Data Mining, Project 3 – Link Analysis

劉俊毅, N16064103

1. Implementation detail

This project is to conduct link analysis on graphs, and the graphs are in txt form. So first of all, the graphs are loaded and processed in function ***graphProcess***. In this function, all nodes are found and because each row represents link of 2 nodes, parents and children of each node can be built using ***class link***. The function will return all the nodes and the subgraph of the processed graphs.

1.1 HITS

After all the nodes and the subgraph are completed by function *graphProcessed,* HITS can be implemented. **Initial values 1** are set for all nodes of graphs, and then iteration starts. While-loop is used here for the iteration and the error should be less than 1e-6 to leave the loop. Note also that the maximum number of iteration is set here to avoid condition of not convergence. Inside the while-loop for iteration, if a node has at least one parent, its authority is summed by the hub of its parent. Similarly, if a node has at least one child, its hub is summed by the authority of its child. See fig.1 below for details:

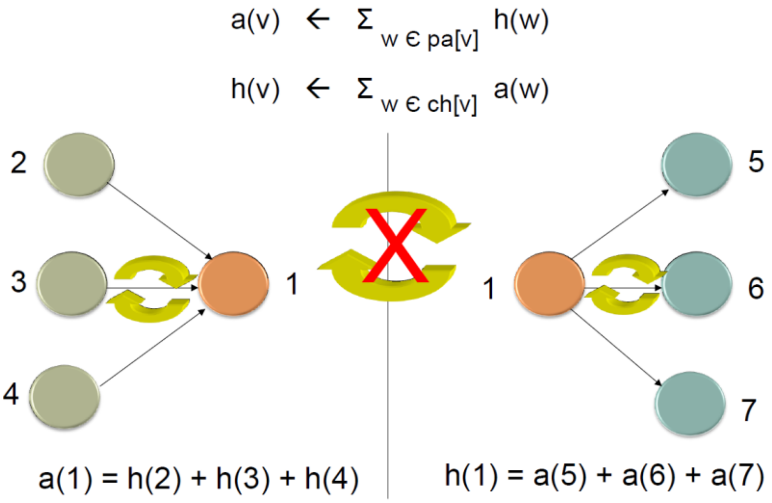


Fig.1 Equations and scheme for authority and hub calculation in HITS

The calculated authority and hub of nodes are **normalized** and the **difference** with values of previous step is calculated. Not until $\|a_t - a_{t-1}\| + \|h_t - h_{t-1}\| < \varepsilon$ would iteration stop, where the convergent criteria ε can be adjusted. Finally, function ***show_AuthHub*** shows the final authorities and hubs of nodes in descending order.

1.2 PageRank

Again, graphs are processed as in HITS implementation. The initial value of each node is now **1/ (number of nodes)** instead of 1. In the while-loop, if a node in subgraph has at least one parent, the PageRank value is calculated by the sum of its PageRank values (of previous iteration step) of its parents divided by its parents' outdegree (number of links outward). Moreover, the random jumping probability is considered here, for the linking from a random page, see eq. (1) for generalized formula

$$PR(u) = \frac{\lambda}{N} + (1-\lambda) \cdot \sum_{v \in B_u} \frac{PR(v)}{L_v} \qquad \text{eq.(1)}$$

The difference of PageRank value for nodes are calculated to get the error used in judgement of stopping the iteration loop, here the criteria is again set as 1e-6 or steps over 1e3 to jump out of iteration.

1.3 SimRank

Same graph procession as HITS and PageRank is conducted. The initial value of the similarity is set as an **identity matrix** because the property S(a,a) = 1. In the for-loop, the calculation of S(a,b) is bypassed if a equals to b or node a or b has no parent again utilizing the property of identity matrix. For rest of pairs of a and b. The similarity values are calculated eq. (2)

$$S(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} S(I_i(a), I_j(b)) \qquad \text{eq.(2)}$$

the similarity matrix is returned with the set decay factor C and the number of iteration. Here set C = 0.8 and iteration = 100.

2. Result analysis and discussion

- **Graph_1**

Graph_1 is a straight line, meaning that node 1 has no authority because of no parent and node 6 has no hub due to no child. Also, the values of authority and hub are the same with just a node translation. See table 1 for the results of HITS and PageRank, for this and all graphs, the convergent criteria are both 1e-6 of error. The result of SimRank for graph_1 is an identity matrix with size 6*6.

Table 1 Results of HITS and PageRank for graph_1

Node#: value		
	HITS	PageRank
Authority	6: 0.347994 5: 0.241080 4: 0.174838 3: 0.132312 2: 0.103778	
Hub	5: 0.347994 4: 0.241080 3: 0.174838 2: 0.132312 1: 0.103778	
PageRank		6: 0.103808 5: 0.0927158 4: 0.0796656 3: 0.0643125 2: 0.04625 1: 0.025

- **Graph_2**

Graph_2 is a circle, in which the last node links back to the first. See table 2 for the results. Note that for HITS, it fails to converge, having the same error after 30 iterations. So, the while-loop should be terminated by the utmost iteration steps. However, for PageRank, the error is 0 and converge after 1 step. The result of SimRank for graph_2 is again a 5*5 identity matrix.

Table 2 Results of HITS and PageRank for graph_2

	HITS	PageRank
Authority	5: 0.334495 4: 0.230556 3: 0.166732 1: 0.142227 2: 0.12599	
Hub	5: 0.269089 4: 0.267634 3: 0.196147 2: 0.149434 1: 0.117696	
PageRank		All 0.2

- **Graph_3**

Graph_3 is like a chain with node 1 and 4 being two ends. See table 3 for the results. The result of SimRank for graph_3 is shown in the form of similarity matrix, see table 4. Here the decay factor is set as 0.8 and the number of iteration is 100. From table 4, S(1,3) = S(2,4) = 2/3 and S(1,2) = S(2,3) = S(3,4) = 0.

Table 3 Results of HITS and PageRank for graph_3

	HITS	PageRank
Authority	3: 0.414055 2: 0.322562 4: 0.314672 1: 0.1651572	
Hub	3: 0.414055 2: 0.322562 4: 0.314672 1: 0.1651572	
PageRank		2: 0.324561 3: 0.324561 1: 0.175439 4: 0.175439

Table 4 Similarity matrix of graph_3 (C=0.8, iteration step = 100)

	1	2	3	4
1	1.0	0	2/3	0
2	0	1.0	0	2/3
3	2/3	0	1.0	0
4	0	2/3	0	1.0

- **Graph_4**

The result is shown in table 5. From the table, node 5 has highest authority and hub values by HITS, however, PageRank of node 1 is larger than node 5. See also, table 6 for result of SimRank for graph_4. It can be seen that S(4,6) and S(4,7) has the largest value. From the graph, node 4 and 6 are both linked by and linking to node 5, and node 4 and 7 are both linked by node1 and linking to node 5.

Table 5 Results of HITS and PageRank for graph_4

	HITS	PageRank
Authority	5: 0.445389 3: 0.248488 4: 0.205901 6: 0.170966 1: 0.165608 7: 0.165362 2: 0.164249	
Hub	5: 0.310805 6: 0.295922 1: 0.277494 4: 0.274272 7: 0.254559 3: 0.0981589 2: 0.0476097	
PageRank		1: 0.280288 5: 0.184198 2: 0.158765 3: 0.138882 4: 0.108220 7: 0.0690775 6: 0.0605706

Table 6 Similarity matrix of graph_3 (C=0.8, iteration step = 100)

	1	2	3	4	5	6	7
1	1.0	0.3603	0.3490	0.3537	0.3377	0.4151	0.2924
2	0.3603	1.0	0.4068	0.3697	0.4122	0.2854	0.4541
3	0.3490	0.4068	1.0	0.4496	0.3901	0.4481	0.4510
4	0.3537	0.3697	0.4496	1.0	0.3427	0.5351	0.5351
5	0.3377	0.4122	0.3901	0.3427	1.0	0.2731	0.4122
6	0.4151	0.2854	0.4481	0.5351	0.2731	1.0	0.2701
7	0.2924	0.4541	0.4510	0.5351	0.4122	0.2701	1.0

- **Graph_5**

The result is shown in table 7. For HITS, the authority values differ slightly for some nodes, however, the hub value of node 468 is way larger than the rest. For PageRank, all values are small. The result of SimRank is a 469*469 matrix, because of its large size, it’s not shown here.

Table 7 Results of HITS and PageRank for graph_5

	HITS	PageRank
Authority	461: 0.209373 438: 0.181190 430: 0.158892 ...	
Hub	468: 0.704729 412: 0.040284 ...	
PageRank		61: 0.002863 122: 0.002818 ...

- **Graph_6**

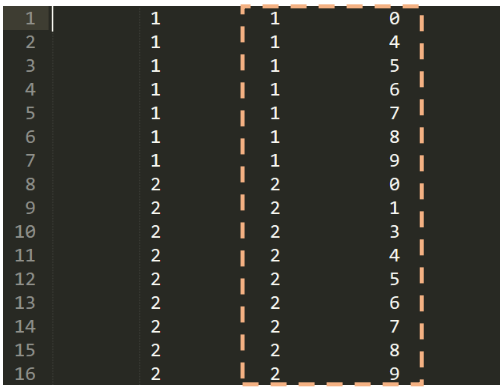
The result is shown in table 8. For PageRank, all values are so small.

Table 8 Results of HITS and PageRank for graph_6

	HITS	PageRank
Authority	709: 0.195676 ...	
Hub	1183: 0.459847 ...	
PageRank		1052: 0.000692381 ...

- **Graph_IBM**

The IBM transaction data of project I is used here. See fig. 2 for part of the data, here the second and third column is my first and second node of every edge, so, the graph has 10 nodes (0-9) and 16 edges. Table 9 shows the result of this graph from IBM transaction data.



```
        1       1       0
        1       1       4
        1       1       5
        1       1       6
        1       1       7
        1       1       8
        1       1       9
        2       2       0
        2       2       1
        2       2       3
        2       2       4
        2       2       5
        2       2       6
        2       2       7
        2       2       8
        2       2       9
```

Fig. 2 Graph from IBM transaction data (10 nodes, 16 edges)

Table 9 Results of HITS and PageRank for graph_IBM

	HITS	PageRank
Authority	9: 0.385314 3: 0.337942 …	
Hub	2: 0.805292 1: 0.446906	
PageRank		0,4,6,7,8,9: 0.0186087 5: 0.017015 1,3: 0.0165938 2: 0.015

The graph can be viewed as two centers (node 1 and 2) linking outward to other nodes while node 2 has a link to node 1. So, the link from node 2 to 1 make hub value of node 2 the largest.

3. Computation performance analysis

3.1 Comparison of execution time and number of iteration

For HITS and PageRank, the convergent criteria ε is set to be 1e-6 and the maximum iteration step is set as 1000 to avoid infinite loop of iteration. For SimRank, the number of iteration is 100 and the decay factor C is 0.8.

Table 10 Execution time of the 7 graphs using three algorithms (in second)

	HITS	PageRank	SimRank
Graph_1	0.001	0.0	0.004
Graph_2	not converge	0.0	0.003
Graph_3	0.005	0.001	0.003
Graph_4	0.002	0.001	0.016
Graph_5	0.113	0.008	59.6544
Graph_6	1.08	0.03	
Graph_IBM	0.001	0.0	

Table 11 Number of iteration for the 7 graphs using HITS and PageRank

	HITS	PageRank
Graph_1	13	7
Graph_2	not converge	1
Graph_3	75	16
Graph_4	28	17
Graph_5	60	14
Graph_6	108	11
Graph_IBM	11	4

3.2 Computational complexity

See table 12 for the complexity of time and space for the three algorithms.

	HITS	PageRank	SimRank
Time	$O(\mathrm{n}^2)$	$O(\mathrm{n}^2)$	$O(\mathrm{n}^3)$
Space	O(n)	O(n)	O(n)

4. Discussion

In doing this project, the first challenge I came across is to implement the linking relation between nodes. And I referred to my friend's way to do, which is using class to add child or parent based on the txt files of graph, and then using such relation of node to do the rest work of the algorithms. After the programming and the tests required for the completion of this report, I think I have get more acquainted with how links work.

To be frank, both the materials like PDFs and lectures by professor are a bit limited for i.e. the understanding of exactly how PageRank and SimRank works, some further materials or consults with TA are needed.

5. Questions & discussion (bonus)

1. There are some limitations of some link analysis algorithms, like for HITS, the one of circular graph fail to converge, so for graphs which are like a circle may be harder to converge for authority and hub values of nodes. In addition, for larger edges or nodes, or when the links are more complicated, SimRank is not efficient to implement. It's not included on the results above, but SimRank of graph_6 took more than 1000 seconds to execute. And the results are giant matrix.

2. Like what the professor have said on lecture or the material I read online, it’s not difficult for people with purpose to create webpage having lots of linking history, especially nowadays. So, this is a large problem, the more accessible the information, the more seriously we should take for their truthfulness and importance.

3. In real web, the number of nodes or webpages are extremely way more than these graphs we have coped with in the project. So, the issue might be the efficiency in doing link analysis and the complexity these analysis would deal with.

4. The effect of C in SimRank is like a parameter tuning the value of the final similarity matrix. Also, it can affect the efficiency of the calculation if inside the loop, 0 will be returned when the value is lower than a certain number. In my case, it is set as 0.8. It can be adjusted with the number of iteration set for distinguishability of result.