Data Mining Project 3

P76061425 林聖軒

Usage

link_analysis.py

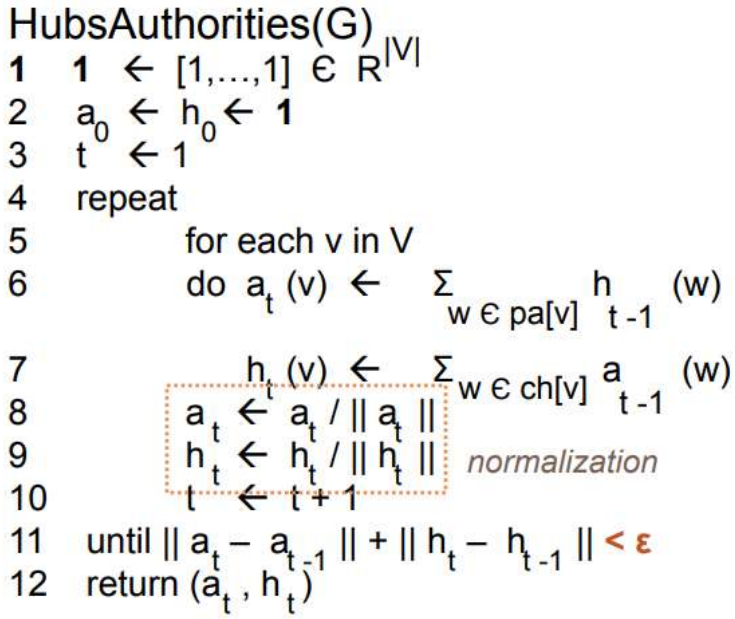
```
$ python3 link_analysis.py [-h]
```

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



  利用兩層for迴圈進行計算，第一層回圈對每一個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

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

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。

- 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: 0.8098
604354579024}
```

- 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: 0.442193534236998l4, 3: 0.49913837843536446, 4: 0.34840643183576686, 5: 0.500635020035182, 7: 0.2089
9872237286285, 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: 0.16186
24944799001, 6: 0.2739497228217984 7}
```

- PageRank

```
PageRank:
best node: 1 value: 0.6897307449446073
{1: 0.6897307449446073, 2: 0.3809396414535482, 3: 0.32583077962000323, 4: 0.24718966745450782, 5: 0.4206993927446729, 7: 0.14680
844458168535, 6: 0.11553078354758817}
```

- SimRank
  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: 8.0538964851
: 1.6711069155232595e-24, 235: 2.4716118411437983e-24, 296: 2.4716118411437983e-24,
47607530715013e-13, 136: 1.946695083121902e-13, 217: 1.3527743382664407e-13, 265: 4.
-13, 300: 3.2806262036016433e-13, 344: 1.3527743382664407e-13, 351: 1.3527743382664
67467060311323e-13, 457: 3.806972617554789e-13, 5: 0.0, 7: 1.5856304994682065e-12, 1
82065e-12, 46: 1.6297473212007835e-11, 187: 1.6297473212007835e-11, 191: 1.353156518
: 3.3146008735269256e-12, 436: 1.8026443586066656e-11, 444: 1.8026443586066656e-11, 26
```

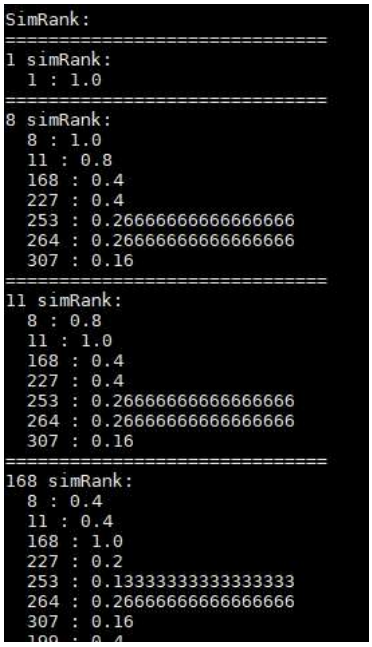
```
hub:
best node: 274 value: 0.19194388446558547
{1: 1.2318679317390329e-20, 8: 0.0, 11: 1.1064176421022567e-20, 168: 0.0, 227: 3.5154175655997557e-
, 264: 5.792616904009446e-25, 307: 5.421469103985068e-20, 2: 7.769692186900134e-37, 9: 0.0, 10: 0.0
 14: 0.0, 3: 1.3968880625423045e-24, 6: 1.5015218013521525e-24, 219: 1.7666234695980047e-49, 223: 1.
8479505468704896e-39, 296: 0.0, 336: 0.0, 4: 3.2942403866017905e-13, 40: 0.0, 136: 0.0, 217: 0.0, 2
87: 0.0, 300: 3.0488515835869463e-13, 344: 3.095502310032329e-13, 351: 2.7283037668010906e-13, 363:
 2.6650010870749144e-13, 457: 0.0, 5: 6.625657820428537e-12, 7: 1.1193244187505134e-11, 12: 1.05502
 0.0, 187: 0.0, 191: 0.0, 244: 0.0, 306: 1.0884517999544842e-11, 436: 0.0, 444: 0.0, 26: 0.0, 448:
.70487794251448 36e-43, 143: 3.70487794251448 36e-43, 256: 0.0, 258: 9.466163877704537e-12, 380: 1.00
 449: 0.0, 459: 0.0, 468: 4.56563031613582e-16, 199: 4.2245029591656 61e-21, 414: 4.6848632685780 41e
```

- PageRank

```
PageRank:
best node: 96 value: 0.42529688179019226
{1: 0.0003838391336762 1503, 8: 0.0004726806194426 4985, 11: 0.0004726806194426 4985, 168: 0.000625846627741
317253511, 253: 0.000911008576494035, 264: 0.001563429435756411, 307: 0.0010755521075624286, 2: 0.0003838
539311733767476, 10: 0.0005393111733767476, 13: 0.0005393111733767476, 14: 0.0005393111733767476, 3: 0.0003
0048748753373705574, 219: 0.0007392657830702924, 223: 0.0006003189596915332, 235: 0.0007392657830702924,
924, 336: 0.0007392657830702924, 4: 0.0003838391336762 1503, 40: 0.0035258921769683 48, 136: 0.0010728327 60
644683477815, 265: 0.0023321703465915993, 287: 0.0035258921769683 48, 300: 0.0018901087898008544, 344: 0.0
: 0.0009091644683477815, 363: 0.0028934340409779115, 454: 0.002863192058449047, 457: 0.0025198666922375 48
21503, 7: 0.0004460281737121 2719 45, 12: 0.0004460281737121 2719 45, 15: 0.0004460281737121 2719 45, 46: 0.0009767 05
767050856315414, 191: 0.0008585478053000892, 244: 0.0007033368062918031, 306: 0.0005496214477813752, 436
 444: 0.0010802983597001973, 26: 0.0004966705596306926, 448: 0.0004966705596306926, 124: 0.0005595354810 9
4130159532, 256: 0.0007608006870840 09, 258: 0.000663128755167949, 380: 0.0007971100673166266, 398: 0.0007
.0004354568586158453, 459: 0.0009145160455950368, 468: 0.0010181093196636925, 199: 0.0005370051419748751
8751, 60: 0.0008424129722406145, 162: 0.0008207315198459613, 469: 0.0008207315198459613, 27: 0.0005227859
41860510113236, 86: 0.0004284186051011 3236, 107: 0.0004284186051011 3236, 180: 0.0004284186051011 3236, 19
6, 239: 0.0004284186051011 3236, 271: 0.0004284186051011 3236, 292: 0.0004284186051011 3236, 305: 0.0004284
042841860510113236, 364: 0.0004284186051011 3236, 370: 0.0004284186051011 3236, 397: 0.0004284186051011 3230
113236, 437: 0.0004284186051011 3236, 446: 0.0004284186051011 3236, 447: 0.0004284186051011 3236, 462: 0.000
```

- SimRank

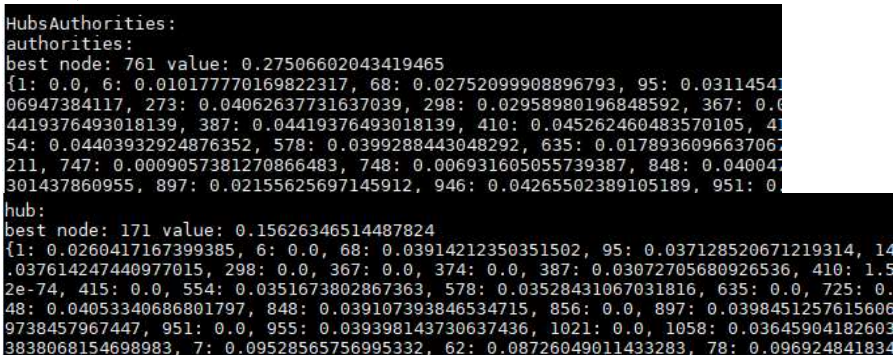
  部分結果



graph_6.txt

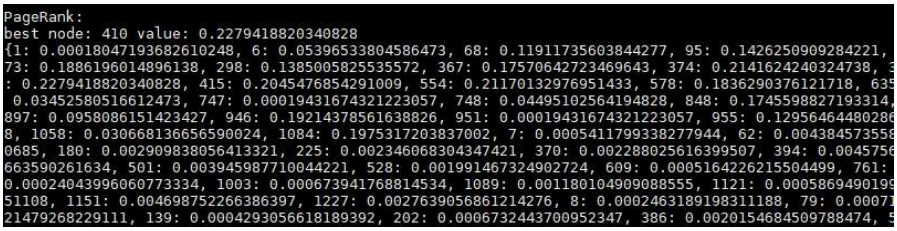
- HITS

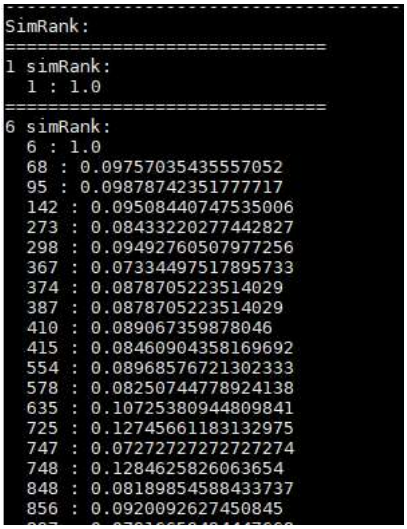
  部分結果



- PageRank

  部分結果



- SimRank
  部分結果


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束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。