**HW3**

**P76074402 張圜華**

**實作環境 :**

python

**執行方式 :**

1. python hit.py : 會先輸出 graph 1~6 再來修改過的 graph 1~3 最後是direct 跟bi-direct 的值跟圖(要把圖關掉，才會接著跑)
2. python pagerank.py : 會先輸出 graph 1~6 再來修改過的 graph 1~3 最後是direct 跟bi-direct 的值跟圖(要把圖關掉，才會接著跑)
3. python simrank.py : 會先輸出 graph 1~6 再來修改過的 graph 1~3 最後是direct 跟bi-direct 的值跟圖(要把圖關掉，才會接著跑)

**修改的dataset:**

存在 hw3dataset 中，分別為

1. graph\_add1.txt: 增加 [(2,1) (1,3)]
2. graph\_add2.txt: 增加 [(2,1) (1,3)]
3. graph\_add3.txt: 增加 [(1,3) (3,1)]

要提升 1 的Hub和Authority很顯然的就是要增加指向1和從1指出的數目，那更明確的說，就是如果要提高hub值就要讓1指向擁有最高Authority 的node; 如果要提高Authority值就要讓擁有最高hub的node 指向 1

**結果與圖片:**

hit 的結果跟圖片:

存在 hit result and picture 資料夾中，結果存在hit\_result.txt中

pagerank 的結果跟圖片:

存在 pagerank result and picture 資料夾中，結果存在pagerank \_result.txt中

simrank 的結果，存在 simrank result 資料夾中的 simrank\_result.txt中

Hit

**介紹:**

1. **Hub與Authority**

Hub和Authority是HITS算法最基本的兩個定義。

Authority 是指與某個領域或者某個話題相關的高質量網頁，比如搜索引擎領域，Google和百度首頁即該領域的高質量網頁，比如視頻領域，優酷和土豆首頁即該領域的高質量網頁。

Hub 是指的是包含了很多指向高質量“Authority”頁面鏈接的網頁，比如 hao123首頁可以認為是一個典型的高質量“Hub”網頁。

HITS算法的目的即是通過一定的技術手段，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。

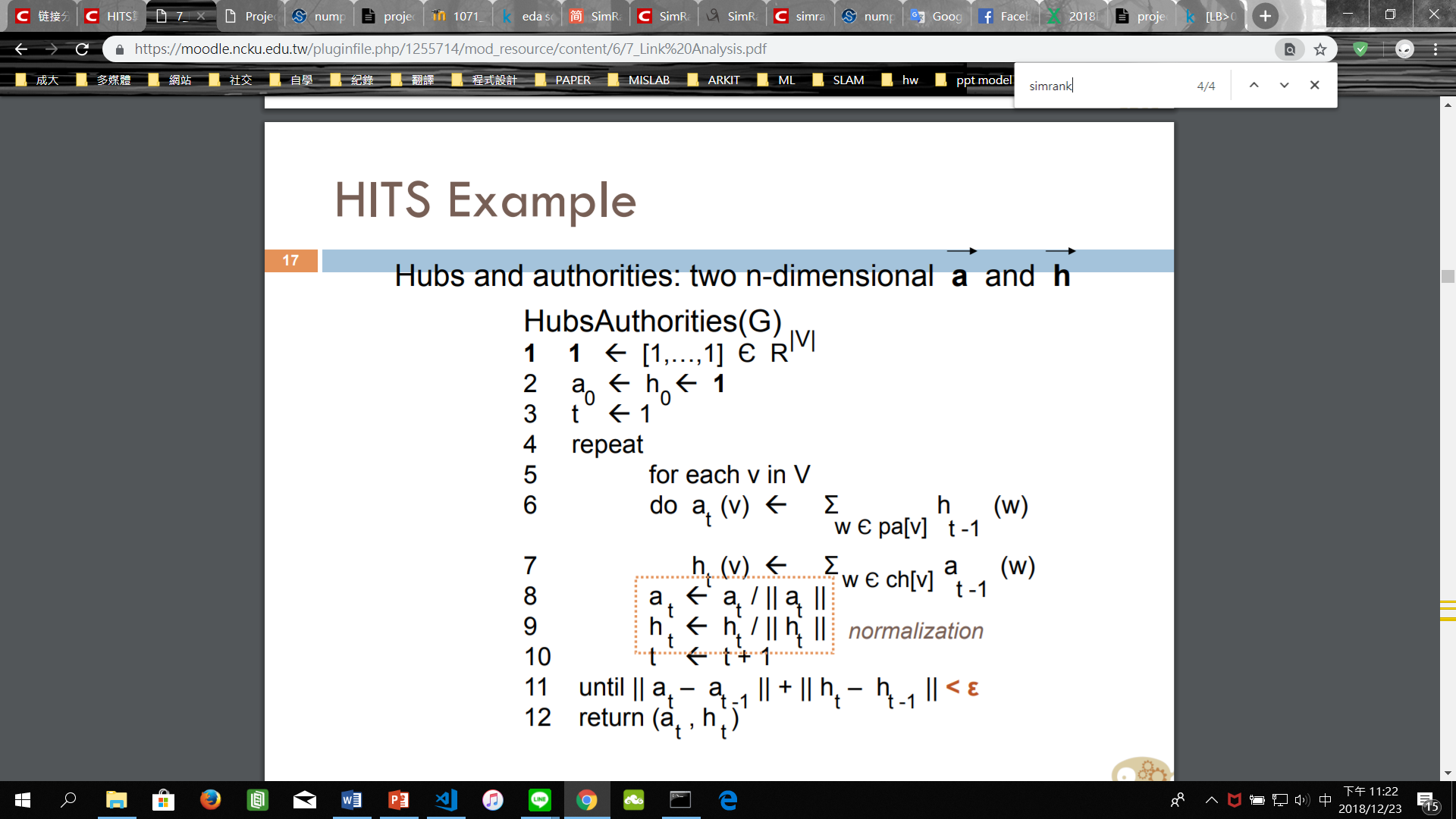
1. **基本思想**

一個好的“Authority”頁面會被很多好的“Hub”頁面指向；

一個好的“Hub”頁面會指向很多好的“Authority”頁面；

**實作細節:**

依照下列演算法實作:



實作程式碼:

def hit(a, h):

a\_ = np.dot(a.T,h)

h\_ = np.dot(a,h)

a\_ = a\_/np.sum(a\_)

h\_ = h\_/np.sum(h\_)

e = 1

while (e > 0.00001):

past\_a = a\_

past\_h = h\_

a\_ = np.dot(a.T,past\_h)

h\_ = np.dot(a,past\_a)

a\_ = a\_/np.sum(a\_) #normalize

h\_ = h\_/np.sum(h\_) #normalize

e = np.linalg.norm(a\_-past\_a)+np.linalg.norm(h\_-past\_h)

return a\_, h\_

變數說明:

1. a是矩陣，表示網頁之間的關係：aij=1表示網頁i指向網頁j，否則為0
2. a\_ 是向量，存所有頁面的authority值，其中第i個分量表示網頁i的authority值
3. h\_ 是向量，存所有頁面的hub值，其中第i個分量表示網頁i的hub值
4. e 是指誤差

實作主要公式:

a\_(t) = np.dot(a.T, h(t-1))

h\_(t) = np.dot(a, a\_(t-1))

直到 ||a\_(t)-a\_(t-1)||+ ||h\_(t)-h\_(t-1)|| < e 才停止

**實作結果分析:**

1. graph 1

分析:

authority 大小順序: 2 3 4 5 6 1

hub 大小順序: 1 2 3 4 5 6

authority : 2 3 4 5 6 相同是因為都有一個node指向他們，而 1 沒有

hub : 1 2 3 4 5 相同是因為他們都有指向一個node，而 6 沒有

1. 修改graph1: graph\_add1.txt : 增加 [(2,1) (1,3)]

比較:

原本 1 的Authority:0 Hub:0.2

後來 1 的Authority: 2.499984e-01 Hub: 4.999958e-01都有增加

1. graph 2

分析:

authority 大小順序: 1 2 3 4 5

hub 大小順序: 1 2 3 4 5

authority : 1 2 3 4 5 相同是因為都有一個node指向他們

hub : 1 2 3 4 5 相同是因為他們都有指向一個node

1. 修改graph2: graph\_add2.txt增加 [(2,1) (1,3)]

比較:

原本 1 的Authority:0.2 Hub:0.2

後來 1 的Authority: 3.568971e-01 Hub: 3.568971e-01都有增加

1. graph 3

分析:

authority 大小順序: 2 3 4 1

hub 大小順序: 2 3 4 1

authority 跟 hub 的值相同，因為是雙向的

1. 修改graph3: graph\_add3.txt增加 [(1,3) (3,1)]

比較:

原本 1 的Authority: 1.909836e-01 Hub: 1.909836e-01

後來 1 的Authority: 2.695941e-01 Hub: 2.695941e-01都有增加

1. graph 4

分析:

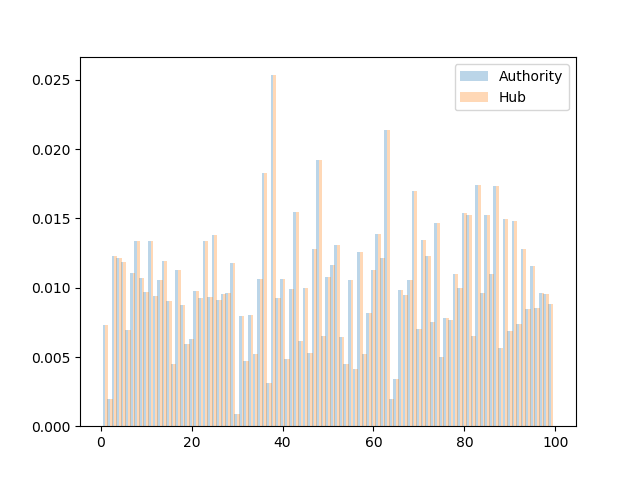
authority 大小順序 : 5 3 2 4 1 7 6

hub 大小順序 : 1 4 5 6 3 7 2

authority : 5 最大是因為都有最多的node指向他

hub : 1 最大是因為他指向最多個node

1. (bidirected)



分析:

從上圖可以很清楚看出authority 跟 hub 的值相同，因為是雙向的

**完整輸出以及圖片都在hit result and picture 資料夾中**

**結果討論:**

HITS算法整體而言是個效果很好的算法，目前不僅應用在搜索引擎領域，而且被“自然語言處理”以及“社交分析”等很多其它計算機領域借鑒使用，並取得了很好的應用效果。然而歸納起來，HITS算法還有以下幾個方面存在不足：

1. 計算效率較低

因為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。

1. 主題漂移問題

如果在擴展網頁集合裡包含部分與查詢主題無關的頁面，而且這些頁面之間有較多的相互鏈接指向，那麼使用HITS算法很可能會給予這些無關網頁很高的排名，導致搜索結果發生主題漂移，這種現像被稱為“緊密鏈接社區現象”（Tightly-Knit CommunityEffect）。

1. 易被作弊者操縱結果

HITS從機制上很容易被作弊者操縱，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鏈接指向作弊網頁，於是可以提升作弊網頁的Authority得分。

1. 結構不穩定

所謂結構不穩定，就是說在原有的“擴充網頁集合”內，如果添加刪除個別網頁或者改變少數鏈接關係，則HITS算法的排名結果就會有非常大的改變。

Page Rank

**介紹:**

PageRank是Google用於用來標識網頁的等級/重要性的一種方法，是Google用來衡量一個網站的好壞的唯一標準。在揉合了諸如Title標識和Keywords標識等所有其它因素之後，Google通過PageRank來調整結果，使那些更具“等級/重要性”的網頁在搜索結果中另網站排名獲得提升，從而提高搜索結果的相關性和質量。其級別從0到10級，10級為滿分。 PR值越高說明該網頁越受歡迎（越重要）。例如：一個PR值為1的網站表明這個網站不太具有流行度，而PR值為7到10則表明這個網站非常受歡迎（或者說極其重要）。一般PR值達到4，就算是一個不錯的網站了。 Google把自己的網站的PR值定到10，這說明Google這個網站是非常受歡迎的，也可以說這個網站非常重要。

page rank 簡單來說:

        就是一個頁面的得票數由所有鏈向它的頁面的重要性來決定，到一個頁面的超鏈接相當於對該頁投一票。一個頁面的PageRank是由所有鏈向它的頁面（鏈入頁面）的重要性經過遞歸算法得到的。一個有較多鏈入的頁面會有較高的等級，相反如果一個頁面沒有任何鏈入頁面，那麼它沒有等級。

對於某個互聯網網頁A來說，該網頁PageRank的計算基於以下兩個基本假設：

數量假設：

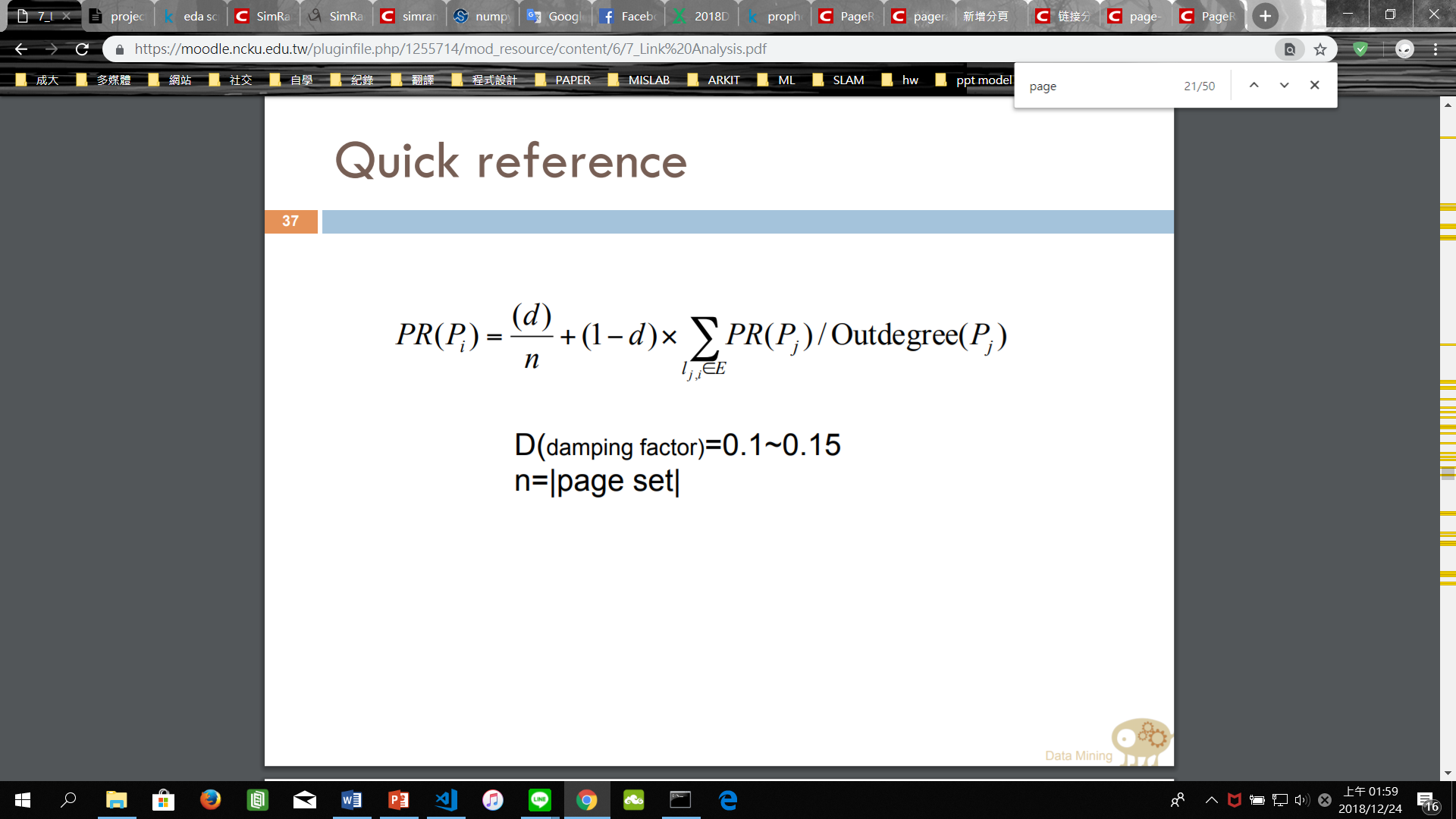
在Web圖模型中，如果一個頁面節點接收到的其他網頁指向的入鏈數量越多，那麼這個頁面越重要。

質量假設：

指向頁面A的入鏈質量不同，質量高的頁面會通過鏈接向其他頁面傳遞更多的權重。所以越是質量高的頁面指向頁面A，則頁面A越重要。

       利用以上兩個假設，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，直到得分穩定為止。 PageRank計算得出的結果是網頁的重要性評價，這和用戶輸入的查詢是沒有任何關係的，即算法是主題無關的。假設有一個搜索引擎，其相似度計算函數不考慮內容相似因素，完全採用PageRank來進行排序，那麼這個搜索引擎的表現是什麼樣子的呢？這個搜索引擎對於任意不同的查詢請求，返回的結果都是相同的，即返回PageRank值最高的頁面。

**實作細節:**



實作程式碼:

def pagerank(a, x):

r = np.dot(a,x)

r = r / np.sum(r)

e = 1

while (e > 0.00001):

past\_r = r

r = np.dot(a,r)

r = r / np.sum(r)

e = np.linalg.norm(r-past\_r)

return r

ee = ee/count

a = 0.85 \* pt + 0.15 \* ee

x = np.ones(count)

變數說明:

1. a 是使用上圖公式，並帶入 d=0.15
2. r 是各個頁面的 PageRank 值
3. e 是誤差

實作主要公式:

r = np.dot(a,r)

直到 ||r-past\_r|| < e 為止，才停止

**實作結果分析:**

1. graph 1

分析:

page rank 大小是: 6 5 4 3 2 1

這個graph 是單向傳遞

1 -> 2

2 -> 3

3 -> 4

4 -> 5

5 -> 6

所以依照page rank的計算方式，6的值是會最大的

1. 修改graph1: graph\_add1.txt

比較:

page rank 大小是: 6 5 4 3 1 2

原本的 1 的 page rank 為 0001 : 3.480503e-02

後來的 1 的 page rank 為 0001 : 7.910035e-02

1. graph 2

分析:

page rank 大小是: 1 2 3 4 5

都相同的原因是這個 graph 也是一個漂亮沒分岔的迴圈，因此 page rank 相同

1. 修改graph2: graph\_add2.txt

比較:

page rank 大小是: 1 3 4 5 2

原本的 1 的 page rank 為 0001 : 2.000000e-01

後來的 1 的 page rank 為 0001 : 2.596471e-01

1 會變大的原因是因為多了2指向1，所以page rank值增加

1. graph 3

分析:

page rank 大小是: 2 3 1 4

從下圖可以很明顯知道 2 3 會相同，1 4會相同

1 -> 2 3 -> 4

1 <- 2 3 -> 4

2 -> 3

2 <- 3

1. 修改graph3: graph\_add3.txt

比較:

page rank 大小是: 3 1 2 4

原本的 1 的 page rank 為 0001 : 1.754375e-01

後來的 1 的 page rank 為 0001 : 2.459272e-01

1 會變大的原因是因為多了3指向1，所以page rank值增加

**完整輸出以及圖片都在pagerank result and picture 資料夾中**

**結果討論:**

PageRank

優點：

        是一個與查詢無關的靜態算法，所有網頁的PageRank值通過離線計算獲得；有效減少在線查詢時的計算量，極大降低了查詢響應時間。

缺點：

1. 人們的查詢具有主題特徵，PageRank忽略了主題相關性，導致結果的相關性和主題性降低
2. 舊的頁面等級會比新頁面高。因為即使是非常好的新頁面也不會有很多上游鏈接，除非它是某個站點的子站點。
3. 沒有區分站內導航鏈接。很多網站的首頁都有很多對站內其他頁面的鏈接，稱為站內導航鏈接。這些鏈接與不同網站之間的鏈接相比，肯定是後者更能體現PageRank值的傳遞關係。
4. 第沒有過濾廣告鏈接和功能鏈接（例如常見的“分享到微博”）。這些鏈接通常沒有什麼實際價值，前者鏈接到廣告頁面，後者常常鏈接到某個社交網站首頁。

**HITS與PageRank比較**

     HITS算法和PageRank算法可以說是搜索引擎鏈接分析的兩個最基礎且最重要的算法。從以上對兩個算法的介紹可以看出，兩者無論是在基本概念模型還是計算思路以及技術實現細節都有很大的不同。

    1.HITS算法是與用戶輸入的查詢請求密切相關的，而PageRank與查詢請求無關。所以，HITS算法可以單獨作為相似性計算評價標準，而PageRank必須結合內容相似性計算才可以用來對網頁相關性進行評價；

2.HITS算法因為與用戶查詢密切相關，所以必須在接收到用戶查詢後實時進行計算，計算效率較低；而PageRank則可以在爬蟲抓取完成後離線計算，在線直接使用計算結果，計算效率較高；

    3.HITS算法的計算對像數量較少，只需計算擴展集合內網頁之間的鏈接關係；而PageRank是全局性算法，對所有互聯網頁面節點進行處理；

    4.從兩者的計算效率和處理對象集合大小來比較，PageRank更適合部署在服務器端，而HITS算法更適合部署在客戶端；

    5.HITS算法存在主題泛化問題，所以更適合處理具體化的用戶查詢；而PageRank在處理寬泛的用戶查詢時更有優勢；

6.HITS算法在計算時，對於每個頁面需要計算兩個分值，而PageRank只需計算一個分值即可；在搜索引擎領域，更重視HITS算法計算出的Authority權值，但是在很多應用HITS算法的其它領域，Hub分值也有很重要的作用；

    7.從鏈接反作弊的角度來說，PageRank從機制上優於HITS算法，而HITS算法更易遭受鏈接作弊的影響。

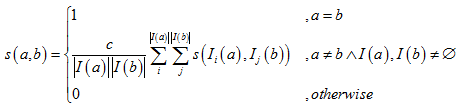
8. HITS算法結構不穩定，當對“擴充網頁集合”內鏈接關係作出很小改變，則對最終排名有很大影響；而PageRank相對HITS而言表現穩定，其根本原因在於PageRank計算時的“遠程跳轉”。

Sim Rank

**介紹:**

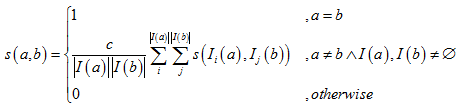
SimRank算法是一種用於衡量結構上下文中個體相似度的方法，其基本思想是：如果兩個對象a和b分別與另外兩個對象c和d關聯，且已知c與d是相似的，則a與b也是相似的；並且任意節點與其自身擁有最大的相似度值為1。 SimRank算法的主要出發點是利用已有個體的相似度來推算其他與之有關聯個體的相似度。

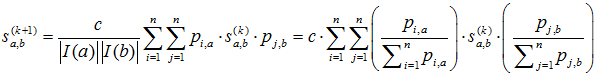
SimRank算法基於一個簡單和直觀的圖論模型，它把對象和對象之間的關係建模為一個有向圖G = (V, E)，其中V是有向圖的節點集合，代表應用領域中的所有對象；E是有向圖的邊的集合，表示對象間的關係。對於圖中的一個節點a，與其所有入邊關聯的鄰節點集合(in-neighbors)記為I(a)，同時，其出邊對應的鄰節點集合(out-neighbors)集合記為O(a )。用s(a, b)表示對象a和對象b之間的相似性，其計算公式為：



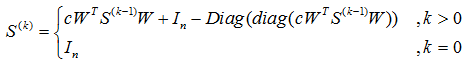
從該計算公式可以看出，個體a, b的相似度取決於所有與a, b相連節點的相似度。式中c是一個常量衰減因子。

**實作細節:**



上述公式可以用矩陣的形式表示出來。假設S表示有向圖G的SimRank分數矩陣，其中s(i,j)表示對象i和j之間的相似性分數； P表示G的連接矩陣，其中P(i,j)表示從頂點i到頂點j的邊數，則

用矩陣的符號表示，即為：



其中，矩陣W表示按列歸一化的P矩陣， I是n×n的單位矩陣。對於一個矩陣X，diag(X)表示獲得由所有X的主對角線上的元素構成的向量；對於一個向量x，Diag(x)操作得到相應的對角矩陣，即x = diag(Diag( x))。由於任意對象和自己的相似度值為1，所以加上項

26c78193c9c6478d0e74909889e0d3496a51df9b (241Ã34)

，其作用是把矩陣 的主對角線元素設為1

實作程式碼:

def simrank(w, s, i\_):

e = 1

while (e > 0.00001):

past\_s = s

s = 0.8 \* np.dot(np.dot(w.T, s),w) + i\_ -np.diag(np.diag(0.8 \* np.dot(np.dot(w.T, s),w)))

e = np.linalg.norm(s-past\_s)

return s

變數說明:

1. s 是矩陣，存在 node 和node間的相似性
2. w 是轉移概率矩陣，它的每一行和為1，如果從節點i可以轉移到節點j，並且這樣的節點i一共有n個，則Qi,j=1/n
3. i 是單位矩陣
4. e 是誤差

實作公式:

s = 0.8 \* np.dot(np.dot(w.T, s),w) + i\_ -np.diag(np.diag(0.8 \* np.dot(np.dot(w.T, s),w)))

直到 ||s-past\_s|| < e 為止，才停止

**實作結果分析:**

1. graph 1

結果會是identity的原因是因為graph1 是單向的，所以每一層找parent都不會有交集

1. graph 2

結果會是identity的原因是因為每一層找parent都不會有交集

**完整輸出都在simrank result and picture 資料夾中**

**結果討論:**

SimRank是一種經典的基於結構的相似度算法,具有快速收斂的優點,但其預計算時間和存儲開銷非常巨大,不適用當前大規模和快速增長環境下的Web頁面網絡。

加速SimRank計算的優化技術：

1. 基本節點選擇可以消除具有先驗零分數的一部分節點對的計算。
2. 部分和記憶可以通過高速緩存部分相似性求和以便以後重用來有效地減少不同節點對之間的相似性的重複計算。
3. 對相似性的閾值設置使得能夠進一步減少要計算的節點對的數量。

**計算效率分析:**

hit

graph1: count: 6 executing hit function time: 9.973049e-04 hit function iteration counts: 1

graph2: count: 5 executing hit function time: 9.984970e-04 hit function iteration counts: 1

graph3: count: 4 executing hit function time: 9.980202e-04 hit function iteration counts: 11

graph4: count: 7 executing hit function time: 9.975433e-04 hit function iteration counts: 19

graph5: count: 469 executing hit function time: 2.912357e-01 hit function iteration counts: 17

graph6: count: 1228 executing hit function time: 7.715178e-02 hit function iteration counts: 95

direct: count: 99 executing hit function time: 2.094507e-02 hit function iteration counts: 11

bi\_direct: count: 99 executing hit function time: 1.293278e-02 hit function iteration counts: 8

add graph1: count: 6 executing hit function time: 1.647472e-02 hit function iteration counts: 21

add graph2: count: 5 executing hit function time: 2.991438e-03 hit function iteration counts: 28

add graph3: count: 4 executing hit function time: 1.994610e-03 hit function iteration counts: 28

pagerank

graph1: count: 6 executing hit function time: 9.980202e-04 hit function iteration counts: 25

graph2: count: 5 executing hit function time: 9.963512e-04 hit function iteration counts: 1

graph3: count: 4 executing hit function time: 9.975433e-04 hit function iteration counts: 12

graph4: count: 7 executing hit function time: 9.977818e-04 hit function iteration counts: 13

graph5: count: 469 executing hit function time: 4.986525e-03 hit function iteration counts: 44

graph6: count: 1228 executing hit function time: 2.593040e-02 hit function iteration counts: 34

direct: count: 99 executing hit function time: 1.097107e-02 hit function iteration counts: 8

bi\_direct: count: 99 executing hit function time: 4.987001e-03 hit function iteration counts: 8

add graph1: count: 6 executing hit function time: 1.561546e-02 hit function iteration counts: 14

add graph2: count: 5 executing hit function time: 0.000000e+00 hit function iteration counts: 11

add graph3: count: 4 executing hit function time: 9.977818e-04 hit function iteration counts: 22

simrank

graph1: count: 6 executing simrank function time: 2.991676e-03 simrank function iteration counts: 2

graph2: count: 5 executing simrank function time: 2.991676e-03 simrank function iteration counts: 2

graph3: count: 4 executing simrank function time: 3.988504e-03 simrank function iteration counts: 12

graph4: count: 7 executing simrank function time: 4.986525e-03 simrank function iteration counts: 26

graph5: count: 469 executing simrank function time: 9.863617e-01 simrank function iteration counts: 34

從上面的資訊看來，執行時間與以下兩點有關:

1. 矩陣大小
2. 幾次迴圈後會收斂

而從上方 hit 、 pagerank 、 simrank 互相作比較，執行時間由快到慢分別為pagerank > hit > simrank

我認為這是因為在迴圈中

pagerank:

r = np.dot(a,r)

hit:

a\_ = np.dot(a.T,past\_h)

h\_ = np.dot(a,past\_a)

sim rank:

s = 0.8 \* np.dot(np.dot(w.T, s),w) + i\_ -np.diag(np.diag(0.8 \* np.dot(np.dot(w.T, s),w)))

pagerank 比 hit快的原因:

少了一行內積處理

hit 比 sim rank快的原因:

sim rank 是矩陣的處理，而Hit 是向量的處理，所以要處理的資料量較小，時間也就比較快速

**Questions & Discussion (optional, but recommended)**

1. **More limitations about link analysis algorithms**

HITS算法結構不穩定，當對“擴充網頁集合”內鏈接關係作出很小改變，則對最終排名有很大影響；而PageRank相對HITS而言表現穩定，其根本原因在於PageRank計算時的“遠程跳轉”。

1. **Can link analysis algorithms really find the “important” pages from Web?**

PageRank沒有區分站內導航鏈接。很多網站的首頁都有很多對站內其他頁面的鏈接，稱為站內導航鏈接。這些鏈接與不同網站之間的鏈接相比，肯定是後者更能體現PageRank值的傳遞關係。因此，can not **r**eally find the “important” pages from Web

1. **What are practical issues when implement these algorithms in a real Web?**

**Performance discussion (time cost)**

SimRank是一種經典的基於結構的相似度算法,具有快速收斂的優點,但其預計算時間和存儲開銷非常巨大,不適用當前大規模和快速增長環境下的Web頁面網絡。

1. **What is the effect of “C” parameter in SimRank?**

c 是影響力的衰退常數，也就是說，他跟影響力有關，而因此有較接近的共同parent node 者，會有較佳的相似性，反之，較遠的會因c影響而相似性低

1. **Design a new link-based similarity measurement**
2. 基本節點選擇可以消除具有先驗零分數的一部分節點對的計算。
3. 部分和記憶可以通過高速緩存部分相似性求和以便以後重用來有效地減少不同節點對之間的相似性的重複計算。
4. 對相似性的閾值設置使得能夠進一步減少要計算的節點對的數量。