

# Data Mining Project 3

P76061425 林聖軒

## Usage

### link\_analysis.py

```
$ python3 link_analysis.py [-h]
```

optional Options	Description
-h --help	show this help message and exit
-f GRAPH_FILE	graph file,(default="./hw3dataset/graph_1.txt")
-mode MODE	ha=HubsAuthorities, pr=PageRank, sr=SimRank, all=all above, (default=all)
-d D	PageRank d, (default=0.1)
-c C	SimRank c, (default=0.8)

## Implementation detail

三種演算法都寫在link\_analysis.py檔案中。

- HITS

依照投影片所寫的演算法，如下

```
HubsAuthorities(G)
1   $\mathbf{1} \leftarrow [1, \dots, 1] \in \mathbb{R}^{|V|}$ 
2   $\mathbf{a}_0 \leftarrow \mathbf{h}_0 \leftarrow \mathbf{1}$ 
3   $t \leftarrow 1$ 
4  repeat
5      for each  $v$  in  $V$ 
6          do  $a_t(v) \leftarrow \sum_{w \in pa[v]} h_{t-1}(w)$ 
7              $h_t(v) \leftarrow \sum_{w \in ch[v]} a_{t-1}(w)$ 
8              $\mathbf{a}_t \leftarrow \mathbf{a}_t / \|\mathbf{a}_t\|$ 
9              $\mathbf{h}_t \leftarrow \mathbf{h}_t / \|\mathbf{h}_t\|$  normalization
10             $t \leftarrow t + 1$ 
11 until  $\|\mathbf{a}_t - \mathbf{a}_{t-1}\| + \|\mathbf{h}_t - \mathbf{h}_{t-1}\| < \epsilon$ 
12 return  $(\mathbf{a}_t, \mathbf{h}_t)$ 
```

利用兩層for迴圈進行計算，第一層迴圈對每一個node迭代，第二層迴圈則計算單個node的authorites值和hub值，authorites用該node每個parent的hub值相加，hub則用該node每個chid的authorites值相加，再對所有的authorites值和hub值除以2norm來做normalization，一直迭代到authorites值和hub值前一次結果差值的2norm加總小於epsilon(這邊設為1e-10)則結束迭代。

- PageRank

同樣依照投影片所寫的公式計算，如下

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

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

對每個node做迭代，用上面的公式計算pageRank值，D值設定為0.1，並做2norm normalization，一直迭代到和上一次結果差值的2norm加總小於epsilon(這邊設為1e-10)則結束迭代。

- SimRank

$$S(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} S(I_i(a), I_j(b))$$

依照上面的公式定義，對每個點與其它所有的計算相似度，給定初始值後，用一個二維矩陣來做計算，對每個點迭代計算與其它所有不同點的結果，迭代到與上次誤差不大則結束迭代。

## Result analysis and discussion

以下呈現 graph 1~6的結果

### graph\_1.txt

- HITS

```
HubsAuthorities:
authorities:
best node: 2 value: 0.447213595499958
{1: 0.0, 2: 0.447213595499958, 3: 0.447213595499958, 4: 0.447213595499958, 5: 0.447213595499958, 6: 0.447213595499958}
hub:
best node: 1 value: 0.447213595499958
{1: 0.447213595499958, 2: 0.447213595499958, 3: 0.447213595499958, 4: 0.447213595499958, 5: 0.447213595499958, 6: 0.0}
```

- PageRank

```
PageRank:
best node: 6 value: 0.8098604354579024
{1: 0.02987343505244505, 2: 0.07806422962433922, 3: 0.15580395647032097, 4: 0.2812110011932284, 5: 0.4835133182936731, 6: 0.8098604354579024}
```

- SimRank

```
SimRank:
=====
1 simRank:
1 : 1.0
=====
2 simRank:
2 : 1.0
=====
3 simRank:
3 : 1.0
=====
4 simRank:
4 : 1.0
=====
5 simRank:
5 : 1.0
=====
6 simRank:
6 : 1.0
=====
```

### graph\_2.txt

- HITS

```
HubsAuthorities:
authorities:
best node: 1 value: 0.447213595499958
{1: 0.447213595499958, 2: 0.447213595499958, 3: 0.447213595499958, 4: 0.447213595499958, 5: 0.447213595499958}
hub:
best node: 1 value: 0.447213595499958
{1: 0.447213595499958, 2: 0.447213595499958, 3: 0.447213595499958, 4: 0.447213595499958, 5: 0.447213595499958}
```

- PageRank

```
PageRank:
best node: 1 value: 0.4472135954999579
{1: 0.4472135954999579, 2: 0.4472135954999579, 3: 0.4472135954999579, 4: 0.4472135954999579, 5: 0.4472135954999579}
```

- SimRank

```
SimRank:
=====
1 simRank:
 1 : 1.0
=====
2 simRank:
 2 : 1.0
=====
3 simRank:
 3 : 1.0
=====
4 simRank:
 4 : 1.0
=====
5 simRank:
 5 : 1.0
=====
```

## graph\_3.txt

- HITS

```
HubsAuthorities:
authorities:
best node: 2 value: 0.6015009550106639
{1: 0.37174803445513915, 2: 0.6015009550106639, 3: 0.6015009550106639, 4: 0.37174803445513915}
hub:
best node: 2 value: 0.6015009550106639
{1: 0.37174803445513915, 2: 0.6015009550106639, 3: 0.6015009550106639, 4: 0.37174803445513915}
```

- PageRank

```
PageRank:
best node: 2 value: 0.6288503045177238
{1: 0.3233377406180201, 2: 0.6288503045177238, 3: 0.6288503045177238, 4: 0.3233377406180201}
```

- SimRank

```
SimRank:
=====
1 simRank:
 1 : 1.0
 3 : 0.666665950838784
=====
2 simRank:
 2 : 1.0
 4 : 0.6666663803355136
=====
3 simRank:
 1 : 0.666665950838784
 3 : 1.0
=====
4 simRank:
 2 : 0.6666663803355136
 4 : 1.0
=====
```

## graph\_4.txt

- HITS

```
HubsAuthorities:
authorities:
best node: 5 value: 0.500635020035182
{1: 0.34668186714993793, 2: 0.44219353423699814, 3: 0.49913837843536446, 4: 0.34840643183576686, 5: 0.500635020035182, 7: 0.20899872237286285, 6: 0.13940770946290396}
hub:
best node: 1 value: 0.6464257201947676
{1: 0.6464257201947676, 2: 0.11208722834189842, 3: 0.2550547508483144, 4: 0.4662086257344565, 5: 0.43118315727820944, 7: 0.1618624944799001, 6: 0.27394972282179847}
```

- PageRank

```
PageRank:
best node: 1 value: 0.6897307449446073
{1: 0.6897307449446073, 2: 0.3809396414535482, 3: 0.32583077962000323, 4: 0.24718966745450782, 5: 0.4206993927446729, 7: 0.14680844458168535, 6: 0.11553078354758817}
```

- ## 部分結果

```

SimRank:
=====
1  simRank:
  1 : 1.0
  2 : 0.36026281492356466
  3 : 0.34895923549390523
  4 : 0.353732790334621
  5 : 0.3376569689372631
  7 : 0.292390591672489
  6 : 0.4150753327709355
=====
2  simRank:
  1 : 0.36026281492356466
  2 : 1.0
  3 : 0.4067901545437175
  4 : 0.3697456432140367
  5 : 0.4121804875460285
  7 : 0.45405120688760703
  6 : 0.28544007954046635
=====
3  simRank:
  1 : 0.34895923549390523
  2 : 0.4067901545437175
  3 : 1.0
  4 : 0.4495651492700736
  5 : 0.39005261515889406
  7 : 0.4510372691391691
  6 : 0.4480930294009781
=====
4  simRank:
  1 : 0.353732790334621

```

graph\_5.txt

- HITS

```
HubsAuthorities:
authorities:
best node: 61 value: 0.4913507493678007
(1: 0.0, 8: 5.0551588352906e-21, 11: 5.0551588352906e-21, 168: 9.581214346542798e-21,
37547726e-21, 264: 9.974639126503919e-21, 307: 8.624809002850715e-20, 2: 0.0, 9: 1.0,
-36, 13: 1.0303114758841021e-36, 14: 1.0303114758841021e-36, 3: 0.0, 6: 8.0538964851e-
1.6711069155232595e-24, 235: 2.45716118411437983e-24, 296: 2.4716118411437983e-24,
47607530715013e-13, 136: 1.946695081312002e-13, 217: 1.3527743382664407e-13, 265: 4.0,
-13, 300: 3.2806262036016433e-13, 344: 1.3527743382664407e-13, 351: 1.3527743382664407e-13,
67467060311323e-13, 457: 3.806972617554789e-13, 5: 0.0, 7: 1.5856304994682065e-12,
82065e-12, 46: 1.6297473212007835e-11, 187: 1.6297473212007835e-11, 191: 1.35315651e-11,
: 3.3146008735269256e-12, 436: 1.802644358606656e-11, 444: 1.802644358606656e-11, 200: 1.802644358606656e-11, 201: 1.802644358606656e-11, 202: 1.802644358606656e-11, 203: 1.802644358606656e-11, 204: 1.802644358606656e-11, 205: 1.802644358606656e-11, 206: 1.802644358606656e-11, 207: 1.802644358606656e-11, 208: 1.802644358606656e-11, 209: 1.802644358606656e-11, 210: 1.802644358606656e-11, 211: 1.802644358606656e-11, 212: 1.802644358606656e-11, 213: 1.802644358606656e-11, 214: 1.802644358606656e-11, 215: 1.802644358606656e-11, 216: 1.802644358606656e-11, 217: 1.802644358606656e-11, 218: 1.802644358606656e-11, 219: 1.802644358606656e-11, 220: 1.802644358606656e-11, 221: 1.802644358606656e-11, 222: 1.802644358606656e-11, 223: 1.802644358606656e-11, 224: 1.802644358606656e-11, 225: 1.802644358606656e-11, 226: 1.802644358606656e-11, 227: 1.802644358606656e-11, 228: 1.802644358606656e-11, 229: 1.802644358606656e-11, 230: 1.802644358606656e-11, 231: 1.802644358606656e-11, 232: 1.802644358606656e-11, 233: 1.802644358606656e-11, 234: 1.802644358606656e-11, 235: 1.802644358606656e-11, 236: 1.802644358606656e-11, 237: 1.802644358606656e-11, 238: 1.802644358606656e-11, 239: 1.802644358606656e-11, 240: 1.802644358606656e-11, 241: 1.802644358606656e-11, 242: 1.802644358606656e-11, 243: 1.802644358606656e-11, 244: 1.802644358606656e-11, 245: 1.802644358606656e-11, 246: 1.802644358606656e-11, 247: 1.802644358606656e-11, 248: 1.802644358606656e-11, 249: 1.802644358606656e-11, 250: 1.802644358606656e-11, 251: 1.802644358606656e-11, 252: 1.802644358606656e-11, 253: 1.802644358606656e-11, 254: 1.802644358606656e-11, 255: 1.802644358606656e-11, 256: 1.802644358606656e-11, 257: 1.802644358606656e-11, 258: 1.802644358606656e-11, 259: 1.802644358606656e-11, 260: 1.802644358606656e-11, 261: 1.802644358606656e-11, 262: 1.802644358606656e-11, 263: 1.802644358606656e-11, 264: 1.802644358606656e-11, 265: 1.802644358606656e-11, 266: 1.802644358606656e-11, 267: 1.802644358606656e-11, 268: 1.802644358606656e-11, 269: 1.802644358606656e-11, 270: 1.802644358606656e-11, 271: 1.802644358606656e-11, 272: 1.802644358606656e-11, 273: 1.802644358606656e-11, 274: 1.802644358606656e-11, 275: 1.802644358606656e-11, 276: 1.802644358606656e-11, 277: 1.802644358606656e-11, 278: 1.802644358606656e-11, 279: 1.802644358606656e-11, 280: 1.802644358606656e-11, 281: 1.802644358606656e-11, 282: 1.802644358606656e-11, 283: 1.802644358606656e-11, 284: 1.802644358606656e-11, 285: 1.802644358606656e-11, 286: 1.802644358606656e-11, 287: 1.802644358606656e-11, 288: 1.802644358606656e-11, 289: 1.802644358606656e-11, 290: 1.802644358606656e-11, 291: 1.802644358606656e-11, 292: 1.802644358606656e-11, 293: 1.802644358606656e-11, 294: 1.802644358606656e-11, 295: 1.802644358606656e-11, 296: 1.802644358606656e-11, 297: 1.802644358606656e-11, 298: 1.802644358606656e-11, 299: 1.802644358606656e-11, 300: 1.802644358606656e-11, 301: 1.802644358606656e-11, 302: 1.802644358606656e-11, 303: 1.802644358606656e-11, 304: 1.802644358606656e-11, 305: 1.802644358606656e-11, 306: 1.802644358606656e-11, 307: 1.802644358606656e-11, 308: 1.802644358606656e-11, 309: 1.802644358606656e-11, 310: 1.802644358606656e-11, 311: 1.802644358606656e-11, 312: 1.802644358606656e-11, 313: 1.802644358606656e-11, 314: 1.802644358606656e-11, 315: 1.802644358606656e-11, 316: 1.802644358606656e-11, 317: 1.
```

- PageRank

[illegible]



- SimRank

部分結果

```
SimRank:
=====
1 simRank:
  1 : 1.0
=====
8 simRank:
  8 : 1.0
 11 : 0.8
 168 : 0.4
 227 : 0.4
 253 : 0.2666666666666666
 264 : 0.2666666666666666
 307 : 0.16
=====
11 simRank:
  8 : 0.8
 11 : 1.0
 168 : 0.4
 227 : 0.4
 253 : 0.2666666666666666
 264 : 0.2666666666666666
 307 : 0.16
=====
168 simRank:
  8 : 0.4
 11 : 0.4
 168 : 1.0
 227 : 0.2
 253 : 0.1333333333333333
 264 : 0.2666666666666666
 307 : 0.16
 389 : 0.4
```

graph\_6.txt

- HITS

部分結果

```
HubsAuthorities:
authorities:
best node: 761 value: 0.27506602043419465
{1: 0.0, 6: 0.010177770169822317, 68: 0.02752099908896793, 95: 0.03114541
06947384117, 273: 0.04062637731637039, 298: 0.02958980196848592, 367: 0.0
4419376493018139, 387: 0.04419376493018139, 410: 0.045262460483570105, 41
54: 0.04403932924876352, 578: 0.0399288443048292, 635: 0.0178936096637067
211, 747: 0.0009057381270866483, 748: 0.006931605055739387, 848: 0.040047
301437860955, 897: 0.02155625697145912, 946: 0.04265502389105189, 951: 0
hub:
best node: 171 value: 0.15626346514487824
{1: 0.0260417167399385, 6: 0.0, 68: 0.03914212350351502, 95: 0.037128520671219314, 14
.037614247440977015, 298: 0.0, 367: 0.0, 374: 0.0, 387: 0.03072705680926536, 410: 1.5
2e-74, 415: 0.0, 554: 0.0351673802867363, 578: 0.03528431067031816, 635: 0.0, 725: 0.
48: 0.04053340686801797, 848: 0.039107393846534715, 856: 0.0, 897: 0.0398451257615606
9738457967447, 951: 0.0, 955: 0.039398143730637436, 1021: 0.0, 1058: 0.03645904182603
3838068154698983, 7: 0.09528565756995332, 62: 0.08726049011433283, 78: 0.096924841834
```

- PageRank

部分結果

```
PageRank:
best node: 410 value: 0.2279418820340828
{1: 0.00018047193682610248, 6: 0.05396533804586473, 68: 0.11911735603844277, 95: 0.1426250909284221,
73: 0.1886196014896138, 298: 0.1395005825535572, 367: 0.17570642723469643, 374: 0.2141624240324738, 3
: 0.2279418820340828, 415: 0.2045476854291000, 554: 0.21170132976951433, 578: 0.1836290376121718, 635
0.03452580516612473, 747: 0.00019431674321223057, 748: 0.04495102564194828, 848: 0.1745598827193314,
897: 0.0958086151423427, 946: 0.19214378561638826, 951: 0.00019431674321223057, 955: 0.12956464480286
8, 1058: 0.030668136656590024, 1084: 0.1975317203837002, 7: 0.0005411799338277944, 62: 0.004384573558
0685, 180: 0.002909838056413321, 225: 0.002346068304347421, 370: 0.002288025616399507, 394: 0.0045756
663590261634, 501: 0.003945987710044221, 528: 0.001991467324902724, 609: 0.0005164226215504499, 761:
0.00024043996060773334, 1003: 0.000673941768814534, 1089: 0.001180104909088555, 1121: 0.0005869490199
51108, 1151: 0.004698752266386397, 1227: 0.0027639056861214276, 8: 0.0002463189198311188, 79: 0.00071
21479268229111, 139: 0.0004293056618189392, 202: 0.0006732443700952347, 386: 0.0020154684509788474, 5
```

- SimRank

部分結果

```
SimRank:
=====
1 simRank:
  1 : 1.0
=====
6 simRank:
  6 : 1.0
  68 : 0.09757035435557052
  95 : 0.09878742351777717
  142 : 0.09508440747535006
  273 : 0.08433220277442827
  298 : 0.09492760507977256
  367 : 0.07334497517895733
  374 : 0.0878705223514029
  387 : 0.0878705223514029
  410 : 0.089067359878046
  415 : 0.08460904358169692
  554 : 0.08968576721302333
  578 : 0.08250744778924138
  635 : 0.10725380944809841
  725 : 0.12745661183132975
  747 : 0.07272727272727274
  748 : 0.1284625826063654
  848 : 0.08189854588433737
  856 : 0.0920092627450845
  887 : 0.08636658181417668
```

discussion

- 透過上面呈現的結果可以觀察到，像圖1這種直接從1連續連連到5也沒有cycle的圖，authorities會在起始node(0)的位置值為0，因為沒有父節點可以計算出值，hub則是會在結束點(6)位置為0，因為沒有子節點能夠計算出值，而PageRank則會在起始點比較低。
- 在實作SimRank的過程中，發現若依照遞迴式直接coding，在遇到有cycle的圖片時會無法結束，所以會用給予每個node對應其他node的相似度初始值，再依照公式計算，直到誤差夠小就結束迭代的這種計算方式來實作此演算法。

Computation performance analysis

HITS

- time

graph	time
graph_1	0m0.091s
graph_2	0m0.092s
graph_3	0m0.091s
graph_4	0m0.093s
graph_5	0m0.129s
graph_6	0m0.813s

PageRank

- time

graph	time
graph_1	0m0.090s
graph_2	0m0.091s
graph_3	0m0.090s
graph_4	0m0.091s
graph_5	0m0.157s
graph_6	0m0.322s

SimRank

- time

graph	time
graph_1	0m0.092s
graph_2	0m0.092s
graph_3	0m0.092s
graph_4	0m0.096s
graph_5	0m9.897s
graph_6	0m39.897s

## analysis

- 上面的執行時間結果可以觀察到，在圖1~3這種很小的圖時間差距不大，但到圖4這種開始有點複雜度的圖效能差距就慢慢出來了，而圖5、6則可以看到HITS的時間明顯少於其他兩種演算法，PageRank則次之，而SimRank在點數多圖複雜時會很明顯的要花非常多時間。

## Discussion

- 在這個project中要我們實作HITS、PageRank及SimRank三種不同的演算法，此三種方法概念上略有一些差異，但都對搜尋引擎有很大的幫助，可以應用於含有元素之間相互參照的情況，而且不只是一是要考慮經度問題，還要將計算的時間複雜度考量進去，因此在寫程式時上網搜尋作法也會發現一些演算法變體。

## Find a way (e.g., add/delete some links) to increase hub, authority, and PageRank of Node 1 in first 3 graphs respectively

- hub的計算方式是child node的authority值相加出來的，所以若要增加hub，以圖1為例，要增加結束點6(無child或少child)之node的child link數，或是增加影響權重，圖2及圖3也是同理。
- authority的方法也類似，authority的計算方法是parent node的hub值相加出來的，因此要增加authority擇要增加起始點(無或少parent)之node的parent link數，，或是增加影響權重，圖2及圖3也是依此類推。

## Questions & Discussion

### More limitations about link analysis algorithms

- 大部分的演算法，都沒有辦法在圖中很好的找到每個node之間最佳的相關性，評分的標準只用連結來判定可能有些不足，連結數多寡的可能有太多變因，網頁質量和連結數其實相關性是不太足夠的。

### Can link analysis algorithms really find the “important” pages from Web?

- 如上題所述，沒有辦法找到很好的important pages，在實際情況中的連結可能也有很多相干度不高的網頁，甚至是廣告蓋版的問題等等，更舊的網頁分數也會因為演算反可能分數高，但實際重要程度可能不及新網頁的質量。

### What are practical issues when implement these algorithms in a real Web?

- 最常見的就是用在搜尋引擎，做網頁排名，像PageRank是google早期用來對搜尋引擎的搜尋結果中做網頁排名的演算法，而像google這種資料量如此龐大的公司，不僅僅是要考量到演算法的精準度，還要顧及時間複雜度不能夠太高，以免造成效能不佳導致使用者體驗不好的問題，因此也有了許多的演算法變體。

### Any new idea about the link analysis algorithm?

- 可能可以多考慮幾層的關係而不只是一層，但時間複雜度也要有所取捨，或為不同的網頁判斷不同的權重，不然就是加入一些使用者偏好的因素在裡面，如瀏覽紀錄或書籤網站等等，藉此來設計新的演算法。

### What is the effect of “C” parameter in SimRank?

- C在SimRank的演算法中代表著阻尼常數，有衰退的效用，較近的共同父節點有比較強的影響力，而比較遠的會因為此係數的關係影響遞減。