

Data Mining - Project 1 Association Rule

電通甲 Q36061240 洗鈺淇

I. Dataset1 : Use IBM Quest Synthetic Data Generator

Parameter settings :

1. Number of transactions in database(*1000)
2. Average transaction length
3. Number of items(*1000)

共產生不同參數的六組 dataset 來測試

- (a) ntrans=0.1,tlen=20,nitems=0.1
- (b) ntrans=0.1,tlen=40,nitems=0.1
- (c) ntrans=0.1,tlen=20,nitems=0.5
- (d) ntrans=0.1,tlen=40,nitems=0.5
- (e) ntrans=0.5,tlen=20,nitems=0.1
- (f) ntrans=0.5,tlen=40,nitems=0.1

Preprocessing :

data.ntrans_0.1.tlen_40.nitems_0.5.txt - 記事本

檔案(F)	編輯(E)	格式(O)	檢視(V)	說明(H)
1	1	1	326	
1	1	1	360	
1	1	1	368	
1	1	1	385	
1	1	1	388	
1	1	1	399	
1	1	1	400	
1	1	1	412	
1	1	1	420	
1	1	1	432	
1	1	1	444	
1	1	1	481	
1	1	1	486	
1	1	1	493	
2	2	2	19	
2	2	2	28	
2	2	2	28	

經由 IBM Quest Synthetic Data Generator 產生出的 data 格式為 <CustID, TransID, Item>(圖 1)，轉換成每一行代表一個 transaction 並以逗號隔開每個 item 的格式（見圖 2）

圖 1: data.ntrans_0.1.tlen_40.nitems_0.5.txt 資料格式

data.ntrans_0.1.tlen_40.nitems_0.5DAT.txt - 記事本

檔案(F)	編輯(E)	格式(O)	檢視(V)	說明(H)
3, 14, 19, 21, 23, 31, 39, 51, 52, 78, 83, 136, 139, 146, 150, 155, 176, 188, 201, 204, 221, 222, 240, 246, 247, 259, 265, 272, 274, 284, 292, 296, 314, 316, 318, 3				
19, 28, 29, 33, 34, 36, 40, 43, 67, 69, 73, 89, 103, 147, 160, 167, 169, 172, 176, 186, 192, 193, 208, 213, 214, 217, 220, 274, 278, 281, 293, 318, 344, 353, 385, 4				
3, 4, 15, 21, 28, 48, 49, 51, 61, 63, 66, 69, 81, 85, 86, 95, 104, 107, 125, 132, 155, 205, 221, 238, 284, 293, 297, 302, 316, 367, 371, 374, 377, 395, 432, 439, 472				
3, 8, 11, 17, 26, 27, 51, 52, 60, 63, 69, 76, 87, 97, 127, 147, 163, 168, 185, 191, 216, 222, 260, 274, 277, 281, 288, 289, 298, 299, 306, 335, 336, 339, 360, 368, 3				
3, 8, 11, 14, 40, 61, 78, 87, 119, 123, 127, 132, 167, 199, 210, 213, 217, 219, 222, 232, 233, 238, 282, 322, 339, 347, 353, 371, 374, 377, 390, 402, 403, 409, 412				
9, 18, 23, 24, 48, 71, 107, 127, 140, 148, 149, 155, 192, 209, 238, 248, 278, 283, 286, 287, 293, 306, 316, 318, 325, 399, 414, 419, 429, 437, 457,				
5, 9, 17, 18, 29, 36, 38, 66, 68, 74, 87, 89, 107, 118, 123, 127, 132, 138, 154, 182, 192, 201, 209, 238, 253, 259, 277, 279, 283, 298, 300, 311, 360, 361, 404, 419				
5, 21, 28, 49, 50, 63, 107, 118, 119, 127, 129, 137, 145, 147, 209, 214, 219, 274, 306, 318, 399, 414, 416, 439, 442, 472, 490,				
8, 33, 48, 61, 73, 87, 89, 95, 100, 106, 118, 119, 124, 127, 137, 139, 154, 155, 170, 178, 210, 222, 238, 246, 292, 293, 300, 325, 329, 360, 371, 385, 404, 412, 41				
38, 63, 89, 100, 103, 123, 138, 148, 166, 204, 217, 248, 274, 278, 295, 318, 322, 326, 364, 365, 371, 374, 392, 414, 420, 425, 432, 444, 447, 464, 467, 486, 493,				
8, 17, 87, 100, 117, 119, 146, 147, 152, 154, 159, 160, 175, 199, 219, 221, 231, 234, 243, 247, 277, 319, 325, 331, 338, 353, 360, 368, 395, 402, 416, 418, 423, 4				
10, 26, 33, 34, 36, 38, 40, 43, 61, 63, 80, 87, 103, 112, 133, 140, 143, 147, 167, 168, 172, 188, 191, 214, 220, 222, 236, 242, 247, 259, 261, 291, 316, 318, 321, 2				
9, 36, 62, 63, 69, 78, 100, 108, 138, 140, 148, 150, 163, 171, 192, 201, 209, 229, 234, 247, 277, 286, 306, 316, 319, 339, 340, 353, 356, 384, 395, 398, 425, 437,				
4, 9, 10, 11, 12, 31, 38, 39, 55, 61, 73, 81, 101, 117, 120, 132, 142, 159, 182, 206, 227, 231, 233, 238, 241, 274, 281, 285, 316, 318, 360, 368, 372, 374, 395, 402				
3, 14, 16, 28, 29, 48, 59, 60, 62, 63, 67, 80, 100, 103, 111, 117, 124, 130, 154, 155, 166, 178, 191, 198, 221, 231, 233, 234, 325, 335, 336, 337, 346, 371, 374, 37				
3, 28, 34, 38, 39, 43, 53, 55, 63, 67, 71, 72, 73, 78, 80, 107, 112, 123, 128, 135, 142, 150, 173, 174, 176, 179, 182, 191, 205, 221, 231, 273, 274, 281, 299, 306, 3				
25, 38, 67, 69, 90, 92, 106, 111, 112, 119, 123, 124, 127, 142, 152, 160, 163, 188, 222, 234, 247, 259, 263, 266, 302, 359, 377, 404, 423, 435, 436, 443, 461, 469				
9, 19, 55, 61, 66, 80, 96, 99, 100, 144, 147, 168, 171, 192, 202, 204, 215, 274, 284, 338, 371, 382, 389, 399, 410, 412, 425, 453, 455, 459, 470, 472, 480, 486, 48				
25, 34, 38, 62, 69, 83, 87, 93, 104, 106, 111, 129, 146, 147, 171, 176, 185, 201, 214, 216, 221, 238, 246, 260, 300, 319, 324, 336, 338, 339, 352, 353, 374, 388, 3				
3, 11, 14, 43, 60, 63, 66, 86, 98, 106, 107, 117, 132, 137, 142, 146, 147, 154, 159, 164, 168, 174, 180, 188, 209, 212, 238, 246, 247, 248, 254, 266, 274, 289, 290				
11, 14, 35, 36, 39, 63, 74, 87, 95, 98, 101, 108, 112, 123, 124, 132, 140, 148, 176, 178, 209, 213, 234, 238, 247, 257, 283, 284, 290, 360, 387, 405, 437, 443, 455				
21, 28, 29, 33, 36, 47, 62, 63, 71, 81, 85, 96, 104, 118, 127, 132, 135, 141, 150, 155, 188, 192, 221, 240, 289, 303, 306, 319, 325, 339, 359, 374, 379, 395, 416, 4				

圖 2: data.ntrans_0.1.tlen_40.nitems_0.5DAT.txt 資料格式

以下由(b) ntrans=0.1,tlen=40,nitems=0.5 的 dataset 來做範例，Minimum support = 0.08, Confidence = 0.7，會算出所有 Frequent itemset 以及產生 Association Rules，IBMOutput.txt 為此範例完整執行結果。

A. Apriori Algorithm

```
=====
Rules:
(191) ==> (63) confidence = 0.733
(385) ==> (222) confidence = 0.727
(490) ==> (63) confidence = 0.733
(222, 471) ==> (87) confidence = 0.833
(87, 471) ==> (222) confidence = 0.833
(8, 87) ==> (222) confidence = 0.75
(117, 38) ==> (402) confidence = 0.727
(439, 316) ==> (221) confidence = 0.889
(221, 316) ==> (439) confidence = 0.727
(132, 8) ==> (222) confidence = 0.75
(402, 403) ==> (318) confidence = 0.8
(318, 403) ==> (402) confidence = 1.0
(318, 402) ==> (403) confidence = 0.727
=====
1.132249 sec
```

▲圖 3：Dataset1 Implement with Apriori Algorithm

- 執行結果共有 13 條 Association Rules
- 執行時間約為 1.132249 sec

B. FP-growth

```
=====
(117 38) ==> (402 ) confidence= 0.73
(191) ==> (63 ) confidence= 0.73
(316 439) ==> (221 ) confidence= 0.89
(316 221) ==> (439 ) confidence= 0.73
(385) ==> (222 ) confidence= 0.73
(403 318) ==> (402 ) confidence= 1.00
(403 402) ==> (318 ) confidence= 0.80
(318 402) ==> (403 ) confidence= 0.73
(471 222) ==> (87 ) confidence= 0.83
(471 87) ==> (222 ) confidence= 0.83
(490) ==> (63 ) confidence= 0.73
(8 132) ==> (222 ) confidence= 0.75
(8 87) ==> (222 ) confidence= 0.75
0.124147 sec
```

▲圖 4：Dataset1 Implement with FP-growth

- 執行結果共有 13 條 Association Rules
- 執行時間約為 0.124147 sec

C. Compare

嘗試執行其他 dataset 來比較參數與執行效率的關係,下表執行結果，執行時間單位(s)

	A	B	C	D	E	F
<i>nTrans</i>	100	100	500	500	100	100
<i>tLen</i>	20	40	20	40	20	40
<i>nItems</i>	100	100	100	100	500	500
<i>Apriori(0.4,0.7)</i>	0.03952	8.28184	0.229484	3.609302	0.6	0.937805
<i>FPGrowth(0.4,0.7)</i>	0.006212	0.163457	0.020077	0.399853	0	0.013558
<i>Apriori(0.04,0.7)</i>	-	-	27.70589	-	0.684226	4.810117
<i>FPGrowth(0.04,0.7)</i>	3.548225	-	1.136646	-	0.042135	0.261417

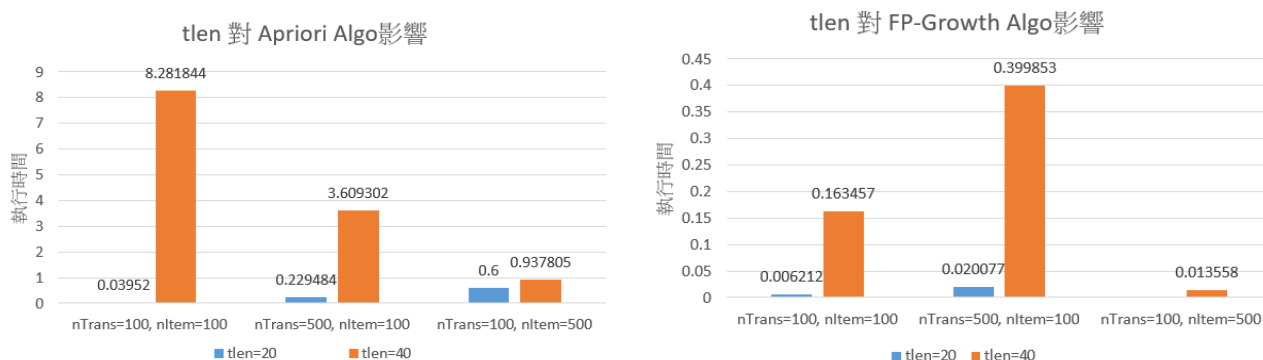
▲圖 5：各 dataset-algo 執行時間

觀察其中比較明顯的特點：單一 transaction 長度(tlen)以及 Minimum support 值來比較演算法的效率。

- Minimum support

由於 E 和 F 的 item 數比較多,於是當 Minimum support 設得值比較大時,找到的 Frequent itemset 會很少於是也得不出任何的 rule，執行時間也比較沒有參考價值，於是 Minimum support 調為 0.04 即可產生比較有意義的 rule；又以 A 和 C 這兩個 dataset 來觀察，若 Minimum support 越小，Apriori 演算法候選項集會越多，FP-Growth 的 FP-tree 會較茂盛，因此兩個演算法的執行時間皆會越長。

- 單一 transaction 長度(tlen)



▲圖 6：tlen 對 Apriori Algorithm 及 FP-Growth 執行效之影響

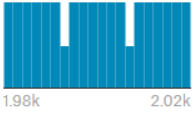

由上圖可得知 tlen 對 Apriori Algorithm 執行效率的影響遠比 FP-Growth 大，對於 transaction 長度比較大的 dataset 若 Minimum support 較大時，此時由於生成的 FP-tree 深度較大，FP 演算法效能顯著下降；若 Minimum support 較小時，此時 Apriori Algo 掃描資次數增加，Apriori 效能會明顯的下降。

II. Dataset2 : Select from kaggle.com

Dataset2 : New Zealand Migration

Migration numbers to and from New Zealand from 1979 to 2016

URL : <https://www.kaggle.com/timoboz/migration-nz>

	Measure	Country	Citizenship	Year	Value
	The signal type given in this row, one of: "Arrivals", "Departures", "Net"	Country from where people arrived into to New Zealand (for Measure = "Arrivals") or to where they left (for Measure = "Departures")	Citizenship of the migrants, one of: "New Zealand Citizen", "Australian Citizen", "Total All Citizenships"	Year of the measurement	Number of migrants
	Arrivals 33% Departures 33% Other (t) 33%	253 unique values	New Zealand Cit... 33% Australian Citizen 33% Other (t) 33%		
1	Arrivals	Oceania	New Zealand Citizen	1979	11817
2	Arrivals	Oceania	Australian Citizen	1979	4436
3	Arrivals	Oceania	Total All Citizenships	1979	19965
4	Arrivals	Antarctica	New Zealand Citizen	1979	10
5	Arrivals	Antarctica	Australian Citizen	1979	0
6	Arrivals	Antarctica	Total All Citizenships	1979	13
7	Arrivals	American Samoa	New Zealand Citizen	1979	17
8	Arrivals	American Samoa	Australian Citizen	1979	4
9	Arrivals	American Samoa	Total All Citizenships	1979	30

▲圖 7: Dataset2 New Zealand Migration

Column Metadata :

Metadata	Type	Description
Year	Int	年份
Measure	String	"Arrivals"移入或"Departures"移出
Country	String	從哪個城市移入至紐西蘭或移出紐西蘭至哪個城市
Citizenship	String	身分
Value	Int	移民數

Preprocessing :

將相同年份、對應城市和公民身分的移出量減去移入量，若所得的值>1000 再把該對應城

```

Migdataset.txt - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)
1979 Oceania
1979 Australia
1979 New-Zealand
1979 Samoa
1979 Asia
1979 Europe
1979 UK
1979 Americas
1979 USA
1979 Not-stated
1980 Oceania
1980 Australia
1980 New-Zealand
  
```

市納入考量，此用意是為了計算出從紐西蘭每年移出至各個城市人口之間的關聯性，經過轉換後所得

Migdataset.txt<year,country>再轉換為每年為一個 transaction，執行結果為 Migoutput.txt

◀圖 8: Migdataset.txt 資料格式

Migdat.txt - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)

```
Oceania,Australia,New-Zealand,Samoa,Asia,Europe,UK,Americas,USA,Not-stated,
Oceania,Australia,New-Zealand,Samoa,Asia,Singapore,Vietnam,Europe,UK,Netherlands,Americas,USA,Not-stated,
Oceania,Australia,New-Zealand,Asia,Europe,UK,Netherlands,Americas,USA,Not-stated,
Oceania,Australia,New-Zealand,Samoa,Asia,Singapore,Europe,UK,Netherlands,Americas,Canada,USA,Not-stated,
Oceania,Australia,New-Zealand,Asia,Europe,UK,Netherlands,Americas,Canada,USA,Not-stated,
Oceania,Australia,New-Zealand,Samoa,Asia,Europe,UK,Americas,USA,Not-stated,
Oceania,Australia,New-Zealand,Samoa,Asia,Europe,UK,Americas,USA,Not-stated,
Oceania,Australia,New-Zealand,Samoa,Asia,Europe,UK,Americas,Canada,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,Singapore,Europe,UK,Americas,Canada,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,New-Zealand,Samoa,Asia,Japan,Malaysia,Singapore,Europe,UK,Americas,USA,Not-stated,
Oceania,Australia,Fiji,New-Zealand,Samoa,Asia,Hong-Kong,Japan,Malaysia,Singapore,Taiwan,Europe,UK,Americas,USA,Not-stated,
Oceania,Australia,Fiji,New-Zealand,Samoa,Asia,Hong-Kong,Japan,Malaysia,Taiwan,Europe,UK,Americas,USA,Not-stated,
Oceania,Australia,Fiji,Asia,Hong-Kong,Japan,Malaysia,Taiwan,Europe,UK,US,USA,Americas,Canada,USA,Not-stated,
Oceania,Australia,Fiji,Asia,Hong-Kong,Japan,Malaysia,Taiwan,Europe,UK,US,USA,Americas,Canada,USA,Not-stated,
Oceania,Australia,Fiji,Asia,Hong-Kong,Japan,South-Korea,Malaysia,Taiwan,Europe,UK,Americas,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,Hong-Kong,India,Japan,South-Korea,Malaysia,Taiwan,Europe,UK,Americas,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,Hong-Kong,India,Japan,South-Korea,Malaysia,Taiwan,Europe,UK,Americas,Canada,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,Hong-Kong,India,Japan,South-Korea,Malaysia,Philippines,Taiwan,Europe,UK,Americas,Canada,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,Hong-Kong,India,Japan,South-Korea,Malaysia,Taiwan,Europe,UK,Americas,Canada,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,Hong-Kong,India,Japan,Malaysia,Taiwan,Europe,UK,Americas,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,Hong-Kong,India,Japan,Taiwan,Europe,UK,Americas,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,India,Japan,South-Korea,Europe,UK,Americas,USA,Africa-and-the-Middle-East,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,India,Japan,South-Korea,Europe,UK,Americas,USA,Africa-and-the-Middle-East,Not-stated,
```

▲圖 9: Migdat.txt 資料格式

A. Apriori Algorithm

```
(UK, Asia, USA, Australia, Americas) ==> (Oceania) confidence = 0.974
(UK, Oceania, USA, Australia, Americas) ==> (Asia) confidence = 0.974
(UK, Oceania, Asia, Australia, Americas) ==> (USA) confidence = 0.974
(UK, Oceania, Asia, USA, Americas) ==> (Australia) confidence = 0.974
(UK, Oceania, Asia, USA, Australia) ==> (Americas) confidence = 0.974
(UK, Oceania, Asia, USA, Not-stated, Americas) ==> (Australia) confidence = 1.0
(UK, Oceania, Asia, USA, Not-stated, Australia) ==> (Americas) confidence = 1.0
(Oceania, Europe, Asia, USA, Not-stated, Australia, Americas) ==> (UK) confidence = 1.0
(UK, Europe, Asia, USA, Not-stated, Australia, Americas) ==> (Oceania) confidence = 1.0
(UK, Oceania, Asia, USA, Not-stated, Australia, Americas) ==> (Europe) confidence = 1.0
(UK, Oceania, Europe, USA, Not-stated, Australia, Americas) ==> (Asia) confidence = 1.0
(UK, Oceania, Europe, Asia, Not-stated, Australia, Americas) ==> (USA) confidence = 1.0
(UK, Oceania, Europe, Asia, USA, Australia, Americas) ==> (Not-stated) confidence = 0.974
(Oceania, Europe, Asia, USA, Australia, Americas) ==> (UK) confidence = 0.974
(UK, Europe, Asia, USA, Australia, Americas) ==> (Oceania) confidence = 0.974
(UK, Oceania, Asia, USA, Australia, Americas) ==> (Europe) confidence = 0.974
(UK, Oceania, Europe, USA, Australia, Americas) ==> (Asia) confidence = 0.974
(UK, Oceania, Europe, Asia, Australia, Americas) ==> (USA) confidence = 0.974
(UK, Oceania, Europe, Asia, USA, Americas) ==> (Australia) confidence = 0.974
(UK, Oceania, Europe, Asia, USA, Australia) ==> (Americas) confidence = 0.974
(UK, Oceania, Europe, Asia, USA, Not-stated, Americas) ==> (Australia) confidence = 1.0
(UK, Oceania, Europe, Asia, USA, Not-stated, Australia) ==> (Americas) confidence = 1.0
=====
0.177532 sec
```

▲圖 10: Dataset2 Implement with Apriori Algorithm

- 執行時間約為 0.177532 sec

B. FP-growth

```
(Oceania USA) ==> (Not-stated ) confidence= 0.97
(Not-stated) ==> (UK ) confidence= 1.00
(UK) ==> (Not-stated ) confidence= 0.97
(Not-stated) ==> (UK USA ) confidence= 1.00
(UK) ==> (Not-stated USA ) confidence= 0.97
(USA) ==> (Not-stated UK ) confidence= 0.97
(Not-stated UK) ==> (USA ) confidence= 1.00
(Not-stated USA) ==> (UK ) confidence= 1.00
(UK USA) ==> (Not-stated ) confidence= 0.97
(Not-stated) ==> (USA ) confidence= 1.00
(USA) ==> (Not-stated ) confidence= 0.97
(Oceania) ==> (UK ) confidence= 1.00
(UK) ==> (Oceania ) confidence= 1.00
(Oceania) ==> (UK USA ) confidence= 1.00
(UK) ==> (Oceania USA ) confidence= 1.00
(USA) ==> (Oceania UK ) confidence= 1.00
(Oceania UK) ==> (USA ) confidence= 1.00
(Oceania USA) ==> (UK ) confidence= 1.00
(UK USA) ==> (Oceania ) confidence= 1.00
(Oceania) ==> (USA ) confidence= 1.00
(USA) ==> (Oceania ) confidence= 1.00
(UK) ==> (USA ) confidence= 1.00
(USA) ==> (UK ) confidence= 1.00
0.048869 sec
```

▲圖 11: Dataset1 Implement with FP-growth

- 執行時間約為 0.048869 sec

由於移出量設置的閾值太小(Avg 大約是 2500 但我設 1000)，造成取出來的城市多半為每年都

會有固定移民從紐西蘭移出至這些城市，不夠具有因年代而改變移民數的代表性，導致算出來的 Frequent itemset 很多，然後 confidence 多半也都是 1。