Ranking of Network Data and its Application to Chinese Mainland University Data

MA Chiyu WU Yuqia

Dept. of Applied Mathematics
The Hong Kong Polytechnic University

CSIC 5011, May 2021

Table of Contents

Table of Contents

Introduction

World wide web $G = (V, E, W)$ W_{ij}: number of links from website i to j

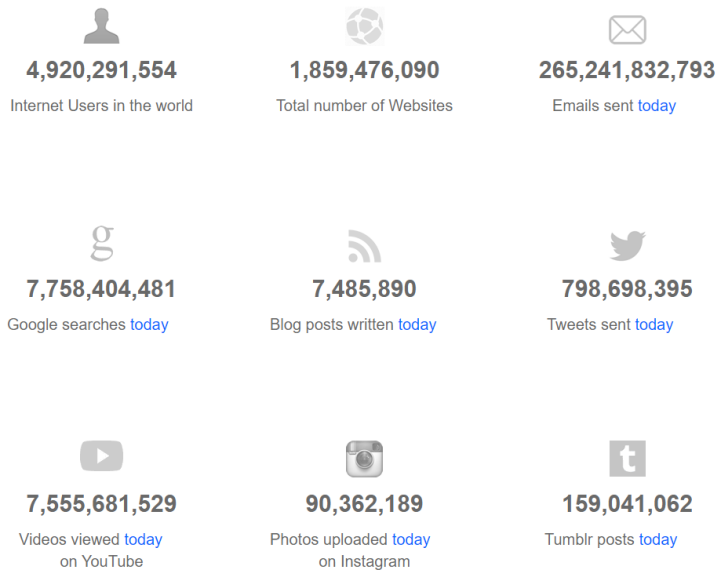


Figure: Screenshot in May 17th, from https://www.internetlivestats.com/

Question: How to rank websites by the importance?

Data

76 universities of Chinese mainland
`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/univ_cn.mat`

	1	1	3	4	5
ResearcRank	pku	tsinghua	fudan	nju	zju

Rank correlation coefficient:

- Spearman's ρ:

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

- Kendall's τ:

$$\frac{2(\text{(number of concordant pairs)} - \text{(number of discordant pairs)})}{n(n-1)}$$

Table of Contents

PageRank Ranking

- Visits to a websites $\rightarrow$ Markov chain on graph $G = \{V, E, W\}$
- Transition probability $P = \{\text{Prob}\,(x_{t+1} = j \mid x_t = i)\} \in \Re^{|V|\times|V|}$
- $\pi \geq 0$, $\pi^T P = \pi$ and $1^T \pi = \pi$(Perron theorem).
  π: the equilibrium distribution, determines the ranking.

$$P_1 = D^{-1} W, \quad D := diag\left(\sum_{j=1}^{|V|} \omega_{ij}\right).$$

$$P_\alpha = \alpha P_1 + (1-\alpha) E.$$

Spearman's ρ among PageRank_α with different α

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.85	0.9
0.1	1.000	0.913	0.679	0.634	0.490	0.500	0.505	0.536	0.514
0.2	0.913	1.000	0.765	0.716	0.569	0.584	0.561	0.557	0.540
0.3	0.679	0.765	1.000	0.857	0.716	0.637	0.642	0.631	0.600
0.4	0.634	0.716	0.857	1.000	0.813	0.680	0.625	0.626	0.605
0.5	0.490	0.569	0.716	0.813	1.000	0.705	0.700	0.608	0.611
0.6	0.500	0.584	0.637	0.680	0.705	1.000	0.890	0.786	0.794
0.7	0.505	0.561	0.642	0.625	0.700	0.890	1.000	0.756	0.744
0.85	0.536	0.557	0.631	0.626	0.608	0.786	0.756	1.000	0.965
0.9	0.514	0.540	0.600	0.605	0.611	0.794	0.744	0.965	1.000

Kendall's τ among PageRank_α with different α

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.85	0.9
0.1	1.000	0.814	0.533	0.476	0.344	0.363	0.378	0.389	0.379
0.2	0.814	1.000	0.663	0.590	0.419	0.448	0.434	0.413	0.396
0.3	0.533	0.663	1.000	0.754	0.575	0.488	0.523	0.478	0.450
0.4	0.476	0.590	0.754	1.000	0.691	0.552	0.498	0.488	0.462
0.5	0.344	0.419	0.575	0.691	1.000	0.584	0.585	0.469	0.461
0.6	0.363	0.448	0.488	0.552	0.584	1.000	0.780	0.695	0.682
0.7	0.378	0.434	0.523	0.498	0.585	0.780	1.000	0.649	0.634
0.85	0.389	0.413	0.478	0.488	0.469	0.695	0.649	1.000	0.913
0.9	0.379	0.396	0.450	0.462	0.461	0.682	0.634	0.913	1.000

Comparison between PageRank_α and ResearchRank

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.85	0.9
Kendall's τ	0.481	0.483	0.488	0.496	0.500	0.508	0.502	0.510	0.511
Spearman's ρ	0.666	0.668	0.675	0.682	0.686	0.698	0.693	0.703	0.704

	1	1	3	4	5
ResearcRank	pku	tsinghua	fudan	nju	zju
	1	2	3	4	5
PageRank_0.85	tsinghua	pku	sjtu	nju	uestc
PageRank_0.9	tsinghua	pku	sjtu	nju	uestc

Table of Contents

Traffic Ranking and Temperature Ranking

- WWW: an unweighted directed graph $G = \{V, E\}$.
- y_{ij} = number of users following (i, j) per unit time, $\quad (i, j) \in E$.
- $\sum_{j|(i,j)\in E} y_{ij} - \sum_{j|(j,i)\in E} y_{ji} = 0 \quad (i = 1, \ldots, |V|)$.

Entropy Maximum Model with Network Flow Constraints

$$\max - \sum_{(i,j)\in E} p_{ij} \log p_{ij}$$

$$\text{s.t.} \sum_{(i,j)\in E} p_{ij} = 1;$$

$$\sum_{i|(i,j)\in E} p_{ij} - \sum_{i|(j,i)\in E} p_{ji} = 0, \quad j \in V;$$

$$p_{ij} \geq 0, \quad (i, j) \in E.$$

Traffic Ranking and Temperature Ranking

Cross Entropy with Prior Distribution

- Weighted Graph $G = \{V, E, W\}$.
- Prior Distribution: $\omega_{ij} := \frac{W_{ij}}{\sum_{(i,j)\in E} W(i,j)}$.

$$-\sum_{(i,j)\in E} p_{ij} \log p_{ij} \quad \rightarrow \quad -\sum_{(i,j)\in E} p_{ij} \log \left(\frac{p_{ij}}{\omega_{ij}} \right)$$

Random Surfer on WWW

$$W \quad \rightarrow \quad W + 11^T - I$$

I is the identical matrix. Minus I is the requirment of network flow.

Traffic Ranking and Temperature Ranking

- Traffic Ranking: $H_j := \sum_{i|(i,j)\in E} p_{ij}$
- Temperature Ranking: λ_i: Lagrange multipliers of conservation equations constraints.Temperature ranking: reversed order of λ_i.

	1	2	3	4	5
Traffic rank	pku	tsinghua	zsu	sjtu	ustc
Temperature rank	hzau	sjtu	hit	tju	swufe

- TrafficRank and ResearchRank:
  Spearman's ρ: 0.584. Kendall's τ: 0.420.
- TemperatureRank and ResearchRank:
  Spearman's ρ: 0.160. Kendall's τ: 0.116.

Table of Contents

Introduction to HITS algorithm

Classification:

- Webpages pointed by others: Authority
- Webpages point others: Hub

Write the score of each authority webpage by x_i, that of hub webpage by y_i. And form directed graph:

- All webpages form a directed graph $G(V, E)$
- V: webpages, E: hyperlinks.

Introduction to HITS algorithm

Run the iteration to obtain final score:

- $L_{ij} = w_{ij}$, if $e_{ij} \in E$. Otherwise, $L_{ij} = 0$.
- $x' = L^T y, y' = Lx', x = x'/\|x'\|, y = y'/\|y'\|$.

The iteration also can be written as:

- $c_1 x^{t+1} = L^T L x, c_2 y^{k+1} = L L^T y^k$.

The final scores are the principal eigenvector of $L^T L$ and LL^T, so we can use SVD to compute the stable scores:

- $L = UDV^T, y^* = U_1, x^* = V_1^T$.

Numerical result

Table: Hub ranking and authority ranking by HITS algorithm

	1	2	3	4	5
Hub rank	pku	ustc	zsu	sjtu	zju
Authority rank	tsinghua	pku	uestc	sjtu	nju

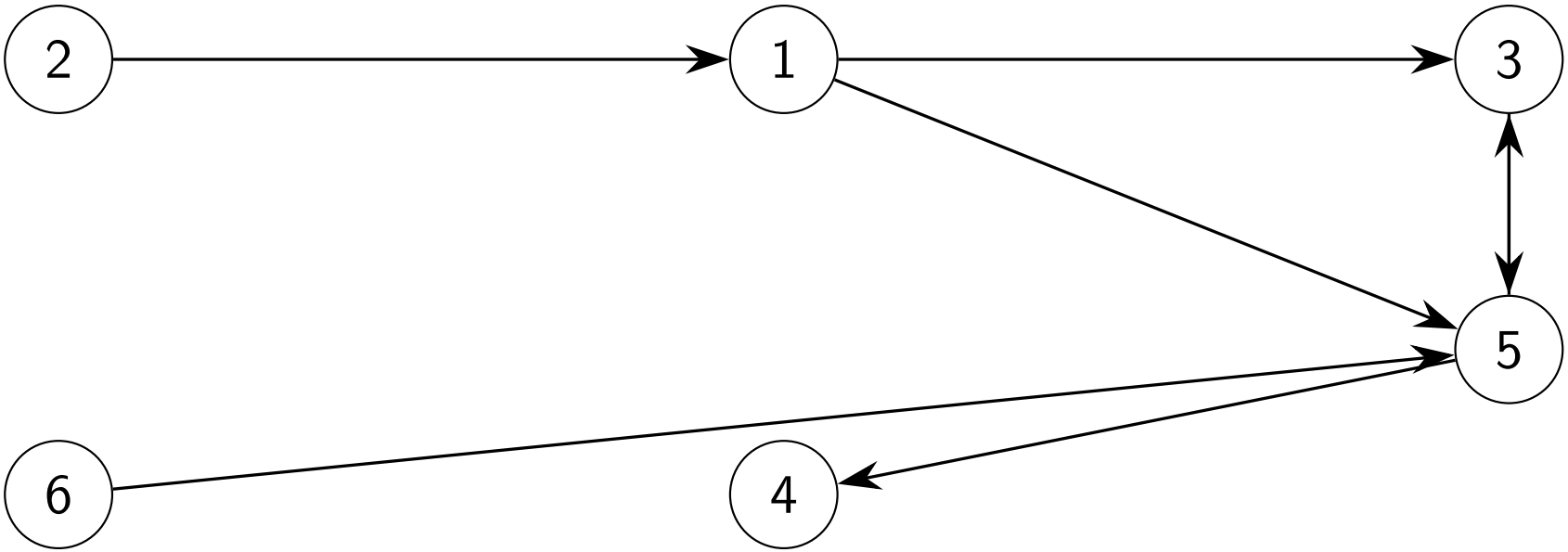
Table: Comparison between HITS and ResearchRank

	Hub ranking	Authority ranking
Spearman's ρ	0.5426	0.7487
Kendall's τ	0.3885	0.5741

Table of Contents

Introduction to SALSA algorithm

We introduce SALSA algorithm by a simple example:



Denote V_h the hub set and V_a the authority set. Then we can easily have

$$V_h = \{1, 2, 3, 5, 6\}, V_a = \{1, 3, 4, 5\}.$$

Introduction to SALSA algorithm

Preprocess data:

$$L = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$L_r = \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, L_c = \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \end{pmatrix}.$$

Introduction to SALSA algorithm

- Calculate the hub matrix H and authority matrix A.
- $H = (L_r L_c^T)_{V_h}, A = (L_c^T Lr)_{V_a}$.

$$H = \begin{pmatrix} \frac{5}{12} & 0 & \frac{2}{12} & \frac{3}{12} & \frac{2}{12} \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & 0 & \frac{3}{4} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{pmatrix}, A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{6} & 0 & \frac{5}{6} \end{pmatrix}.$$

- Find the connected component.
- $H_1 = \{2\}, H_2 = \{1, 3, 5, 6\}$.
- $A_1 = \{1\}, A_2 = \{3, 4, 5\}$.

Introduction to SALSA algorithm

Calculate the authority score: (hub score is similar)

- $score_i = \frac{C_j}{|C|}\frac{B_i}{E_j}$
- C_j: number of nodes of connected component
- $|C|$: number of nodes in graph
- B_i: number of outdegree of node i
- B_i: number of outdegree of the connected component

Final result:

- $\pi_A = (0.25, 0.25, 0.125, 0.375)$
- $\pi_H = (0.2667, 0.2, 0.1333, 0.2667, 0.1333)$
- Authority ranking: (1/5 2 3/6)
- Hub ranking: (5 1/3 4)

Numerical result

Table: Hub ranking and authority ranking by SALSA algorithm

	1	2	3	4	5
Hub rank	pku	ustc	zsu	njaun	sjtu
Authority rank	tsinghua	pku	uestc	sjtu	nju

Table: Comparison between SALSA and ResearchRank

	Hub ranking	Authority ranking
Spearman's ρ	0.4399	0.7220
Kendall's τ	0.3127	0.5508

Discussion

- The authority ranks of two ranking methods are similar
- 4-th and 5-th hub rank of HITS: sjtu (647), zju (383)
- 4-th hub 5-th rank of SALSA: njaun(688), sjtu(647)
- zju point 1/3 of its outdegree to pku
- This provide a way to cheak in ranking
- The correlation coefficient of results of HITS is relatively high

Table of Contents

Conclusion

- PageRank with large α, authority ranking by HITS and authority ranking by SALSA are more related to Research Rank.
- Traffic ranking, temperature ranking, hub ranking by HITS and hub ranking by SALSA perform worse.

Thank You!