**Project 3 Report**

工資所 R36074121 辛政達

1. 程式架構

本次作業主要在探討基於圖論模型進行網頁重要度排名的HITS與PageRank演算法，以及計算節點相似度的SimRank演算法。

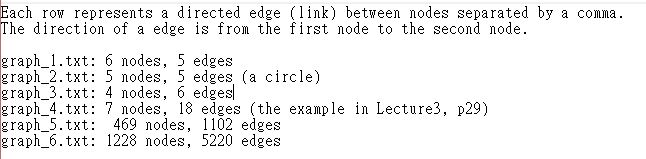
此次使用JAVA進行程式的撰寫，並且針對不同的演算法獨立撰寫不同的模組。為了能夠完成輸入資料網路圖型化的部分，我使用了JGraphT套件中所提供的圖型操作之基本方法的部分。

此次實作的三種演算法都是屬於類似的收斂演算流程，在輸入網路圖型與相關參數後，首先進行迭代陣列的初始化，依據各個演算法設定不同的初始值。在迭代運算的部分，主要是根據各演算法的公式，找出每點或每組點關係的新答案，此部分原則上會需要查找該點的輸入節點或輸出節點，因此與圖論模型的完善程度有關，此也為選擇JGraphT套件的原因之一。

完成該次迭代後，便會進入迭代檢查中，迭代檢查分為兩部分，分別是相似度檢測與最大迭代數檢查。相似度檢測採用此迭代與前代陣列的差異做幾何平均計算其誤差值，若小於一定程度後就結束迭代過程。而若無法再最大迭代數前收斂，則會在抵達最大迭代數時停止運算，以防止陷入無窮迴圈的情況。完成迭代後即輸出結果，結束該演算法。

1. 資料集說明

本次作業總共使用了八筆資料圖型，前面六筆為此次作業給予的資料集(gragh\_1.txt~graph\_6.txt)，此六筆資料集的資訊如下所示:



另外兩筆資料圖型則為曾經在project1中使用過的以IBM Quest data Generator所產生的資料集，以及其關聯規則(FP-Growth)作為輸入圖型，且參數設置如下:

|  |  |
| --- | --- |
| IBM-Quest-Data-Generator 資料產生參數 | |
| 交易紀錄平均長度 | 10 |
| 關聯規則平均長度 | 4 |
| 交易紀錄數量 | 10萬 |
| FPGrowth 演算參數 | |
| Support | 0.005 |
| Confidence | 0.6 |
| 只顯示最關鍵關聯法則 | 是 |

由於一筆交易資料內可能有很多個項目，因此在本次的圖形定義中，將每筆交易資料的開始項目設為出發點，之後的每一個項目為一筆從出發點連至該點的邊。完成定義且修改檔案格式並且產生關連規則後，得到原始交易紀錄與關連規則的兩筆資料圖型，下面則會針對每一張圖的結果進行結果的分析。

1. HITS、PageRank、SimRank演算法結果分析

本次作業三種演算法相關參數的設定如下:

|  |  |
| --- | --- |
| 通用部分 | |
| 跳脫迭代過程的最大差異值 | 0.00005 |
| 計算與輸出精度 | 小數點後第9位四捨五入 |
| PageRank | |
| Damping Factor | 0.85 |
| SimRank | |
| Damping Factor | 0.8 |

進行運算後，此次作業提供的六筆資料圖型可以得到以下結果:

1. graph\_1.txt(輸出：ORGINAL\_1)

在圖一中，由於是一條單向的直向連接，在PageRank的方面，很明顯的可以看出V1的PageRank值僅有來自Damping Factor所給與基礎分數，而越後面的結點則繼承了前面的分數而越來越高。

在HITS演算法中，V1由於沒有節點接入，所以其Authority為0，而同理沒有節點流出的V6其HUB值也是0。其他部分則因為相同的流入流出結構而獲得一樣多的分數。而這分數的呈現，在進行公式推導後，我認為應該取決於將迭代陣列轉換成向量後得到的距離有關，在節點數為n的單向的直向連接中，除了首點Authority與末點HUB外，其他值應該為，其n值越大節點，分數就會越低。

|  |  |  |  |
| --- | --- | --- | --- |
|  | PageRank | Authority | HUB |
| Vertex 1: | 0.025 | 0 | 0.447214 |
| Vertex 2: | 0.04625 | 0.447214 | 0.447214 |
| Vertex 3: | 0.064313 | 0.447214 | 0.447214 |
| Vertex 4: | 0.079666 | 0.447214 | 0.447214 |
| Vertex 5: | 0.092716 | 0.447214 | 0.447214 |
| Vertex 6: | 0.103808 | 0.447214 | 0 |

在這一張圖中，由於節點間並沒有有相同的母節點，所以在SimRank演算法中，除了自己與自己節點關係為1外，其餘節點間的關係皆為0。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SimRank | V1 | V2 | V3 | V4 | V5 | V6 |
| V1 | 1 | 0 | 0 | 0 | 0 | 0 |
| V2 | 0 | 1 | 0 | 0 | 0 | 0 |
| V3 | 0 | 0 | 1 | 0 | 0 | 0 |
| V4 | 0 | 0 | 0 | 1 | 0 | 0 |
| V5 | 0 | 0 | 0 | 0 | 1 | 0 |
| V6 | 0 | 0 | 0 | 0 | 0 | 1 |

1. graph\_2.txt(輸出：ORGINAL\_2)

在圖二中，圖形從原本的單向連接變成了一個環狀單向連接，使得PageRank值藉由彼此的連接而變得一模一樣。根據我的計算，在節點數為n的環狀單向連接下，所有節點的PR值皆為，而所有節點的Authority與HUB皆為。

|  |  |  |  |
| --- | --- | --- | --- |
|  | PageRank | Authority | HUB |
| Vertex 1: | 0.2 | 0.447214 | 0.447214 |
| Vertex 2: | 0.2 | 0.447214 | 0.447214 |
| Vertex 3: | 0.2 | 0.447214 | 0.447214 |
| Vertex 4: | 0.2 | 0.447214 | 0.447214 |
| Vertex 5: | 0.2 | 0.447214 | 0.447214 |

但是就算所有節點都接在一起了，任意兩個節點間還是沒有相同的母節點存在，因此呈現的結果也跟圖一一樣。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SimRank | V1 | V2 | V3 | V4 | V5 |
| V1 | 1 | 0 | 0 | 0 | 0 |
| V2 | 0 | 1 | 0 | 0 | 0 |
| V3 | 0 | 0 | 1 | 0 | 0 |
| V4 | 0 | 0 | 0 | 1 | 0 |
| V5 | 0 | 0 | 0 | 0 | 1 |

1. graph\_3.txt(輸出：ORGINAL\_3)

圖三的結構與圖一十分的相似，是一個直向連接的圖型，只不過此圖的連接是雙向的。從此圖可以發現，PageRank和HITS的結果均是位處於中間的節點分數較高，位於首尾的節點分數則較低。

|  |  |  |  |
| --- | --- | --- | --- |
|  | PageRank | Authority | HUB |
| Vertex 1: | 0.1754325 | 0.371744 | 0.371744 |
| Vertex 2: | 0.3245675 | 0.601503 | 0.601503 |
| Vertex 3: | 0.3245675 | 0.601503 | 0.601503 |
| Vertex 4: | 0.1754325 | 0.371744 | 0.371744 |

在SimRank的結果中，不難發現有相同母節點的V1與V3、V2與V4具有較高的相似度，雖然系出同源，但沒有相同母節點的V1與V4，其相似度為0。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SimRank | V1 | V2 | V3 | V4 |
| V1 | 1 | 0 | 0.66665548 | 0 |
| V2 | 0 | 1 | 0 | 0. 66665548 |
| V3 | 0.66665548 | 0 | 1 | 0 |
| V4 | 0 | 0.66665548 | 0 | 1 |

1. graph\_4.txt(輸出：ORGINAL\_4)

此圖型為講義上的範例，但本次使用的Damping Factor與講義範例中不同，因此數字並不一致。我們可以從此圖發現，V1的PageRank值最高，代表在網路環境中V1具有較高的重要性。而在SimRank的測試中，可以看到最高相似度的組合為V4與V6以及V4與V7。

|  |  |  |  |
| --- | --- | --- | --- |
|  | PageRank | Authority | HUB |
| Vertex 1: | 0.280278 | 0.346708 | 0.646419 |
| Vertex 2: | 0.158768 | 0.442186 | 0.112094 |
| Vertex 3: | 0.138883 | 0.499136 | 0.25506 |
| Vertex 4: | 0.10822 | 0.348409 | 0.466203 |
| Vertex 5: | 0.184201 | 0.500624 | 0.431194 |
| Vertex 6: | 0.060568 | 0.139417 | 0.273954 |
| Vertex 7: | 0.069081 | 0.208992 | 0.161859 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SimRank | V1 | V2 | V3 | V4 | V5 | V6 | V7 |
| V1 | 1 | 0.3602 | 0.3489 | 0.3537 | 0.3376 | 0.4151 | 0.2924 |
| V2 | 0.3602 | 1 | 0.4067 | 0.3697 | 0.4122 | 0.2854 | 0.4540 |
| V3 | 0.3489 | 0.4067 | 1 | 0.4496 | 0.3900 | 0.4481 | 0.4510 |
| V4 | 0.3537 | 0.3697 | 0.4496 | 1 | 0.3426 | 0.5351 | 0.5351 |
| V5 | 0.3376 | 0.4122 | 0.3900 | 0.3426 | 1 | 0.2731 | 0.4122 |
| V6 | 0.4151 | 0.2854 | 0.4481 | 0.5351 | 0.2731 | 1 | 0.2701 |
| V7 | 0.2924 | 0.4540 | 0.4510 | 0.5351 | 0.4122 | 0.2701 | 1 |

1. graph\_5.txt(輸出：ORGINAL\_5)

此張圖屬於一個較大的資料集，由於輸出資料較多，因此詳細輸出附於輸出資料內，不在此份報告中呈現。

雖然資料集較大，但整體而言，三種演算法均能在1秒之內完成運算並輸出。

1. graph\_6.txt(輸出：ORGINAL\_6)

此張圖的資料更加龐大，HITS與PageRank演算法總共花了約2秒的時間才能完成運算，由於資料量過於龐大，因此在此張圖的部分並沒有測試SimRank演算法

1. 原始交易紀錄(輸出：CUSTOM\_GRAPH)

本次作業所使用的資料，共有871個節點與107433個邊，資料過於龐大，因此不對此執行SimRank演算法。此外，可以發現多數的PageRank都低於1/1000，最高的數值也只有約2/1000，其原因為節點數量較多且連線過於密集，導致分數普遍性的被拉低。

1. FP-Growth關聯規則(輸出: CUSTOM\_GRAPH\_FPGROWTH)

透過關聯規則產生出來的圖型，能夠減低許多節點與邊的數量，因此分數與第七張圖相比相對的提升很多。而仔細觀察可以發現，PageRank的分數與其關聯規則的SUP值有關聯。

1. 問題討論

此部分將針對作業後所附的”Question & Discussion”的部分問題來進行研究與說明。

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

PageRank與HITS演算法確實能夠找出在網路結構相對重要的頁面。但其功能也就是探測網路節點在整個圖形中的重要性而已，對於該節點的性質與其他因素，演算法並沒有考慮到，因此該演算法只能在完全將所有節點視為相同性質的狀況下使用。

除此之外，若在數量非常多的網路中，每個節點的重要度都被拉低，分數差異非常的小，相較來說透過這兩種演算法找出重要頁面的難度就會更高。

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

此三種演算法其實不太適用於真實網路，這是因為在真實網路中，頁面的數量與連結複雜度均有大幅度的提升，而PageRank、HITS、SimRank三種演算法都必須對整張圖進行多次的迭代運算才能完成收斂並輸出。

以圖六為例子，我們可以發現光是1228個節點5220個邊的資料量就已經無法實作SimRank(實行時間超過5分鐘)，這也表示這類演算法的時間複雜度是非常高的。若要在真實網路環境下使用，我認為這些演算法都必須進行一些算法上的改善才可以。

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

在SimRank中，C(factor of limited confidence) 代表的是相似程度的衰減指數，當兩點共同的母節點距離越遠時，其相似度會隨著衰減指數的次方而增加。