聯**

1. 有機石榴石甘藷(Organic Garnet Sweet Potato) 🡪有機嬰兒菠菜(Organic Baby Spinach)的Confident rate=50%
2. 有機嬰兒芝麻菜(Organic Baby Arugula) 🡪有機嬰兒菠菜(Organic Baby Spinach)的Confident rate=47%
3. Honeycrisp蘋果(Honeycrisp Apple) 🡪香蕉(Banana)的Confident rate=43%
4. 有機意大利荷蘭芹束(Organic Italian Parsley Bunch) 🡪有機大蒜(Organic Garlic)的Confident rate=42%
5. 哈斯鱷梨(Hass Avocados) 🡪有機草莓(Organic Strawberries的Confident rate=42%

* **學生從關聯中揣測獲得的可能訊息**

1. 在這個美國賣場中**買蔬菜或水果都會一次買多樣**->推測美國到超市的距離較遠，一次車程都會買未來一個禮拜要買的食物量。
2. 這個地區大賣場**義大利麵銷售很高**->當地**口味喜歡芹菜搭配大蒜**。
3. **作業遇到的困難與成長**

* **關聯法觀念的理解**

1. 從完全不知道怎麼做關聯法 -> 上完課後**懂得關聯法的四大步驟**。
2. 從原本程式速度非常慢 -> **學會了FP growth後程式速度快了三倍**。

* **資料處理能力**

1. 從原本資料只會單存讀取txt,csv檔案。->現在學會**自己寫程式將不同檔案類型任意做轉換**，且學會用IBM Quest Data Converter將資料轉換成二維稀疏矩陣。

* **程式撰寫能力**

1. 學會在寫較難的程式時，一開始不要直接下手**先透過紙筆擬出架構**再進行撰寫速度會快非常多。
2. 學會將**程式進行邏模組化**，使得在寫程式時可以透過直接配對創造出更多

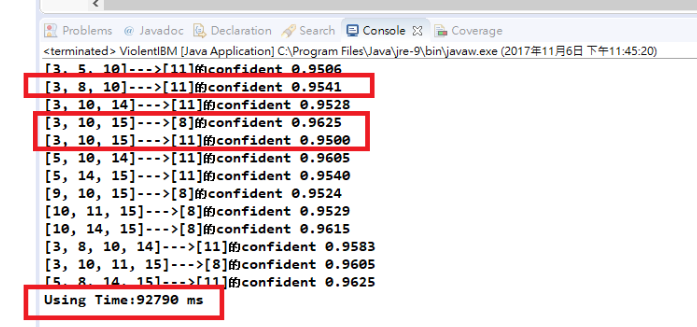
有彈性且不同功能的程式。

1. **學會利用樹的結構**，且再利用樹的結構前若**先進行排序可以提高Search的速度**。
2. **提升debug的能力**，因為學生在**這項作業的努力時間至少100個小時以上**，學生明顯感受到對程式的熟悉度以及思考錯誤以及修正錯誤的能力大幅提升。

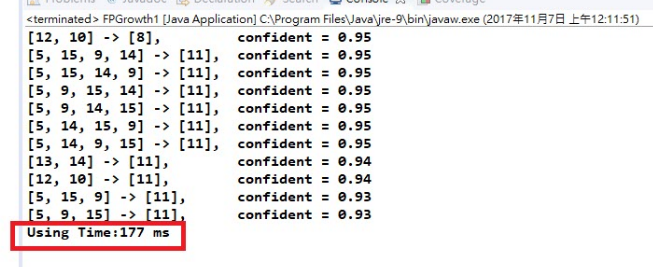
* **軟體應用能力**

1. 學會Weka吃得資料格式如何轉換，以及**學會應用weka來做關聯資料分析**。
2. **附上實驗結果圖片**

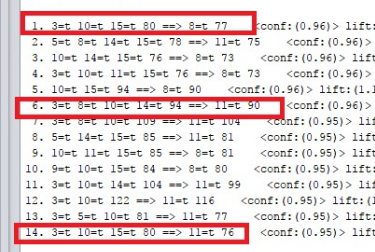
* **P JAVA用暴力法confident=0.95 support=0.25**



* **P JAVA用FP growth confident=0.95 support=0.25**



* **P WEKA用Apriori法confident=0.95 support=0.25**





* **P WEKA用FP growth confident=0.95 support=0.25**

