Data Mining-HW1

P76061425 林聖軒

主程式: assoc_analysis.py

執行參數:

```
shengxuan@gpuserval-System-Product-Name:~/DM/DM_HW1$ python3 assoc_analysis.py -h
usage: assoc_analysis.py [-h] [-p POLICY] [-msp MIN_SUPPORT]
                         [-mcf MIN_CONFIDENCE] [-f DATA_FILE] [-ftp FILE_TYPE]

optional arguments:
  -h, --help            show this help message and exit
  -p POLICY             input policy,default=fpg,fpg=(fpGroeth),apr=apriori
  -msp MIN_SUPPORT      input min_support,default=3
  -mcf MIN_CONFIDENCE   input min_confidence,default=0.5
  -f DATA_FILE          input database,default=data.ntrans_1
  -ftp FILE_TYPE        input file type,default=i, i=(IBM-Quest-Data-Generator)
                        k=(kaggle data)
```

Result compare:

Dataset: data.ntrans_1 (IBM Quest Data)
Number_of_transactions:1000

Minimum Support = 3 ,Minimum Confidence = 0.5

執行時間:

Apriori Algorithm:

```
real    1m50.782s
```

FP-Growth:

```
real    0m5.904s
```

Minimum Support = 4 ,Minimum Confidence = 0.5

執行時間:

Apriori Algorithm:

```
real    0m13.639s
```

FP-Growth:

```
real    0m1.349s
```

Minimum Support = 5 ,Minimum Confidence = 0.5

執行時間:

Apriori Algorithm:

```
real    0m0.947s
```

FP-Growth:

```
real    0m0.610s
```

Kaggle DataSet (bonus):

Dataset: kaggle_data.csv
Number_of_transactions:9684

Minimum Support = 5 ,Minimum Confidence = 0.5

執行時間:

Apriori Algorithm:

```
shengxuan@gpuserval-System-Product-Name:~/DM/DM_HW1$ time py assoc_analysis.py -p apr
-msp 5 -msc 0.5 -f kaggle_data.csv -ftp k
```

```
real    1m15.463s
```

FP-Growth:

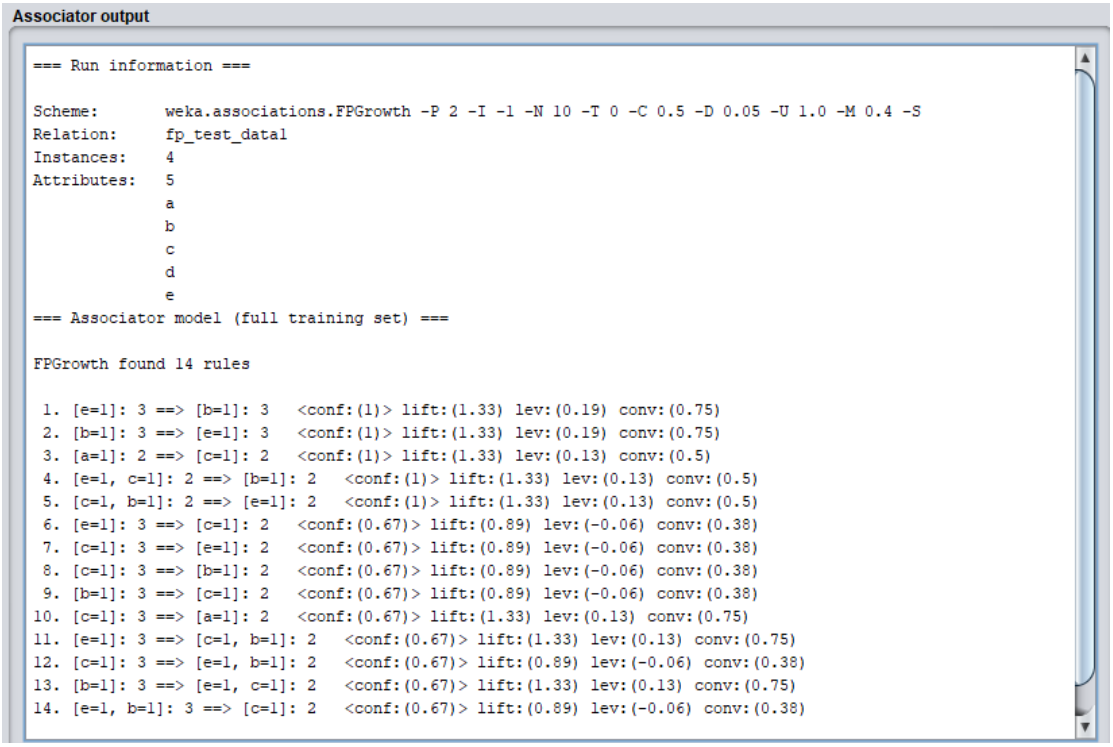
```
shengxuan@gpuserval-System-Product-Name:~/DM/DM_HW1$ time py assoc_analysis.py -p fpg
-msp 5 -msc 0.5 -f kaggle_data.csv -ftp k
```

```
real    0m0.564s
```

上面幾個例子可以很明顯的看到，FP-Growth 演算法的執行時間較少，效率明顯高於 Apriori Algorithm，而 minimum support 越低則需要越長的時間來找 association rule。

程式驗證:

與 weka 結果相同




```
shengxuan@gpuserval-System-Product-Name:~/DM/DM_HW1$ py assoc_analysis.py -p fpg -msp 2 -f example_data.txt
policy: fpg
min_support: 2
min_conf: 0.5
data_file: example_data.txt

frequent itemset:
1_itemset_count: 4
2_itemset_count: 4
3_itemset_count: 1
all_itemset_count: 9

association_list:
('b', 'c') -> [('e',)]
('c', 'e') -> [('b',)]
('b', 'e') -> [('c',)]
('c',) -> [('b', 'e'), ('a',), ('b',), ('e',)]
('b',) -> [('c', 'e'), ('c',), ('e',)]
('e',) -> [('c', 'b'), ('c',), ('b',)]
('a',) -> [('c',)]
found 14 rules
```