

**科目：1071\_資料探勘 DATA MINING**

**學生：楊沛霖**

**學號：Q36071156**

**日期：2018-10-23**

**前言：**

1. 本次作業利用不同的工具( 自己方法、WEKA )以及資料 ( 老師課堂資料、IBM、Kaggle ) 使用 FP-growth 來尋找 freqItems 並比較不同工具的結果。
2. 本次作業的 inputData 都放在 use\_data 資料夾內，outputData 則放置在 save\_data 資料夾。

## Topic 1:

先以老師上課資料確認程式運行是否正確。

Min\_sup : 2

執行時間 : 1.1854257

使用資料(IBM Quest Data Generator):

FP-growth:

老師講義上的圖(圖 一)與這次程式實作出來的圖(圖 二)

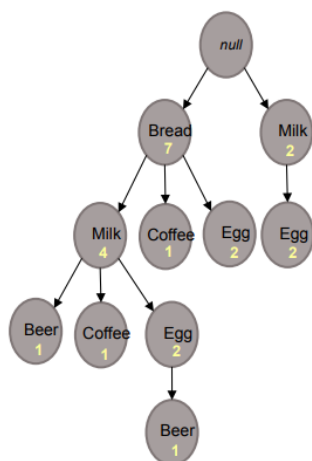


圖 二

```
Null Set 1
  Bread 7
    Milk 4
      Beer 1
      Coffee 1
      Egg 2
        Beer 1
    Coffee 1
    Egg 2
      Milk 2
      Egg 2
```

圖 一

## Association analysis:

分別為自己寫的程式(圖 三)與 WEKA 的 FP-tree(圖 四)的關聯顯示,可以看到兩邊答案相同。 對應 function 為 start.py 中的 `associate(freqItems, simpDat)`

```
{'Coffee'} >>> {'Bread'}
conf : 1.0
{'Beer'} >>> {'Milk'}
conf : 1.0
{'Beer'} >>> {'Milk', 'Bread'}
conf : 1.0
{'Beer'} >>> {'Bread'}
conf : 1.0
{'Milk', 'Beer'} >>> {'Bread'}
conf : 1.0
{'Beer', 'Bread'} >>> {'Milk'}
conf : 1.0
{'Milk'} >>> {'Egg'}
conf : 0.6666666666666666
{'Milk'} >>> {'Bread'}
conf : 0.6666666666666666
{'Milk', 'Bread'} >>> {'Beer'}
conf : 0.5
{'Milk', 'Bread'} >>> {'Egg'}
conf : 0.5
{'Egg'} >>> {'Milk'}
conf : 0.6666666666666666
{'Egg'} >>> {'Bread'}
conf : 0.6666666666666666
{'Milk', 'Egg'} >>> {'Bread'}
conf : 0.5
{'Egg', 'Bread'} >>> {'Milk'}
conf : 0.5
{'Bread'} >>> {'Milk'}
conf : 0.5714285714285714
{'Bread'} >>> {'Egg'}
conf : 0.5714285714285714
rules_count : 16
```

圖 三

```
Scheme:      weka.associations.FPGrowth -P 2 -I -1 -N 30 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 2.0
Relation:     weka_IBM
Instances:    9
Attributes:   5
              Bread
              Milk
              Beer
              Coffee
              Egg

=== Associator model (full training set) ===

FPGrowth found 16 rules (displaying top 16)

1. [Coffee=T]: 2 ==> [Bread=T]: 2 <conf:(1)> lift:(1.29) lev:(0.05) conv:(0.44)
2. [Beer=T]: 2 ==> [Bread=T]: 2 <conf:(1)> lift:(1.29) lev:(0.05) conv:(0.44)
3. [Beer=T]: 2 ==> [Milk=T]: 2 <conf:(1)> lift:(1.5) lev:(0.07) conv:(0.67)
4. [Beer=T]: 2 ==> [Bread=T, Milk=T]: 2 <conf:(1)> lift:(2.25) lev:(0.12) conv:(1.11)
5. [Bread=T, Beer=T]: 2 ==> [Milk=T]: 2 <conf:(1)> lift:(1.5) lev:(0.07) conv:(0.67)
6. [Milk=T, Beer=T]: 2 ==> [Bread=T]: 2 <conf:(1)> lift:(1.29) lev:(0.05) conv:(0.44)
7. [Milk=T]: 6 ==> [Bread=T]: 4 <conf:(0.67)> lift:(0.86) lev:(-0.07) conv:(0.44)
8. [Egg=T]: 6 ==> [Bread=T]: 4 <conf:(0.67)> lift:(0.86) lev:(-0.07) conv:(0.44)
9. [Milk=T]: 6 ==> [Egg=T]: 4 <conf:(0.67)> lift:(1) lev:(0) conv:(0.67)
10. [Egg=T]: 6 ==> [Milk=T]: 4 <conf:(0.67)> lift:(1) lev:(0) conv:(0.67)
11. [Bread=T]: 7 ==> [Milk=T]: 4 <conf:(0.57)> lift:(0.86) lev:(-0.07) conv:(0.58)
12. [Bread=T]: 7 ==> [Egg=T]: 4 <conf:(0.57)> lift:(0.86) lev:(-0.07) conv:(0.58)
13. [Bread=T, Milk=T]: 4 ==> [Egg=T]: 2 <conf:(0.5)> lift:(0.75) lev:(-0.07) conv:(0.44)
14. [Bread=T, Egg=T]: 4 ==> [Milk=T]: 2 <conf:(0.5)> lift:(0.75) lev:(-0.07) conv:(0.44)
15. [Milk=T, Egg=T]: 4 ==> [Bread=T]: 2 <conf:(0.5)> lift:(0.64) lev:(-0.12) conv:(0.3)
16. [Bread=T, Milk=T]: 4 ==> [Beer=T]: 2 <conf:(0.5)> lift:(2.25) lev:(0.12) conv:(1.04)
```

圖 四

## 比較結果:

程式經過老師範例的洗禮過後,能正常運行並且和分析工具 **WEKA** 所顯示的資料相同,因此可以加入 **IBM** 資料去處理了。

## Topic 2:

實際運行 IBM Quest Data Generator 所生成的資料

Min\_sup : 2

執行時間 : 2.124941

### 使用資料(IBM Quest Data Generator):

使用參數 :

```
lit -ntrans 1 -ascii -tlen 5 -nitems 10 -npats 3
```

圖 五為 IBM Quest Data Generator ， 表一 為切換成老師投影片的形式。

1	1	247
1	1	506
1	1	578
2	2	247
2	2	506
2	2	578
3	3	247
3	3	506
4	4	247
4	4	506
4	4	578
5	5	247
5	5	506
5	5	578
6	6	247
6	6	506
6	6	578

圖 五

TID	ItemSet
1	{247,506,578}
2	{247,506,578}
3	{247,506,247}
4	{247,506,578}
5	{247,506,578}
6	{247,506,578}

表 一

### FP-growth:

有程式生成的 FP-growth(圖 六)

```
[[ '247', '506', '578'], [ '247', '506', '578'], [ '247', '506'], [ '247', '506', '578'], [ '247', '506', '578']]
Null Set 1
  247 5
    506 5
      578 4
```

圖 六

## Association analysis:

分別為自己寫的程式(圖 八)與 WEKA 的 FP-tree(圖 七)的關聯顯示,可以看到兩邊答案相同。

對應 function 為 start.py 中的 `associate(freqItems, simpDat)`

```
{'578'} >>> {'247'}
conf : 1.0
{'578'} >>> {'506'}
conf : 1.0
{'578'} >>> {'247', '506'}
conf : 1.0
{'506', '578'} >>> {'247'}
conf : 1.0
{'247', '578'} >>> {'506'}
conf : 1.0
{'247'} >>> {'578'}
conf : 0.8
{'247'} >>> {'506', '578'}
conf : 0.8
{'247'} >>> {'506'}
conf : 1.0
{'506'} >>> {'578'}
conf : 0.8
{'506'} >>> {'247', '578'}
conf : 0.8
{'506'} >>> {'247'}
conf : 1.0
{'247', '506'} >>> {'578'}
conf : 0.8
rules_count : 12
```

圖 八

```
=== Run information ===

Scheme:      weka.associations.FPGrowth -P 2 -I -1 -N 20 -T 0 -C 0.1 -D 0.05 -U 1.0 -M 0.1
Relation:    weka_IBM
Instances:    5
Attributes:   3
              247
              506
              578

=== Associator model (full training set) ===

FPGrowth found 12 rules (displaying top 12)

1. [506=T]: 5 ==> [247=T]: 5 <conf:(1)> lift:(1) lev:(0) conv:(0)
2. [247=T]: 5 ==> [506=T]: 5 <conf:(1)> lift:(1) lev:(0) conv:(0)
3. [578=T]: 4 ==> [506=T]: 4 <conf:(1)> lift:(1) lev:(0) conv:(0)
4. [578=T]: 4 ==> [247=T]: 4 <conf:(1)> lift:(1) lev:(0) conv:(0)
5. [578=T]: 4 ==> [506=T, 247=T]: 4 <conf:(1)> lift:(1) lev:(0) conv:(0)
6. [506=T, 578=T]: 4 ==> [247=T]: 4 <conf:(1)> lift:(1) lev:(0) conv:(0)
7. [247=T, 578=T]: 4 ==> [506=T]: 4 <conf:(1)> lift:(1) lev:(0) conv:(0)
8. [506=T]: 5 ==> [578=T]: 4 <conf:(0.8)> lift:(1) lev:(0) conv:(0.5)
9. [247=T]: 5 ==> [578=T]: 4 <conf:(0.8)> lift:(1) lev:(0) conv:(0.5)
10. [506=T]: 5 ==> [247=T, 578=T]: 4 <conf:(0.8)> lift:(1) lev:(0) conv:(0.5)
11. [247=T]: 5 ==> [506=T, 578=T]: 4 <conf:(0.8)> lift:(1) lev:(0) conv:(0.5)
12. [506=T, 247=T]: 5 ==> [578=T]: 4 <conf:(0.8)> lift:(1) lev:(0) conv:(0.5)
```

圖 七

## 比較結果:

無論資料量大小皆與 WEKA 的分析結果一致,不過在自己 `associate` 的寫法中沒辦法像 WEKA 能夠把除了 True 或 False 的資料顯示出來,只能簡單的用存不存在去做對比。

## Topic 3:

### 使用資料(Kaggle):

使用資料來源：

<https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>

資料目的：

此資料為判斷每種遊戲主機銷量與顧客反應。

### FP-growth ( 部分 ) :

```
Sony Computer Entertainment 1
10 1
Simulation 1
8.2 1
22 1
83 1
PS 1
Namco 1
0.89 1
1997 1
16 1
Simulation 1
Sega 1
82 1
24 1
8.4 1
0.52 1
1999 1
DC 1
Role-Playing 192
2 1
24 1
8.6 1
PS 1
90 1
4 1
1999 1
SquareSoft 1
PC 31
Activision 3
2004 1
7.3 1
57 1
93 1
Blizzard Entertainment 1
6 1
2010 1
2 1
53 1
90 1
5.6 1
Blizzard Entertainment 1
7.5 1
2016 1
62 1
88 1
0.32 1
Blizzard Entertainment 1
```

經處理後的 WEKA 讀取文件 (arff) :

文件名稱 : @relation

Table Header 屬性 : @attribute

屬性分為 : string , numeric , nominal, date

分別為 字串, 數字, 自定義屬性 (ex. temperature {hot, mild, cool}) , 日期

```
1 @relation weka_
2
3 @attribute Name string
4 @attribute Platform string
5 @attribute Year_of_Release numeric
6 @attribute Genre string
7 @attribute Publisher string
8 @attribute NA_Sales numeric
9 @attribute EU_Sales numeric
10 @attribute JP_Sales numeric
11 @attribute Other_Sales numeric
12 @attribute Global_Sales numeric
13 @attribute Critic_Score numeric
14 @attribute Critic_Count numeric
15 @attribute User_Score numeric
16 @attribute User_Count numeric
17 @attribute Developer string
18 @attribute Rating string
19 @data
20 Name,Platform,Year_of_Release,Genre,Publisher,NA_Sales,EU_Sales,JP_Sales,Other_Sales,Global_Sales
21 Wii_Sports,Wii,2006,Sports,Nintendo,41,29,4,8,8253,76,51,8,322,Nintendo,E
22 Mario_Kart_Wii,Wii,2008,Racing,Nintendo,16,13,4,3,3552,82,73,83,709,Nintendo,E
23 Wii_Sports_Resort,Wii,2009,Sports,Nintendo,16,11,3,3,3277,80,73,8,192,Nintendo,E
24 New_Super_Mario_Bros,DS,2006,Platform,Nintendo,11,9,6,3,298,89,65,85,431,Nintendo,E
25 Wii_Play,Wii,2006,Misc,Nintendo,14,9,3,3,2892,58,41,66,129,Nintendo,E
```