P76071129 高至柔 project3 report

[一、 HITS Algorithm 3](#_Toc533450759)

[1 實作結果 3](#_Toc533450760)

[1.1 graph\_1 3](#_Toc533450761)

[1.2 graph\_2 3](#_Toc533450762)

[1.3 graph\_3 4](#_Toc533450763)

[1.4 graph\_4 4](#_Toc533450764)

[1.5 graph\_5 5](#_Toc533450765)

[1.6 graph\_6 5](#_Toc533450766)

[1.7 data from project 1 (bi-directed) 6](#_Toc533450767)

[1.8 data from project 1 (directed) 6](#_Toc533450768)

[2 實作細節與Performance Analysis 7](#_Toc533450769)

[3 討論 7](#_Toc533450770)

[二、 PageRank Algorithm 8](#_Toc533450771)

[1 實作結果 8](#_Toc533450772)

[1.1 graph1 8](#_Toc533450773)

[1.2 graph2 8](#_Toc533450774)

[1.3 graph3 8](#_Toc533450775)

[1.4 graph4 8](#_Toc533450776)

[1.5 graph5 9](#_Toc533450777)

[1.6 graph6 9](#_Toc533450778)

[1.7 data from project 1 (bi-directed) 擷取部分 10](#_Toc533450779)

[1.8 data from project 1 (directed) 擷取部分 10](#_Toc533450780)

[2 實作細節與Performance Analysis 11](#_Toc533450781)

[3 討論 11](#_Toc533450782)

[三、 SimRank Algorithm 12](#_Toc533450783)

[1 實作結果 12](#_Toc533450784)

[1.1 graph1 12](#_Toc533450785)

[1.2 graph2 12](#_Toc533450786)

[1.3 graph3 12](#_Toc533450787)

[1.4 graph4 12](#_Toc533450788)

[1.5 graph5 12](#_Toc533450789)

[2 實作細節與Performance Analysis 13](#_Toc533450790)

# HITS Algorithm

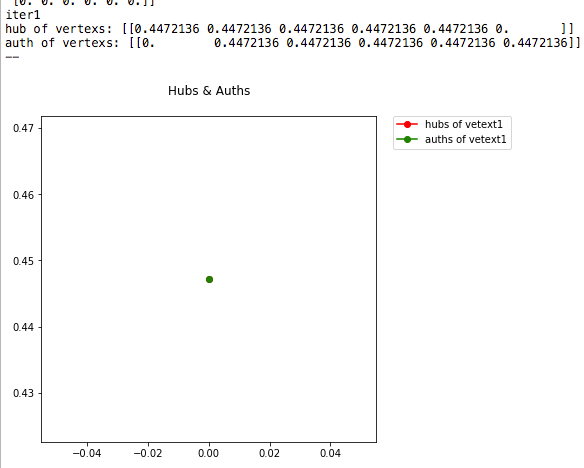
## 實作結果

graph1～graph4呈現出最後每一個點的hub 及 authority值

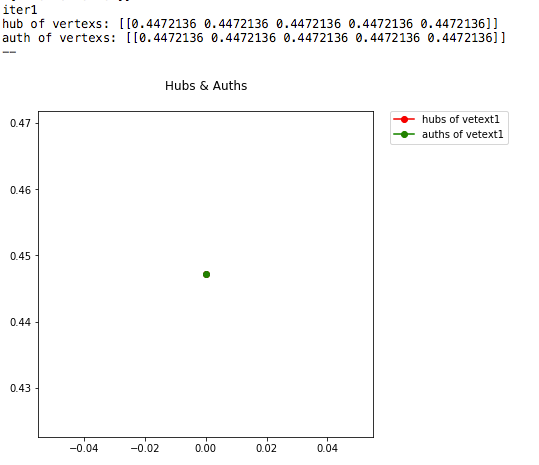
graph 5 6及project1的data 因為點太多了就不貼上來了（程式執行可顯示）

並且將每張圖的Node1 的hub(紅線) 及 authority(綠線) 值呈現於圖表中。

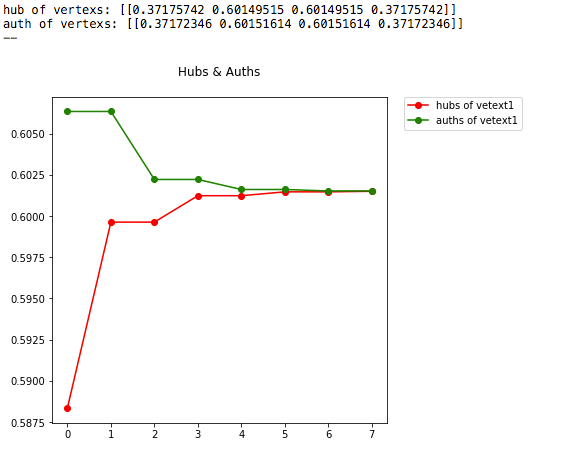
### graph\_1



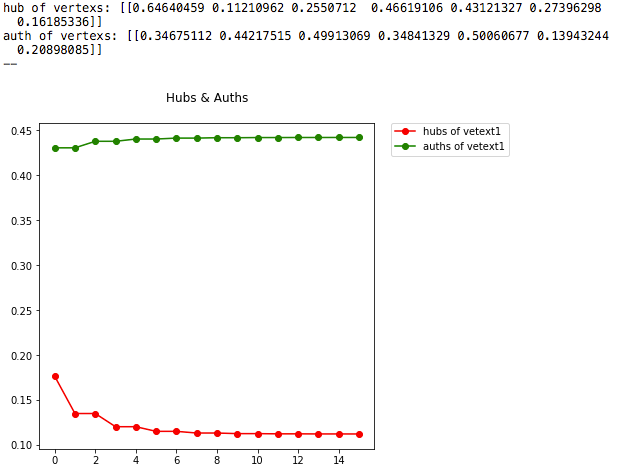
### graph\_2



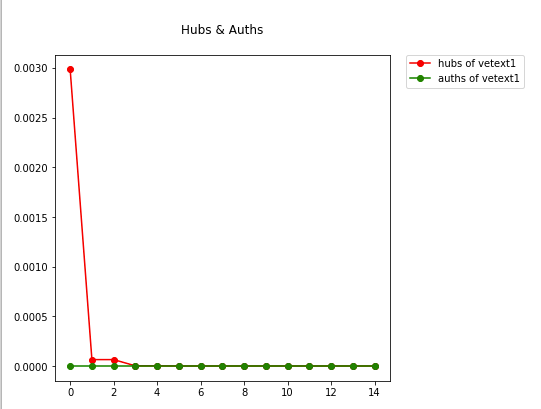
### graph\_3



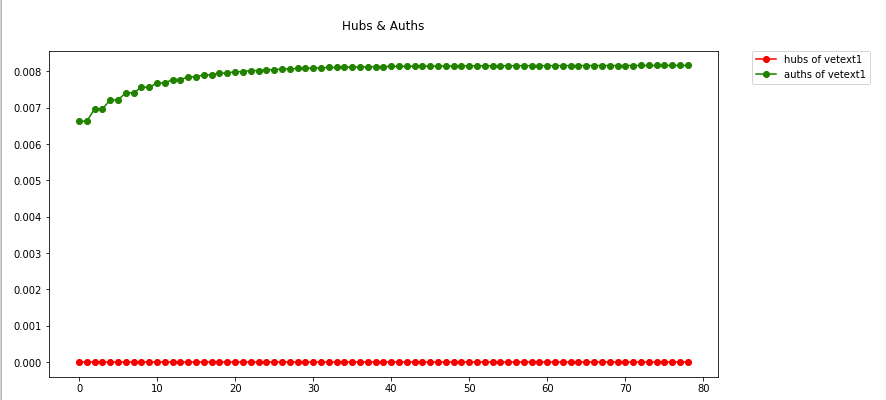
### graph\_4



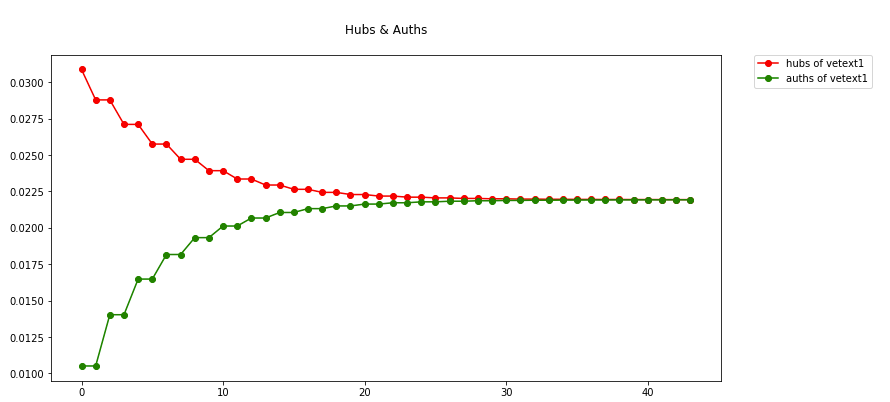
### graph\_5



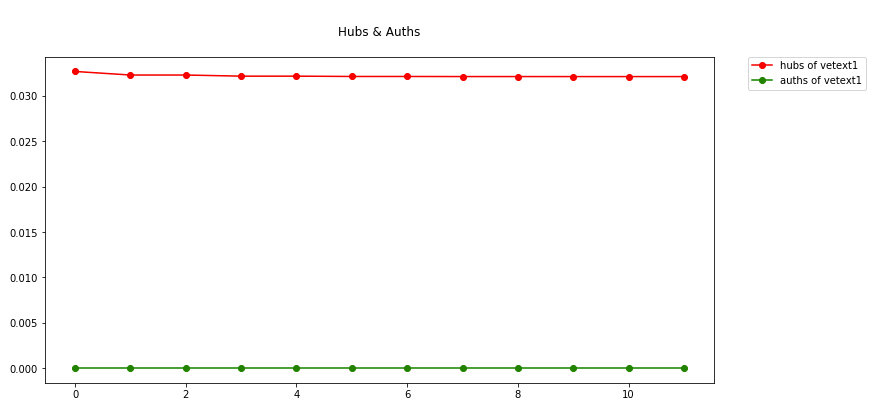
### graph\_6



### data from project 1 (bi-directed)



### data from project 1 (directed)



## 實作細節與Performance Analysis

n: node number e: link number

在網頁分析中e 應該會遠小於n

|  |  |
| --- | --- |
| 實作細節 | Time complexity |
| File\_max 用來讀取node總數建立合適大小的矩陣 | O(n) |
| FiletoAD 讀取檔案轉換成Adjacency Matrix | O(n) |
| 利用csc\_matrix 將矩陣轉換成適合  sparse matrix做運算的形式 |  |
| Hit coherence function：  定義epsilon容許的誤差，  迭代計算hub及authority值，  並作normalization。 | O(1)  O(n)  O(n) |
| 最後作圖呈現vertex1在迭代過程中hub及authority值的變化 |  |
| Total | O(n) |

## 討論

HITS algo其實很快就可以收斂了，用這個方法可以很快的求出具有權威性及能指出權威性的網站，當然這兩者之間也會互相影響，必定能收斂。

如果是無相圖的話可以明顯的看出來最後該網站的hub及authority應該要一樣。

因爲graph6的資料必較像網路連結的情況，可以很明顯地看出兩點震盪的情形。

如果要增加node1 的hub值可以讓node1 增加一條連結到authority較高的node; 如果要增加node1 的authority值可以讓authority較高的node 增加一條連結到node1。

# PageRank Algorithm

## 實作結果

矩陣中表示每一個算出來的PageRank值

(d=0.15，將最後的結果normalization過)

### graph1



### graph2



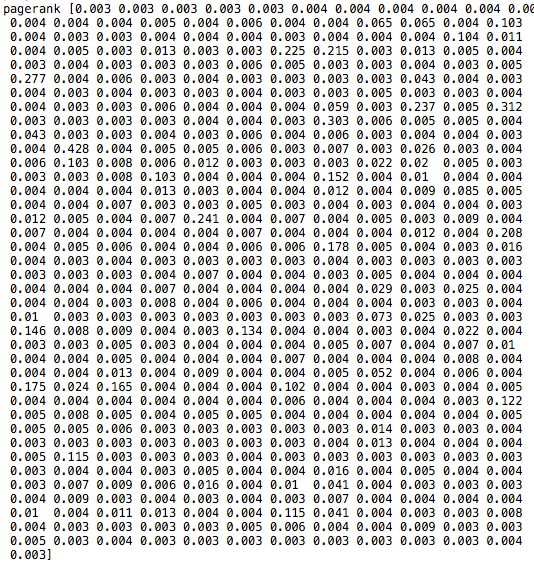
### graph3



### graph4



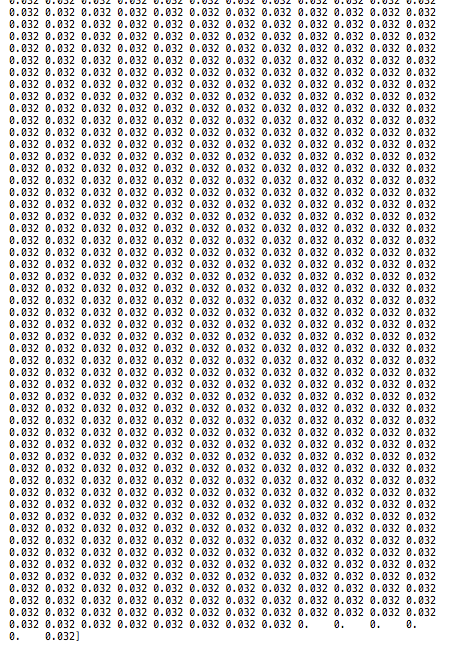
### graph5



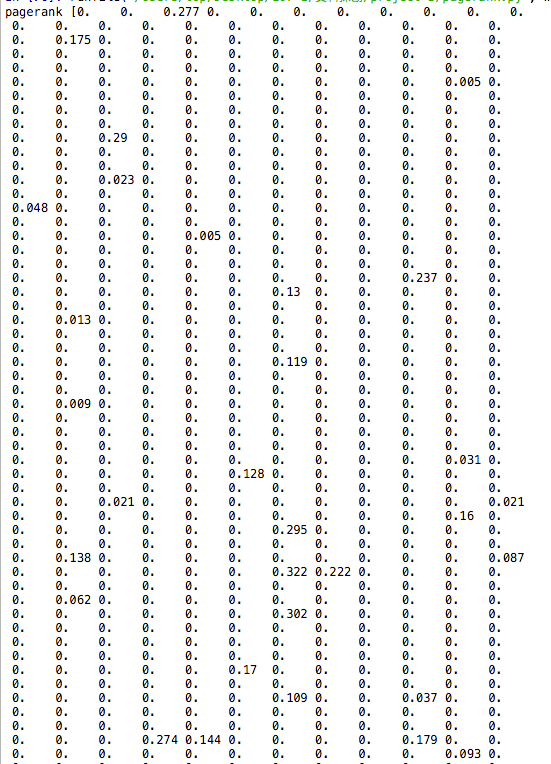
### graph6



### data from project 1 (bi-directed) 擷取部分



### data from project 1 (directed) 擷取部分



## 實作細節與Performance Analysis

n: node number e: link number

|  |  |
| --- | --- |
| 實作細節 | Time complexity |
| File\_max 用來讀取node總數建立合適大小的矩陣 | O(n+e) |
| FiletoAD 讀取檔案轉換成Adjacency Matrix | O(n+e) |
| 利用csc\_matrix 將矩陣轉換成適合  sparse matrix做運算的形式 |  |
| PageRank function：  設定s = (1-damping factor)，容許誤差，  迭代計算PageRank，  並作normalization。 | O(1)  O(n\*e)  O(n) |
| Total | O(n\*e) |

## 討論

PageRank 起初實作的時候沒有設定S值（即1-d），發現這樣無法讓graph\_6收斂，因為graph\_6並不是強連通圖，有些沒有連結到得點會造成無法收斂的結果。

如果是無相圖，則每個點的PageRank會一樣，除了孤立點外。

將其他點都連向node1，或減少連到node1的點連出去的link 可以增加node1的PageRank。

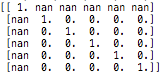
# SimRank Algorithm

## 實作結果

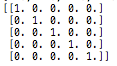
採用C =10 ，矩陣會顯示node i & j 之間的simrank值

若有node的in-degree是0，則兩個node之間無法計算該值顯示為nan

### graph1



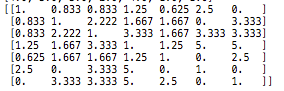
### graph2



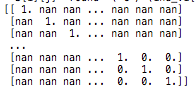
### graph3



### graph4



### graph5



## 實作細節與Performance Analysis

|  |  |
| --- | --- |
| 實作細節 | Time complexity |
| File\_max 用來讀取node總數建立合適大小的矩陣 | O(n) |
| FiletoAD 讀取檔案轉換成Adjacency Matrix | O(n) |
| 利用csc\_matrix 將矩陣轉換成適合  sparse matrix做運算的形式 |  |
| simrank function：  (由於對稱性，先求出下三角的減少程式執行時間)  計算每個node的in-degree，  summation有相同in node，  最後求出兩點間的simrank。 | O(n^2)  O(n^2)  O(n^2) |
| Total | O(n^2) |

## 討論

連結結構相像的的兩點 simrank會較大。

如果有點的in-degree是0 ，無法比較兩點之間的simrank。