Link Analysis Homework 2

Student ID : P76034711
Name : 呂鴻
Environments : Ubuntu 14.04
Language : Java
Execute : Command Line

◆ Implementation detail for each algorithm :

首先,我會先列出全部的 graph 檔給使用者作選擇,選擇完要處理的 graph 之後在選擇 algorithm,然後就會列出該演算法會算出的值,以下會針對不同演算法做細部的解說.

```
neal@neal-K42Jv:~/workspace/LinkAnalysis/bin$ java LinkAnalysis
Please choose File :
1. graph_1.txt
2. graph_2.txt
3. graph_3.txt
4. graph_4.txt
5. graph_5.txt
6. graph_6.txt
7. graph_7.txt
8. graph_8.txt
9. graph_9.txt
10. graph_10.txt
```

1. HITS :

Hyperlink Induced Topic Search(HITS) 最基本的兩個定義 Authority and Hubs , 使用者會給定一個 query point, 根據 query 我們可以透過搜尋系統找到 root set. 這裡我是藉由 query 點向外擴展一層當作 root set, 再來是根據 root set 來擴充成 base set. 這裡我實作方法是將 root set 有相關的 node 也都加進來, 所以簡單來說就是在向外況展一層. 有個特例: 如果整個 data set 小於十個 node, 因為資料太少了所以我將整個 data set 當作是 base set.

然後就去計算 base set 內中所有 node 的 authority value and hubs value. 算出來後根據這兩個定義分別去作排名的動作, 取前五個.

```
Please choose root :
1. node 1       2. node 2       3. node 3       4. node 4       5. node 5       6. node 6
Root :
3
        Authority       Hubs
Node 1 :
        0.17757241      0.71178522
Node 2 :
        0.31268159      0.57442658
Node 3 :
        0.56502274      0.00000004
Node 4 :
        0.00000010      0.40422260
Node 5 :
        0.74259515      0.00000004
Node 6 :
        0.00000010      0.00000000

Authority Ranking for Top 5:
        Node 5 : 0.74259515
        Node 3 : 0.56502274
        Node 2 : 0.31268159
        Node 1 : 0.17757241
Hubs Ranking for Top 5:
        Node 1 : 0.71178522
        Node 2 : 0.57442658
        Node 4 : 0.40422260
        Node 3 : 0.00000004
```

2. PageRank :

舉個簡單的例子來說，PageRank 就是手上握有的投票數，他們的對外連結就是把手上得票平均分出去，也就是說，別人連到你的網站，表示他把票投給你．結論大致上可以說是連結到你的網頁愈多，你的 PageRank 值就愈高，但細節來看，如果連到你的都是弱連結，那麼你的 pageRank 值也比較難累積．

實作方面就是透過演算法算出每個 node 的 PageRank value, 然後取前 10 名列出來．PageRank 有點像是 HITS 的 authority 值．

```
The PageRank ranking is :
Node 1 :
        0.28028763
Node 5 :
        0.18419804
Node 2 :
        0.15876440
Node 3 :
        0.13888175
Node 4 :
        0.10821955
Node 7 :
        0.06907747
Node 6 :
        0.06057065
```

3. SimRank :

SimRank 是一種基於圖的拓撲結構信息來衡量任意兩個對象間相似程度的模型，SimRank 相似度的核心思想為：如果兩個對象和被其相似的對象所引用（即它們有相似的入鄰邊結構），那麼這兩個對象也相似。近年來已在信息檢索領域引起廣泛關注，成功應用於網頁排名、協同過濾、孤立點檢測、網路圖聚類、近似查詢處理等。

實作方面就是讓使用者給定兩個 nodes，然後去計算其結構相似度．

```
Please choose two nodes to calculate similarity :
1. node 1       2. node 2       3. node 3       4. node 4       5. node 5       6. node 6       7. node 7
Pair 1 :
2
Pair 2 :
5

S(2,5) : 0.10846354
```

◆ Characteristics/observation for 2 LP model graphs and 2 graphs you find/generate from IMDB and other sources :

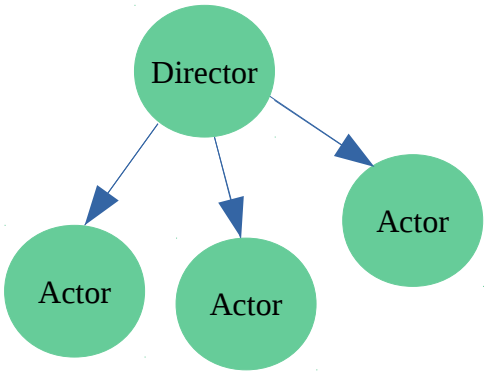
1. LP model graphs :

I random two graphs. For graph 7, I give 50 nodes and random the connections between random nodes. For graph 8, I give 100 nodes and do it as the same way. So everytime I run the program, these two graphs will be different.

2. IMDB graph :

I choose the top 250 movies on IMDB website as the dataset. (link) Each movie contains many informations. I choose the Director and the Actors for each movie. The director connects to each actor. So the node means people who is director or actor. It means famous director will have higher Hubs in HITS algorithm because the director directs lots of movies. And famous actor will have higher Authority value because the actor acts lots of movies. The node will present the people's name.

```
Authority Ranking for Top 5:
        Morgan Freeman : 0.58102816
        Ben Affleck : 0.35834393
        Rosamund Pike : 0.35834393
        Brad Pitt : 0.35834393
Hubs Ranking for Top 5:
        David Fincher : 0.84987426
        Frank Darabont : 0.52698552
        Tim Robbins : 0.00000000
        Michael Clarke Duncan : 0.00000000
```



3. Generate by my own :

I use graph from lecture 3 p.52 for my temporary dataset.

◆ Result analysis and discussion :

1. Limitations about link analysis algorithm :

HITS 算是個不錯的演算法, 不僅用在 search engine, 還被用在 natural language and social network analysis 上, 但我們還是可以歸納 HITS 的一些缺點 :

(1) lower computation effects :

因為 HITS 是與查詢相關的演算法, 在使用者下 query 後才可以作計算, 然而他的演算是用迭代的計算才能得到最後的結果, 導致計算效率較低.

(2) 主題漂移的問題 :

如果在擴展 root set 中包含部份與查詢無關的主題頁面, 而且這些頁面有較多的相互連結方向, 那麼使用 HITS 可能會給予這些頁面很高的排名, 導致搜尋結果主題漂移, 這種現象稱為"緊密連接社區現象"

(3) 易被操縱 :

例如說作弊者可以建立一個網頁, 頁面內容增加很多指向高質量的或者著名網站的網址, 他就會變成一個很好的 Hub 頁面, 之後作弊者在將這個網頁連到作弊的網頁, 於是可以提昇作弊網頁的 Authority.

PageRank 演算法是基於"從許多優質的網頁連結過來的網頁, 必定還是優質的網頁"的回歸關係來判定重要性與否. 也可以用上述的投票例子來闡述 PageRank 的演算流程.

PageRank 的優點在於它對網頁有全局的重要性排序, 並且計算流程可以在 off-line 完成, 這樣有利於迅速回應使用者的請求.

但其缺點在於主題無關性, 沒有區分頁面內的導航連結 廣告連結和功能連結等, 容易對廣告頁面有過高的評價. 另一個弊端是舊的頁面等級會比新的頁面高, 因為新頁面不會有太多連結, 這就是為什麼 PageRank 需要多項算法結合的原因.

Search Engine 藉由 PageRank 值的分數來判斷此網站的重要程度, 進而做為關鍵字搜尋的排名依據參考, PageRank 值越高代表 search engine 對此站的重要度高, 相對的信賴及內容準確度也越高. 但也有人用灌水的方式來灌 PageRank 的值, Google 用了一個方法來辨識虛偽的連結, google 推出了新的屬性 nofollow, 使得網站管理員和網誌作者可以做出一些 Google 不計票的連結, 也就是說這些連結不算作"投票", nofollow 的設置可以抵制評論垃圾.

SimRank 模型定義兩個頁面的相似度是基於如下遞歸的思想：如果指向結點 a 和指向結點 b 的結點相似，那麼 a 和 b 也認為是相似的。這個遞歸定義的初始條件是：每個結點與它自身最相似。 這個方法主要是基於不動點迭代來解 SimRank 值，其優點是計算精度較高，但時間複雜度較大.

- Find a way (e.g., add/delete some links) to increase hub, authority, and pagerank of Node 1 in first 4 graphs respectively :

If page wants to increase hub , just connect the higher authority's node.
If page wants to increase authority , just connect the higher hub's node.
If page wants to increase PageRank, just let higher PageRank's node connect to Node1 .

Increase hub :

graph 1 : 1 → 5
graph 2 : 1 → anynode
graph 3 : 1 → 2
graph 4 : 1 → 5
graph 5 : 1 → 433
graph 6 : 1 → 410
graph 7 : 1 → 不定(graph random created)
graph 8 : 1 → 不定(graph random created)
graph 9 : 1 → Tom Hanks
graph 10 : 1 → 3 or 1 → 5

Increase authority :

graph 1 : 1 → 4
graph 2 : 1 → anynode
graph 3 : 1 → 2 or 1 → 3
graph 4 : 1 → 4
graph 5 : 1 → 396
graph 6 : 1 → 410
graph 7 : 1 → 不定(graph random created)
graph 8 : 1 → 不定(graph random created)
graph 9 : 1 → Frank Darabont
graph 10 : 1 → 2

Increase PageRank :

為了增加 PageRank, 只要把 PageRank 高的頁面指向 PageRank 低的頁面這樣就可以提昇. Pagerank 愈高越好, 然後連出來的連結愈少愈好. 詳細的內容在 Result analysis and discussion 的章節有詳細討論.

◆ What I learned and Comments :

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

我覺得這兩種演算法 HITS and PageRank 都是找重要的 page 一個很重要的參考依據, 但不能完全依據. 所以 Google 提出了很多改善方法, 且不會只單獨運用一個演算法, 會綜合每一種演算法得出最佳的結果.

2. What do the result say for your actor/movie graph ?

I craw IMDB website of top 250 movies web page and get the directors and the actors for each movie. Each node presents director or actor. The director connect to actors for each movie. The result says that the higher hubs means he is a popular director and he direct lots of movie. The higher authority means he is a popular actor and lots of movie he acts. So the higher authority people may be a superstar.

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

C 稱為阻尼係數. SimRank 就是計算兩節點的相似程度, 其意義是兩個體 a b 的相似度取決於 a, b 相連節點的相似程度, s(a,b) 介於[0,1], 當 a=b 時, s(a,b)=1, 現在如果 a!=b, 且只有一個入節點 d, 我們不希望從 d 計算到的 s(a,b)也是 1, 因此 C 也可以作為衰減因子, 介於[0,1], s(a,b)=C. C 值愈小, 表示衰退的速度愈快, 反之則愈慢.

◆ Execution Process :

Ex : choose graph 5 to calculate Node 70's HITS

```
neal@neal-K42Jv:~/workspace/LinkAnalysis/bin$ java LinkAnalysis
Please choose File :
1. graph_1.txt
2. graph_2.txt
3. graph_3.txt
4. graph_4.txt
5. graph_5.txt
6. graph_6.txt
7. graph_7.txt
8. graph_8.txt
9. graph_9.txt
10. graph_10.txt
5
Please choose methods for link analysis :
1. HITS
2. PageRank
3. SimRank
1
```

```
Please choose root :
1. node 1       2. node 2       3. node 3       4. node 4       5. node 5       6. node 6       7. node 7       8. node 8       9. node 9       10. node 10     11. node 11     12. node 12     13. node 13 1
4. node 14      15. node 15     16. node 16     17. node 17     18. node 18     19. node 19     20. node 20     21. node 21     22. node 22     23. node 23     24. node 24     25. node 25     26. node 26 2
7. node 27      28. node 28     29. node 29     30. node 30     31. node 31     32. node 32     33. node 33     34. node 34     35. node 35     36. node 36     37. node 37     38. node 38     39. node 39 4
0. node 40      41. node 41     42. node 42     43. node 43     44. node 44     45. node 45     46. node 46     47. node 47     48. node 48     49. node 49     50. node 50     51. node 51     52. node 52 5
3. node 53      54. node 54     55. node 55     56. node 56     57. node 57     58. node 58     59. node 59     60. node 60     61. node 61     62. node 62     63. node 63     64. node 64     65. node 65 6
6. node 66      67. node 67     68. node 68     69. node 69     70. node 70     71. node 71     72. node 72     73. node 73     74. node 74     75. node 75     76. node 76     77. node 77     78. node 78 7
9. node 79      80. node 80     81. node 81     82. node 82     83. node 83     84. node 84     85. node 85     86. node 86     87. node 87     88. node 88     89. node 89     90. node 90     91. node 91 9
2. node 92      93. node 93     94. node 94     95. node 95     96. node 96     97. node 97     98. node 98     99. node 99     100. node 100   101. node 101   102. node 102   103. node 103   104. node 104
105. node 105   106. node 106   107. node 107   108. node 108   109. node 109   110. node 110   111. node 111   112. node 112   113. node 113   114. node 114   115. node 115   116. node 116   117. node 117
118. node 118   119. node 119   120. node 120   121. node 121   122. node 122   123. node 123   124. node 124   125. node 125   126. node 126   127. node 127   128. node 128   129. node 129   130. node 130
131. node 131   132. node 132   133. node 133   134. node 134   135. node 135   136. node 136   137. node 137   138. node 138   139. node 139   140. node 140   141. node 141   142. node 142   143. node 143
144. node 144   145. node 145   146. node 146   147. node 147   148. node 148   149. node 149   150. node 150   151. node 151   152. node 152   153. node 153   154. node 154   155. node 155   156. node 156
157. node 157   158. node 158   159. node 159   160. node 160   161. node 161   162. node 162   163. node 163   164. node 164   165. node 165   166. node 166   167. node 167   168. node 168   169. node 169
170. node 170   171. node 171   172. node 172   173. node 173   174. node 174   175. node 175   176. node 176   177. node 177   178. node 178   179. node 179   180. node 180   181. node 181   182. node 182
183. node 183   184. node 184   185. node 185   186. node 186   187. node 187   188. node 188   189. node 189   190. node 190   191. node 191   192. node 192   193. node 193   194. node 194   195. node 195
196. node 196   197. node 197   198. node 198   199. node 199   200. node 200   201. node 201   202. node 202   203. node 203   204. node 204   205. node 205   206. node 206   207. node 207   208. node 208
209. node 209   210. node 210   211. node 211   212. node 212   213. node 213   214. node 214   215. node 215   216. node 216   217. node 217   218. node 218   219. node 219   220. node 220   221. node 221
222. node 222   223. node 223   224. node 224   225. node 225   226. node 226   227. node 227   228. node 228   229. node 229   230. node 230   231. node 231   232. node 232   233. node 233   234. node 234
235. node 235   236. node 236   237. node 237   238. node 238   239. node 239   240. node 240   241. node 241   242. node 242   243. node 243   244. node 244   245. node 245   246. node 246   247. node 247
248. node 248   249. node 249   250. node 250   251. node 251   252. node 252   253. node 253   254. node 254   255. node 255   256. node 256   257. node 257   258. node 258   259. node 259   260. node 260
261. node 261   262. node 262   263. node 263   264. node 264   265. node 265   266. node 266   267. node 267   268. node 268   269. node 269   270. node 270   271. node 271   272. node 272   273. node 273
274. node 274   275. node 275   276. node 276   277. node 277   278. node 278   279. node 279   280. node 280   281. node 281   282. node 282   283. node 283   284. node 284   285. node 285   286. node 286
287. node 287   288. node 288   289. node 289   290. node 290   291. node 291   292. node 292   293. node 293   294. node 294   295. node 295   296. node 296   297. node 297   298. node 298   299. node 299
300. node 300   301. node 301   302. node 302   303. node 303   304. node 304   305. node 305   306. node 306   307. node 307   308. node 308   309. node 309   310. node 310   311. node 311   312. node 312
313. node 313   314. node 314   315. node 315   316. node 316   317. node 317   318. node 318   319. node 319   320. node 320   321. node 321   322. node 322   323. node 323   324. node 324   325. node 325
326. node 326   327. node 327   328. node 328   329. node 329   330. node 330   331. node 331   332. node 332   333. node 333   334. node 334   335. node 335   336. node 336   337. node 337   338. node 338
339. node 339   340. node 340   341. node 341   342. node 342   343. node 343   344. node 344   345. node 345   346. node 346   347. node 347   348. node 348   349. node 349   350. node 350   351. node 351
352. node 352   353. node 353   354. node 354   355. node 355   356. node 356   357. node 357   358. node 358   359. node 359   360. node 360   361. node 361   362. node 362   363. node 363   364. node 364
365. node 365   366. node 366   367. node 367   368. node 368   369. node 369   370. node 370   371. node 371   372. node 372   373. node 373   374. node 374   375. node 375   376. node 376   377. node 377
378. node 378   379. node 379   380. node 380   381. node 381   382. node 382   383. node 383   384. node 384   385. node 385   386. node 386   387. node 387   388. node 388   389. node 389   390. node 390
391. node 391   392. node 392   393. node 393   394. node 394   395. node 395   396. node 396   397. node 397   398. node 398   399. node 399   400. node 400   401. node 401   402. node 402   403. node 403
404. node 404   405. node 405   406. node 406   407. node 407   408. node 408   409. node 409   410. node 410   411. node 411   412. node 412   413. node 413   414. node 414   415. node 415   416. node 416
417. node 417   418. node 418   419. node 419   420. node 420   421. node 421   422. node 422   423. node 423   424. node 424   425. node 425   426. node 426   427. node 427   428. node 428   429. node 429
430. node 430   431. node 431   432. node 432   433. node 433   434. node 434   435. node 435   436. node 436   437. node 437   438. node 438   439. node 439   440. node 440   441. node 441   442. node 442
443. node 443   444. node 444   445. node 445   446. node 446   447. node 447   448. node 448   449. node 449   450. node 450   451. node 451   452. node 452   453. node 453   454. node 454   455. node 455
456. node 456   457. node 457   458. node 458   459. node 459   460. node 460   461. node 461   462. node 462   463. node 463   464. node 464   465. node 465   466. node 466   467. node 467   468. node 468
469. node 469
Root :
70
```

```
Authority Ranking for Top 5:
        Node 96 : 0.60381532
        Node 43 : 0.36415201
        Node 92 : 0.32874312
        Node 24 : 0.26473884
Hubs Ranking for Top 5:
        Node 369 : 0.46960050
        Node 70 : 0.45809438
        Node 43 : 0.42851864
        Node 109 : 0.35680787
```

If I choose graph 9(IMDB), the Node will replace to Name :

```
Authority Ranking for Top 5:
        Kevin Spacey : 0.73032067
        Gabriel Byrne : 0.36512444
        Patrick Stewart : 0.36512444
        Ian McKellen : 0.36512444
Hubs Ranking for Top 5:
        Bryan Singer : 0.81647518
        Sam Mendes : 0.40826970
        Curtis Hanson : 0.40826970
        Annette Bening : 0.00000000
```