P76074648 張媤婷

Data Mining - Project1

檔案說明：

Homework1\_Apriori.py : 實作Apriori演算法的程式碼

Homework2\_FP-tree.py : 實作FP-Growth的程式碼

Book\_data : 課程講義上Apriori的範例資料

Book\_data2.txt : 課程講義上FP-tree的範例資料

Data.data : 由IBM Quest Data Generator產生的小型資料

IBM.data : 由IBM Quest Data Generator產生的大型資料

Output\_Testing\_Data.csv.csv : 由Data.data轉換成WEKA所能讀取的資料

實作方法：

Apriori： 用list來存每次產生出來的C(n)與L(n)，並用getWinner(k)來取出特定L(k)，取出的資料可以用來組成新的L(k+1)；另一getLoser()函式，用來找出被pruning掉的element，用這些elements來掃描L(k+1)，不斷重複，直到找不到新的一個Level。

FP-Growth：新增class Tree()，Tree有parent、child、sibling、next、data、num這些屬性，分別用來記錄物件的父親、第一個孩子、父親的下一個孩子，下一個相同值的物件、這個物件出現的次數。建立樹的方向是由上往下，經歷相同prefix的物件時不新增物件，只增加次數(num)；找尋Conditional Patterns由鎖定的item開始往上找，找完就前往next再往上找。最後用找到的pattern去找出frequent patterns。

效能分析：

取十次計算時間取平均 Support=0.01

1. Book\_data : 課程講義上Apriori的範例資料，4筆交易

Apriori ：It takes 2 milliseconds to find the above patterns

FP-Growth ：It takes 1 milliseconds to find the above patterns

→ FP-Growth比Apriori快了大約50%

1. Book\_data2.txt : 課程講義上FP-tree的範例資料，5筆交易

Apriori ：It takes 5.1 milliseconds to find the above patterns

FP-Growth ：It takes 4.5 milliseconds to find the above patterns

→ FP-Growth比Apriori快了大約13.33%

1. Data.data : 由IBM Quest Data Generator產生的小型資料，37筆交易

Apriori ：It takes 33.5 milliseconds to find the above patterns

FP-Growth ：It takes 26 milliseconds to find the above patterns

→ FP-Growth比Apriori快了大約28.85%

1. IBM.data : 由IBM Quest Data Generator產生的大型資料，24154筆交易

Apriori ：It takes 1374288 milliseconds to find the above patterns

FP-Growth ：It takes 320436 milliseconds to find the above patterns

→ FP-Growth比Apriori快了大約328.88%

第一、二、三種Dataset的運算時間為十次下來的平均值，第四種dataset則是取三次運算速度的平均。可以觀察出前三種數據的運算時間差異度不大，大約都在兩三成附近，但在第四種數據上，FP-Growth的速度可以達到Apriori的三倍之多，大幅凸顯出這兩個演算法的時間複雜度差異，尤其Apriori的作法更是讓L(1)與L(2)成為這個演算法最耗時的部分。