COMP631 Introduction to IR
Homework 2 Solution

March 22, 2022

1 HITS and PageRank (35 pts)

Given the figure including a graph as below:

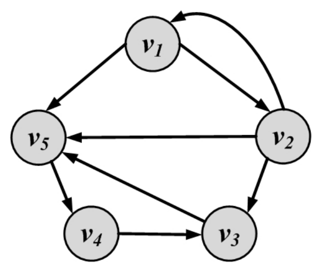

Figure 1: A toy example of a graph

1) (8 pts) For each graph node in Figure 1, calculate the degree centrality *normalized by the degree sum*, as well as the rank with respect to the centrality.

Solution:

We use in-degree as the centrality value. Then, according to table beneath, the normalization factor = 1+1+2+1+3 = 8

Node	In-degree	Centrality	Rank
1	1	1/8	3
2	1	1/8	3
3	2	2/8	2
4	1	1/8	3
5	3	3/8	1

Note: Out-degree or In-degree calculating strategies are both acceptable, but normalization factor depends on which calculating strategies you used.

2) (8 pts) In the context of world wide web or social media, given a concrete real-world example for entities that can be recognized as "authorities".

Solution: Influential people on Instagram.

Note: The answers are various, including the upper example but not limited.

3) (12 pts) Calculate the PageRank values for the graph in Figure 1 by applying the iterative algorithm. (Please show 2 iteration steps.)

Solution:

Note: d is not required. Thus, the formula of PageRank iterations can be either Equation(1) or Equation(2).

- When $d = 1$, PageRank iterations:

$$C_{Page}(v_i) = d\Big(\sum_{v_j \in \mathcal{N}(v_i)} \frac{C_{Page}(v_j)}{d_j^{OUT}}\Big) \tag{1}$$

- When $d \neq 1$, PageRank iterations:

$$C_{Page}(v_i) = d\Big(\sum_{v_j \in \mathcal{N}(v_i)} \frac{C_{Page}(v_j)}{d_j^{OUT}}\Big) + \frac{1-d}{N} \tag{2}$$

  where d is damping factor, N is the number of node, and $N(v_i)$ denotes as the neighborhood set of v_i. In this case, we have $N = 5$, which implies that $(1-d)/N = 0.02$.

Here we demonstrate the answer when $d = 0.9$. We follow the denotation of C_i^t as the node i with t-th step to gain the following result from iterations:

Step	v_1	v_2	v_3
t	C_1^t	C_2^t	C_3^t
t+1	$d(C_2^t/3) + 0.02$	$d(C_1^t/2) + 0.02$	$d(C_2^t/3 + C_4^t) + 0.02$
Step	v_4	v_5	
t	C_4^t	C_5^t	
t+1	$d(C_5^t) + 0.02$	$d(C_1^t/2 + C_2^t/3 + C_3^t) + 0.02$	

After then, according to the upper table, we can have:

Step	v_1	v_2	v_3	v_4	v_5	Sum
1	0.2	0.2	0.2	0.2	0.2	1
2	0.08	0.11	0.26	0.2	0.35	1
3	0.053	0.056	0.233	0.335	0.323	1

4) (7 pts) Explain the limitations of Katz Centrality, and how PageRank overcome these limitations.

Solution:

- **Limitation:** A node has high Katz Centrality will pass all its centrality along its out-links in directed graph.
- **Overcoming strategy:** Divide the value of passed centrality by the degree of a nodes' outgoing links. This means that the formulation can be recast as follows:

$$C_{Katz}^{New}(v_i) = \alpha \sum_j A_{ij} \frac{C_{Katz}^{New}(v_j)}{d_j^{OUT}} + \beta \tag{3}$$

2 Graph Analysis (35 pts)

Given the graph as below:

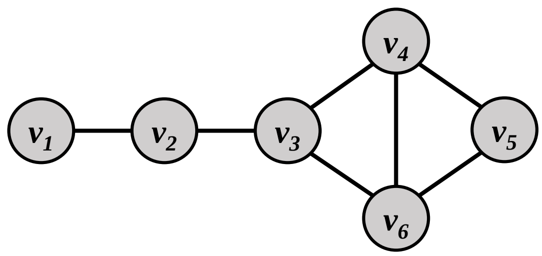

Figure 2: Another toy example of a graph

1) (7 pts) Calculate the Closeness centrality of node v_5.

Solution: $1/((4+3+2+1+1)/5) = \frac{5}{11}$

2) (7 pts) Calculate the Betweenness centrality of node v_4.

Solution: We first find the total number of shortest paths and those that pass through v_4 for all the node pairs.

$v_1v_2 : \frac{0}{1}$
$v_1v_3 : \frac{0}{1}$
$v_1v_5 : \frac{1}{2}$
$v_1v_6 : \frac{0}{1}$
$v_2v_3 : \frac{0}{1}$
$v_2v_5 : \frac{1}{2}$
$v_2v_6 : \frac{0}{1}$
$v_3v_5 : \frac{1}{2}$
$v_3v_6 : \frac{0}{1}$
$v_5v_6 : \frac{0}{1}$

Thus, the Betweenness centrality of node v_4 is $2 \times (\frac{0}{1} + \frac{0}{1} + \frac{1}{2} + \frac{0}{1} + \frac{0}{1} + \frac{1}{2} + \frac{0}{1} + \frac{1}{2} + \frac{0}{1} + \frac{0}{1}) = 3$ or $(\frac{0}{1} + \frac{0}{1} + \frac{1}{2} + \frac{0}{1} + \frac{0}{1} + \frac{1}{2} + \frac{0}{1} + \frac{1}{2} + \frac{0}{1} + \frac{0}{1}) = \frac{3}{2}$

Note: Doubling the sum is not required for calculating the Betweenness centrality .

3) (7 pts) Calculate the Jaccard similarity between v_3 and v_5.

Solution: The neighbors of v_3 are $\{v_2, v_4, v_6\}$. The neighbors of v_5 are $\{v_4, v_6\}$. The Jaccard similarity is $\frac{|\{v_2,v_4,v_6\} \cap \{v_4,v_6\}|}{|\{v_2,v_4,v_6\} \cup \{v_4,v_6\}|} = \frac{|\{v_4,v_6\}|}{|\{v_2,v_4,v_6\}|} = \frac{2}{3}$

4) (7 pts) Calculate the local clustering coefficient of v_4.

Solution: The neighbors of v_4 are $\{v_3, v_5, v_6\}$. The number of pairs of the neighbors are 2, i.e., $\langle v_3, v_6 \rangle$ and $\langle v_5, v_6 \rangle$. There are 3 possible links among $\{v_3, v_5, v_6\}$. Therefore, the local clustering coefficient of v_4 is $\frac{2}{3}$

5) (7 pts) Explain how we can use similarity measure or clustering coefficient for friend recommendation in social network.

Solution: Two users who are not connected in a social network will be likely to know each other if they tend to have a similar friend circle. Thus, we can recommend them to be friends. Formally, if the Jaccard similarity of two users v_i and v_j is large, which means they have many common friends, then we recommend them to be friends. Similarly, if a node v_i has a large local clustering coefficient, which means many of v_i's friends know each other, then we can recommend those neighbors who are not connected to be friends.

3 Results Assembly (30 pts)

1) (12 pts) For a multi-term query q, explain why we can only consider high-idf query terms for index elimination.

Solution: 1) Terms with low-idf values have little influence on the similarity score, e.g., cosine similarity, between query and document. 2) It can reduce the computational costs of calculating similarity scores.

2) (18 pts) Consider a more general form of net score,

$$netscore(q, d) = \alpha \cdot g(d) + \beta \cdot cosine(q, d), \tag{4}$$

where α and β are weights, $0 \geq \alpha \geq 1$, $0 \geq \beta \geq 1$, $\alpha + \beta = 1$, explain how the IR system would behave with different relative weighting.

Solution: We discuss three possible cases below.

- When $\alpha \gg \beta$, $g(d)$ dominates $netscore(q, d)$. We are likely to get important items, which however may not be relevant to our query.
- When $\alpha \ll \beta$, $g(d)$, $cosine(q, d)$ dominates $netscore(q, d)$. We are likely to retrieve relevant items to our query, but they may not come from popular information sources, and the authority of the retrieved items is not guaranteed.
- When $\alpha \approx \beta$, $netscore(q, d)$ coordinates between the quality and relevancy of the retrieved items. We are likely to obtain items that are of high quality and are relevant with our query.