科目: 1071\_資料探勘 DATA MINING

學生:楊沛霖

學號: Q36071156

日期:2018-11-18

## 作業要求:

1. 利用課程給的八個 Graphs 來計算 HITS、PageRank 和 SimRank。

- 2. 運算所需的參數為 random jumping probability
- 3. 討論和分析每個演算法和結果。

# 1. 各圖間的比較

## graph\_1:

是一條線單一方向的圖由  $1\rightarrow 2\rightarrow 3\rightarrow 4\rightarrow 5\rightarrow 6$ 

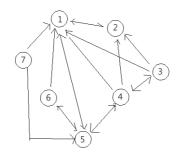
## graph\_2:

是一條線但是方向是雙向的由 154254354

## graph\_3:

是一條線單一方向的圖由 1>2>3....>6

## graph\_4:



## **graph**\_5,6:

已過於複雜

## 2. Hits

### 實作方式:

```
\begin{array}{lll} \text{HubsAuthorities}(G) \\ 1 & 1 & \in [1, \dots, 1] \in \mathbb{R}^{|V|} \\ 2 & a_0 & \in h_0 \in 1 \\ 3 & t & \in 1 \\ 4 & \text{repeat} \\ 5 & \text{for each v in V} \\ 6 & do \ a_t \ (v) \leftarrow \sum_{w \in pa[v]} h \ (w) \\ 7 & & h_t \ (v) \leftarrow \sum_{w \in ch[v]} a \ (w) \\ 8 & & a_t \leftarrow a_t \ / \ \| \ a_t \| \\ 9 & & h_t \leftarrow h_t \ / \ \| \ h_t \| \\ 10 & & \text{normalization} \\ 11 & \text{until } \| \ a_t - a_{t-1} \| + \| \ h_t - h_{t-1} \| \le \epsilon \\ 12 & \text{return } (a_t, h_t) \end{array}
```

1. 資料初始化,看有幾個點就創幾個 a 和 h 而他們的值皆為 1(1.2.3 行)

- 2. 對每個有連接到  $a_t(v)$ 的點的  $h_{t-1}(w)$ 作加總(第 6 行)
- 3. 對每個有連接到 h<sub>t</sub>(v)的點的 a<sub>t-1</sub>(w)作加總(第 7 行)
- 4. 正規化(第8.9行)
- 5. 重複 2-5 直到 a 和 h 與前項相減絕對值得加總小於誤差值
- 6. 取得 at 和 ht

### D 設定為: 0.15

正規化方式為:平方後加總在開根號

## graph\_1:

時間與 cpu:

0.46s user 0.19s system 14% cpu 4.626 total

#### 結果:

收斂執行次數:3

收斂結束時 Error: 0.9190116821894447

```
authority page : {'1': 0.5, '2': 0.5, '3': 0.5, '4': 0.5, '5': 0.0, '6': 0}
hub : {'1': 0.5, '2': 0.5, '3': 0.5, '4': 0.5, '5': 0.0, '6': 0}
```

# graph\_2:

時間與 cpu:

0.46s user 0.15s system 4% cpu 14.856 total

結果:

收斂執行次數:3

收斂結束時 Error: 0.0

authority page :

```
{'1': 0.4472135954999579, '2': 0.4472135954999579, '3':
      0.4472135954999579, '4': 0.4472135954999579, '5':
      0.4472135954999579}
    hub page :
      {'1': 0.4472135954999579, '2': 0.4472135954999579, '3':
      0.4472135954999579, '4': 0.4472135954999579, '5':
      0.4472135954999579}
graph 3:
  時間與 cpu:
    0.46s user 0.15s system 4% cpu 14.856 total
  結果:
    收斂執行次數:3
    收斂結束時 Error: 0.24854976622531513
    authority page :
      {'1': 0.3922322702763681, '2': 0.5883484054145521, '3':
      0.5883484054145521, '4': 0.3922322702763681}
    hub page :
      {'1': 0.3922322702763681, '2': 0.5883484054145521, '3':
      0.5883484054145521, '4': 0.3922322702763681}
graph 4:
  時間與 cpu:
    0.44s user 0.15s system 48% cpu 1.256 total
  結果:
    收斂執行次數:3
    收斂結束時 Error: 0.5012605876917594
    authority page :
      {'1': 0.5063160398440778, '2': 0.23014365447458082, '3':
      0.276172385369497, '4': 0.3222011162644131, '5':
```

```
0.552344770738994, '6': 0.4142585780542455, '7':
0.18411492357966464}
hub page :
{'1': 0.5063160398440778, '2': 0.23014365447458082, '3':
0.276172385369497, '4': 0.3222011162644131, '5':
0.552344770738994, '6': 0.4142585780542455, '7':
0.18411492357966464}
```

## graph\_5:

時間與 cpu:

0.45s user 0.15s system 55% cpu 1.056 total

結果:

收斂執行次數:3

收斂結束時 Error: 1.4092964448407446

#### authority page :

authority page: ('3': 0.04119823789241145, '6': 0.043487028886434305, '1': 0.67781889379677717, '396': 0.03204307391632001, '223': 0.0709525208147086, '412': 0.4463142438344573, '38': 0.25863338232458924, '95': 0.12130592268321146, '179': 0.146482 (2361746292, '236': 0.12130592268321146, '43': 0.05950856584459431, '169': 0.027465491928274296, '369': 0.05950856584459431, '152': 0.13961625063539435, '264': 0.12130592268321146, '273': 0.10070680373706575, '277': 0.12130592268321146, '321': 0.10 070680373706575, '327': 0.1191713168918861, '366': 0.104070680373706575, '327': 0.0546635208321146, '321': 0.10 070680373706575, '327': 0.04666372982068575, '163': 0.06408614783264002, '443': 0.06866372982068575, '458': 0.1533489965995315, '5': 0.08468526677884575, '906': 0.16021536958160 01, '380': 0.150386668668734862, '7': 0.14877141461148577, '60': 0.14877141461148577, '60': 0.18504866868684734862, '7': 0.14877141461148577, '60': 0.14877141461148577, '60': 0.18504866868684734862, '7': 0.14877141461148577, '60': 0.18771474525666866, '12': 0.139616256539435, '258': 0.04119823789241145, '134': 0.0480646108674348024, '176': 0.0869740577728686, '85': 0.04119823789241145, '134': 0.048064610874480024, '176': 0.0869740577728686, '85': 0.04119823789241145, '134': 0.048064610874480024, '176': 0.0869740577728686, '182': 0.0869740577728686, '85': 0.04119823789241145, '134': 0.09155163976091432, '185': 0.061797356838617166, '212': 0.061797356838617166, '283': 0.12359471367722432, '293': 0.1098640671369717, '291': 0.09384043075493717, '299': 0.09384043075493717, '315': 0.07953010280275431, '456': 0.07553010280275431, '456': 0.07553010280275431, '456': 0.07553010280275431, '456': 0.07553010280275431, '456': 0.07553010280275431, '456': 0.07553010280275431, '456': 0.07553010280275431, '456': 0.09384043075493717, '299': 0.09384043075493717, '315': 0.07087818939, '207': 0.12359471367722432, '293': 0.013732745964137148, '259': 0.09584046977388864, '320': 0.08549309838556578, '300': 0.08549309838565489, '300': 0.0854930938756514, '416': 0.0

hub page :

hub: ('3': 0.04119823789241142, '5': 0.08468526677884572, '1': 0.07781889379677713, '2': 0.094577581988045714, '222': 0.070
95522081470857, '227': 0.12359471367723428, '235': 0.074655491928274285, '25': 0.07553010280275428, '412': 0.46431424383445
706, '70': 0.03890944689838857, '94': 0.1213059226832114, '152': 0.13961625063539426, '273': 0.10070680373700569, '277': 0.1
213059226832114, '369': 0.05950856584459428, '386': 0.10070680373700569, '192': 0.03204307391632, '202': 0.01831032795218285
5, '38': 0.25863338232458283, '184': 0.08293647578482284, '458': 0.153780965995314, '12': 0.13961625063539426, '84': 0.1670
817425636686, '258': 0.14419383262344, '306': 0.16021536958160001, '380': 0.13503866864734856, '43': 0.05950856584459428, '1
09': 0.027465491928274285, '58': 0.06866372982068572, '103': 0.06408614783264, '443': 0.06866372982068572, '56': 0.091551639
76091427, '204': 0.113207926283114, '321': 0.10070680373700569, '327': 0.11901713168918855, '439': 0.0866740577786857, '264
': 0.02659911894620571, '307': 0.05493098385654857, '17': 0.1487714146114857, '60': 0.1487714146114857, '148': 0.0366206555944
36571, '274': 0.1327498776533257, '37': 0.086697405777286857, '115': 0.09155163976091428, '176': 0.08667405777286857, '182': 0.08697405777286857, '182': 0.08697405777286857, '182': 0.085721977485057143, '348': 0.13046108665930284, '440': 0.05721977485057143, '85': 0.0752397485057143, '348': 0.0595340180850286, '303': 0.06408614783264, '134': 0.048664610874479996, '179': 0.14648262
817462844, '249': 0.013732745964137143, '13': 0.036620655904455751, '212': 0.066179735603861714, '297': 0.09384043075493714, '297': 0.09384043075493714, '326': 0.0755301028027543, '47': 0.027754282922229714, '11': 0.13732745964137143, '135': 0.061797356038661714, '297': 0.09384043075493713, '135': 0.0705522081470857, '371': 0.1167283438572550143, '456': 0.0755301028027543, '47': 0.027754282922229714, '11': 0.13732745964137143, '135': 0.06179735603866174, '297': 0.09384043075493713, '135': 0.0407758188937967771
3, '119': 0.05493098385

## graph\_6:

時間與 cpu:

0.48s user 0.25s system 57% cpu 1.456 total

結果:

收斂執行次數:3

收斂結束時 Error: 0.8876169361023691

資料量太大其他結果放在: some\_output\hits\_G6.txt

討論:由於 Hits 的算法就算被多個 node 指但沒有對其它 node 指的點 hub 值會是 0 而如果都指向人而沒被指的話 authority 也會是 0,而相較於 PageRank Hits 是比較容易被作弊的,比如作弊者可以建立一個網頁,頁面內容增加很多指向高質量網頁或者著名網站的網址,這就是一個很好的 Hub 頁面,之後作弊者再將這個網頁連結指向作弊網頁,於是可以提升作弊網頁的 Authority 得分。

## 3. PageRank

#### 實作方式:

$$PR(P_i) = \frac{(d)}{n} + (1 - d) \times \sum_{l_{j,i} \in E} PR(P_j) / \text{Outdegree}(P_j)$$

D(damping factor)=0.1~0.15 n=|page set|

因此為 Quick reference 所以只要照著這個公式去實作即可。

$$\sum_{j, \in I} PR(P_j) / \text{Outdegree}(P_j)$$
 為  $P_j$  的  $P_i$  R 除上所有射入  $P_j$  點的數量

## D 設定為: 0.15

## graph\_1:

時間與 cpu:

0.41s user 0.13s system 49% cpu 1.071 total

#### 結果:

{'1': 0.02499999999999998, '2': 0.16666666666666666, '3':

# graph\_2:

時間與 cpu:

0.34s user 0.07s system 92% cpu 0.471 total

結果:

{'1': 0.2, '2': 0.2, '3': 0.2, '4': 0.2, '5': 0.2}

## graph\_3:

時間與 cpu:

0.38s user 0.12s system 72% cpu 0.711 total

```
結果:
   {'1': 0.14375, '2': 0.35624999999999996, '3': 0.3562499999999996, '4':
   0.14375}
graph 4:
  時間與 cpu:
   0.33s user 0.06s system 92% cpu 0.381 total
  結果:
   {'1': 0.2946428571428571, '2': 0.1469047619047619, '3':
   0.11654761904761904, '4': 0.07607142857142857, '5':
   0.045714285714285714}
graph 5:
  時間與 cpu:
   0.41s user 0.05s system 92% cpu 0.781 total
  結果: 由於數量眾多,只取前面常出現的數值,以下格式為:
   (數值,出現次數)
   (0.0005787389582698751, 12), (0.00044065387348969436, 15),
   (0.00044928419128845566, 17), (0.00041521714734597684, 19),
   (0.00045924225028702635, 20), (0.0005463752665245202, 20),
   (0.0004264392324093816, 20), (0.00039534470504619753, 20),
   (0.0003923240938166311, 24), (0.00036752328582650653, 25),
    (0.0004331023454157782, 30), (0.0003663003663003663, 39)
graph 6:
  時間與 cpu:
    0.45s user 0.04s system 92% cpu 0.981 total
```

結果: 由於數量眾多,只取前面常出現的數值,以下格式為:

(數值,出現次數)

(0.00015099077090119435, 21), (0.00013248091788613933, 21),

(0.00015361267397098017, 24), (0.0001452225841476656, 25),

(0.00013788125555226532, 28), (0.00014312506169183694, 29),

(0.00015511090429657206, 32), (0.00014983713355048862, 35),

(0.00014687063750581667, 39), (0.00014378053745928339, 43),

(0.0001425081433224756, 46), (0.00015224472454326583, 48),

(0.00014447830198591993, 48), (0.00014877223753445253, 51),

(0.00013903233494875664, 62)

**討論:** 相對於 Hits PageRank 的算法是比較全面性的,因為 Hits 只需要負責兩個網站之間的關係,但是 PageRank 則會牽扯到整個網路(會去抓指向該 node 的 node 的資訊)

## 4. SimRank

實作方式:

(1) 当
$$a = b$$
时, $s(a,b) = 1$ 

$$(2)$$
 当 $\mathcal{I}(a) = \emptyset$ 或者 $\mathcal{I}(b) = \emptyset$ 时, $s(a,b) = 0$ 

(3) 其他情况下,

$$s(a,b) = \frac{C}{|\mathcal{I}(a)| |\mathcal{I}(b)|} \sum_{i=1}^{|\mathcal{I}(a)|} \sum_{j=1}^{|\mathcal{I}(b)|} s(\mathcal{I}_i(a), \mathcal{I}_j(b))$$

 $\sum_{i=1}^{|\mathcal{I}(a)|} \sum_{j=1}^{|\mathcal{I}(b)|} s(\mathcal{I}_i(a), \mathcal{I}_j(b))$ 

意思為執行每個射入的 node 兩兩做 SimRank 並且加總

C為:06

搜尋深度為·10

## graph 1:

時間與 cpu: 0.45s user 0.12s system 20% cpu 2.814 total

執行結果:除了 node 相同的為 1 以外的 node 與 node 都為 0

討論:

由於 G 1 是一個由 1>2>3....>6 一條線的圖, 因此除了自己和自

己其他的射入點都為 0

## graph\_2:

時間與 cpu: 0.37s user 0.83s system 70% cpu 0.605 total

執行結果:除了 node 相同的為 1 以外的 node 與 node 都為 0

## graph\_3:

時間與 cpu: 7.88s user 0.11s system 90% cpu 8.607 total

執行結果:

除了點(1,3)(3,1)(2,4)(4,2)為 0.42856 \ node 相同的為 1 外,其

他都為 0

## graph\_4:

時間與 cpu: 0.35s user 0.06s system 95% cpu 0.432 total

執行結果:

```
      1 1 1.00000
      2 1 0.14265
      3 1 0.13302

      1 2 0.12686
      2 2 1.00000
      3 2 0.18629

      1 3 0.11582
      2 3 0.18811
      3 3 1.00000

      1 4 0.12226
      2 4 0.15064
      3 4 0.23882

      1 5 0.10949
      2 5 0.18843
      3 5 0.17004

      1 6 0.18525
      2 6 0.06686
      3 6 0.24574

      1 7 0.05927
      2 7 0.23441
      3 7 0.23191
```

4 1 0.13715	5 1 0.12743	6 1 0.19490	7 1 0.07940
4 2 0.14390	5 2 0.19681	6 2 0.04250	7 2 0.24530
4 3 0.23338	5 3 0.17861	6 3 0.22250	7 3 0.24427
4 4 1.00000	5 4 0.13604	6 4 0.31125	7 4 0.33181
4 5 0.12268	5 5 1.00000	6 5 0.04347	7 5 0.20189
4 6 0.33181	5 6 0.06712	6 6 1.00000	7 6 0.06361
4 7 0.31125	5 7 0.20497	6 7 0.02250	7 7 1.00000

## graph\_5:

時間與 cpu: 0.35s user 0.06s system 95% cpu 0.432 total

執行結果:除了 node 相同的為1以外的 node 與 node 都為0

### 討論:

對於 SimRank 的算法比起 Hits 更接近 PageRank,因為他們的關聯度都不單單只是兩個網頁,而是在整個網路上,而 SimRank 更是將跌代的概念用的更為貫徹.

# 5. 增加 node 1 的 hub, authority, and PageRank

# 1. graph\_1

- ► Hits:可以再增加一個點在 node1 的前面指向 node1 即可增加他的 hub 和 authority
- ▶ PageRank:可以拿掉 node6 即可

# 2. graph\_2

- ➤ Hits:只要再隨意找一個 node 並且將它指向 node1 即可
- ▶ PageRank: 只要再隨意找一個 node 並且將它指向 node1 即可

# **3. graph**\_3

- ➤ Hits:只要將 node4 指向 node1 即可
- ➤ PageRank: 只要將 node4 指向 node1 即可

# 6. 問題討論

- 1. More limitations about link analysis algo
  - ➤ 在這些演算法中都必須要設定最大深度才行,否則就會出現永遠跑不完的情況出現,而這個深度的其實也不是越多越好,在這些 dataSet 中常常只要跑個 3.4 次就可以知道有沒有循環的輪迴了。
- 2. What are practical issues when use in real web

➤ 在真實的 web 中應該沒辦法像我們的 dataset 一樣可以看到他的 完整有向圖,因為你在逛網頁的時候那個網頁通常只知道你去了哪裡,而沒辦法知道你是從哪裡來(in\_edges)

## 3. What do the result say for your actor/movie graph

➤ 我們可以從演算法裡面得知這個 graph 的哪個 node 有著比其他 node 都還要高的權重(重要性),比起人用肉眼看還要準確並且有明確性

### 4. What is the effect of 'C' in SimRank

➤ C的值影響性其實沒有我一開始想的還要大,由於在同一個圖中C的值基本上是一樣的,因此對其他 node 來說反而沒意義,除非拿來跟不同的圖的 node 做比較才會顯示出他的用意。不過在同一個圖中也會因為 C 所以不段的滾動某個較長的邊的 node 值

## 7. 心得

這次的 project 和前幾次的很不一樣,前面的都可以用肉眼稍微計算一次就可以得到答案,但這次的需要大量的運算,真的很佩服當初可以想出這些演算法的科學家們。而在這次的 project 中學到最大的應該就是圖形的應用了吧!原本還在思考要用怎麼樣的方式去把圖

給實做出來,但幸好後來找到了 networkx 讓圖這邊可以快速地找到 inoutedge,真的方便了許多。