Data Mining Project 3

<u>Link Analysis Practice</u>

資訊所

P76071349

王鈺云

2018/12/25

- Introduction

　　此次作業是實作三種演算法，分別是HITS、PageRank和SimRank，其中HITS和PageRank是搜索引擎中兩個最基礎和最重要的算法，基本思想都是首先利用網頁之間的link關係構建網頁的網絡圖，通過相鄰節點之間的網頁的重要性，計算出某個網頁的重要性，而SimRank則是用來衡量任意兩個對象間的相似程度。

- Dataset

　　利用project3dataset裡的graph_1~graph_6與project 1中所使用IBM Generator所產生的資料來實作三種演算法，其中資料1,2形式表示有node 1連至node 2之link存在，而我用的IBM Generator產生的資料中node數為10，link數為76。

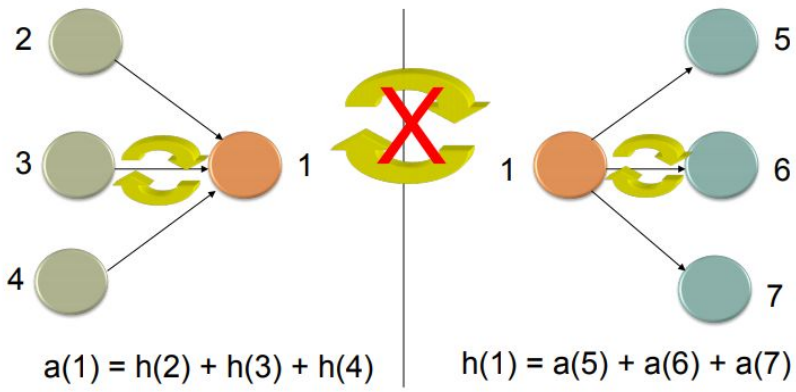
- Implementation and Result analysis

  1. HITS

     Hits算法的基本思想

     A. 好的Hub型網頁指向好的Authority網頁
     B. 好的Authority網頁是由好的Hub型網頁所指向的網頁。

     首先將graph中node之間的link存成dictionary中key與value的關係，例如1,2表示node1到node2之間有link存在，dictionary就存為{1:[2]}，令所有node的初始值(hub,autoiruty) = (1,1)，接著以下圖的方式不斷更新所有node之hub值與authority值，並且每回合皆normalize，不斷重複，直到hub值與authority值皆沒有太大的變化，即為結果。



     以graph_3為例，threshold設定為0.0001，跑了7輪後之結果為下圖。

```
file name: graph_3.txt
number of itrations: 7
time cost: 0.0006868839263916016 s
hub 1 : 0.0827186512118019
hub 2 : 0.21654373024236037
hub 3 : 0.35036880927291886
hub 4 : 0.35036880927291886
authority 1 : 0.0827186512118019
authority 2 : 0.21654373024236037
authority 3 : 0.35036880927291886
authority 4 : 0.35036880927291886
--------------------------------------
```

2. PageRank

PageRank以下圖之公式實作，將graph裡node與node間的link關係存為nxn矩陣型式，n為node數，例如1,2表示node1到node2之間有link存在，矩陣中第0列，第1行就存為1，沒有link存在就存為0，之後將這個矩陣normalize，使得每列之和為1，再將其轉置，成為公式中所需的轉移矩陣，接著以初始pr矩陣不斷迭代，且過程中每回合需將新的pr矩陣normalize，不斷重複，直到pr矩陣皆沒有太大的變化，即為結果。

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

以IBM Generator data為例，有10個node，76條link，d設定為0.15，threshold設定為0.01，跑了4輪後之結果為下圖。

```
file name: project1_data.txt
number of iterations: 4
time cost: 0.0014965534210205078 s
pagerank:
[[0.0376319037]
 [0.0185073415]
 [0.1118535759]
 [0.0990106542]
 [0.1369006687]
 [0.1369006687]
 [0.1217382718]
 [0.1184439522]
 [0.1369006687]
 [0.0821122946]]
--------------------------------------
```

當threshold設定為0.001，跑了6輪後之結果為下圖。

```
file name: project1_data.txt
number of iterations: 6
time cost: 0.0022106170654296875 s
pagerank:
[[0.0373852482]
 [0.0186659781]
 [0.1120934311]
 [0.0987858309]
 [0.1370162965]
 [0.1370162965]
 [0.1217954248]
 [0.1183514574]
 [0.1370162965]
 [0.0818737401]]
-------------------------------------
```

3. SimRank

   SimRank以下圖之公式實作，利用node間的in-neighbor關係遞迴求得點與點之間的相似程度，首先將graph中node的in-neighbor存成dictionary中key與value的關係，例如1,2表示node2有一條in-neighbor的link存在，dictionary就存為{2:[1]}，以遞迴方式寫成function，不斷重複，直到S(a,a) = 1為遞迴中止條件，其中，有些點與點之間的SimRank會陷入無限遞迴，所以我設定了SimRank遞迴如果超過500次，就直接令SimRank為0，

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

   以graph_3為例，c設定為0.6之結果為下圖。

```
file name: graph_3.txt
in neighbors graph:
{1: [2], 2: [1, 3], 3: [2, 4], 4: [3]}
S( 1 , 1 ) =  1
S( 1 , 2 ) =  0.0
S( 1 , 3 ) =  0.428571428571428 5
S( 1 , 4 ) =  0.0
S( 2 , 1 ) =  0.0
S( 2 , 2 ) =  1
S( 2 , 3 ) =  0.0
S( 2 , 4 ) =  0.4285714285714285
S( 3 , 1 ) =  0.4285714285714285
S( 3 , 2 ) =  0.0
S( 3 , 3 ) =  1
S( 3 , 4 ) =  0.0
S( 4 , 1 ) =  0.0
S( 4 , 2 ) =  0.4285714285714285
S( 4 , 3 ) =  0.0
S( 4 , 4 ) =  1
time cost: 0.008866548538208008 s
--------------------------------------
```

以graph_3為例，c設定為0.8之結果為下圖。

```
file name: graph_3.txt
in neighbors graph:
{1: [2], 2: [1, 3], 3: [2, 4], 4: [3]}
S( 1 , 1 ) =  1
S( 1 , 2 ) =  0.0
S( 1 , 3 ) =  0.6666666666666666
S( 1 , 4 ) =  0.0
S( 2 , 1 ) =  0.0
S( 2 , 2 ) =  1
S( 2 , 3 ) =  0.0
S( 2 , 4 ) =  0.6666666666666666
S( 3 , 1 ) =  0.6666666666666666
S( 3 , 2 ) =  0.0
S( 3 , 3 ) =  1
S( 3 , 4 ) =  0.0
S( 4 , 1 ) =  0.0
S( 4 , 2 ) =  0.6666666666666666
S( 4 , 3 ) =  0.0
S( 4 , 4 ) =  1
time cost: 0.030653953552246094 s
---------------------------------------
```

- Computation performance analysis

整體來說，node與link越多，會造成執行時間越長，因遞迴關係，SimRank所需執行時間最長，例如SimRank的graph_5之執行時間為155秒。

- HITS (threshold = 0.0001)

graph	iteration	time cost
graph_1	2	0.00125 s
graph_2	2	0.00068 s
graph_3	7	0.00120 s
graph_4	9	0.00120 s
graph_5	4	0.19071 s
graph_6	6	2.34103 s
IBM Generator data	4	0.00219 s

- PageRank (threshold = 0.01 , d = 0.15)

graph	iteration	time cost
graph_1	18	0.00178 s
graph_2	2	0.00093 s
graph_3	4	0.00133 s
graph_4	5	0.00085 s
graph_5	25	0.16013 s
graph_6	25	0.78340 s
IBM Generator data	4	0.00149 s

- SimRank (c = 0.6)

graph	time cost
graph_1	0.01611 s
graph_2	0.02623 s
graph_3	0.00886 s
graph_4	0.06445 s
graph_5	155.56749 s

- Discussion

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

  A : HITS從機制上很容易作弊，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，之後再將這個網頁鏈接指向作弊網頁，就可以提升作弊網頁的Authority得分。HITS算法的計算對象數量較少，只需計算相關node之間的鏈接關係；而PageRank比較算是全局性算法，需對所有node進行處理。HITS算法在計算時，對於每個node需要計算兩個值，所需執行時間較長，而PageRank只需計算一個值即可。

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

  - 提升Node1 hub方法

    增加node 1出發的link

    1. graph_1

    ```
    file name: graph_1.txt
    number of itrations: 2
    time cost: 0.0012004375457763672 s
    hub 1 : 0.2
    hub 2 : 0.2
    hub 3 : 0.2
    hub 4 : 0.2
    hub 5 : 0.2
    hub 6 : 0.0
    ------------------------------------
    file name: graph_1_newh.txt
    number of itrations: 7
    time cost: 0.003598928451538086 s
    hub 1 : 0.5582401843260578
    hub 2 : 0.14724289317184047
    hub 3 : 0.14724289317184047
    hub 4 : 0.14724289317184047
    hub 5 : 3.1136158420774045e-05
    hub 6 : 0.0
    ------------------------------------
    ```

    2. graph_2

```
file name: graph_2.txt
number of itrations: 2
time cost: 0.0019352436065673828 s
hub 1 : 0.2
hub 2 : 0.2
hub 3 : 0.2
hub 4 : 0.2
hub 5 : 0.2
-----------------------------------
file name: graph_2_newh.txt
number of itrations: 7
time cost: 0.0017149448394775390 s
hub 1 : 0.5582401843260578
hub 2 : 0.14724289317184047
hub 3 : 0.14724289317184047
hub 4 : 0.14724289317184047
hub 5 : 3.1136158420774045e-05
-----------------------------------
```

3. graph_3

```
file name: graph_3.txt
number of itrations: 7
time cost: 0.0004868507385253906 s
hub 1 : 0.0827186512118019
hub 2 : 0.21654373024236037
hub 3 : 0.35036880927291886
hub 4 : 0.35036880927291886
-----------------------------------
file name: graph_3_newh.txt
number of itrations: 11
time cost: 0.0013856887817382812 s
hub 1 : 0.20959642525491046
hub 2 : 0.07190282514356575
hub 3 : 0.3776264776238755
hub 4 : 0.3408742719776483
-----------------------------------
```

- 提升Node1 authority方法

  增加連到node 1的link

  1. graph_1

```
file name: graph_1.txt
number of itrations: 2
time cost: 0.0015964508056640625 s
authority 1 : 0.0
authority 2 : 0.2
authority 3 : 0.2
authority 4 : 0.2
authority 5 : 0.2
authority 6 : 0.2
-----------------------------------
file name: graph_1_newp.txt
number of itrations: 7
time cost: 0.003537893295288086 s
authority 1 : 0.16666543210791032
authority 2 : 7.407352538129347e-06
authority 3 : 0.2083317901348879
authority 4 : 0.2083317901348879
authority 5 : 0.2083317901348879
authority 6 : 0.2083317901348879
-----------------------------------
```

2. graph_2

```
file name: graph_2.txt
number of itrations: 2
time cost: 0.001383066177368164 s
authority 1 : 0.2
authority 2 : 0.2
authority 3 : 0.2
authority 4 : 0.2
authority 5 : 0.2
-----------------------------------
file name: graph_2_newp.txt
number of itrations: 7
time cost: 0.0018014907836914062 s
authority 1 : 0.20870947886746
authority 2 : 1.2823143208863356e-05
authority 3 : 0.26375923266311035
authority 4 : 0.26375923266311035
authority 5 : 0.26375923266311035
-----------------------------------
```

3. graph_3

```
file name: graph_3.txt
number of itrations: 7
time cost: 0.0009329319000244141 s
authority 1 : 0.0827186512118019
authority 2 : 0.21654373024236037
authority 3 : 0.35036880927291886
authority 4 : 0.35036880927291886
------------------------------------
file name: graph_3_newp.txt
number of itrations: 12
time cost: 0.0009181499481201172 s
authority 1 : 0.2096012868138617
authority 2 : 0.07189173630511461
authority 3 : 0.3776452449719405
authority 4 : 0.34086173190908314
------------------------------------
```

- 提升Node1 PageRank方法

  增加連到node 1的link

  1. graph_1

```
file name: graph_1.txt
number of iterations: 18
time cost: 0.002389192581176758 s
pagerank:
[[0.0000000176]
 [0.0000018121]
 [0.000088249 ]
 [0.0027005621]
 [0.0582122171]
 [0.9389971421]]
------------------------------------
file name: graph_1_newp.txt
number of iterations: 7
time cost: 0.002185344696044922 s
pagerank:
[[0.3307139244]
 [0.3385708647]
 [0.1725722853]
 [0.0886228465]
 [0.0461304347]
 [0.0233896444]]
------------------------------------
```

  2. graph_2

```
file name: graph_2.txt
number of iterations: 2
time cost: 0.0038373470306396484 s
pagerank:
[[0.2]
 [0.2]
 [0.2]
 [0.2]
 [0.2]]
------------------------------------
file name: graph_2_newp.txt
number of iterations: 7
time cost: 0.0015432834625244l4 s
pagerank:
[[0.3481524237]
 [0.347922213 ]
 [0.1718536595]
 [0.087212903 ]
 [0.0448588008]]
------------------------------------
```

3. graph_3

```
file name: graph_3.txt
number of iterations: 4
time cost: 0.0011019706726074219 s
pagerank:
[[0.1671432617]
 [0.3328567383]
 [0.3328567383]
 [0.1671432617]]
------------------------------------
file name: graph_3_newp.txt
number of iterations: 6
time cost: 0.0012438297271728516 s
pagerank:
[[0.3072297668]
 [0.3855376188]
 [0.2299680865]
 [0.077264528 ]]
------------------------------------
```

- Conclusion

HITS算法在海量網頁中找到與用戶查詢主題相關的高質量“Authority”頁面和“Hub”頁面，尤其是“Authority”頁面，因為這些頁面代表了能夠滿足用戶查詢的高質量內容，搜索引擎以此作為搜索結果返回給用戶。

PageRank算法剛開始賦予每個網頁相同的重要性得分，通過迭代遞歸計算來更新每個頁面節點的PageRank得分，直到收斂穩定為止，由此方法可得知網頁的重要程度。

HITS與PageRank最大的差別在於PageRank算法與使用者輸入的關鍵字無關，HITS算法與使用者輸入的查詢請求有密切關係。由於HITS算法是與查詢相關的算法，所以必須在接收到用戶查詢後實時進行計算，而HITS算法本身需要進行很多輪迭代計算才能獲得最終結果，這導致其計算效率較低，這是實際應用時必須慎重考慮的問題。

SimRank完全基於圖形理論，可以計算圖中任意兩個節點間的相似度。

- Referance

https://blog.csdn.net/zealfory/article/details/77998568
https://www.cnblogs.com/flippedkiki/p/6557114.html
https://my.oschina.net/osandy/blog/414108
https://blog.csdn.net/rubinorth/article/details/52231620https://tw.saowen.com/a/882eb6108d2e498524f749061d111bdab56e62cb7cd07d9427e601f7d0484e2
https://blog.csdn.net/faith_binyang/article/details/54142495
http://www.cnblogs.com/zhangchaoyang/articles/4575809.html