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

Chiyu Ma
Department of Applied Mathematics
Hong Kong Polytechnic University
`chiyu.ma@connect.polyu.hk`

Yuqia Wu
Department of Applied Mathematics
Hong Kong Polytechnic University
`yuqia.wu@connect.polyu.hk`

Abstract

Given a directed graph (or network), ranking nodes according to their importance is an important problem. It is useful for search engine. One of the famous ranking methods is Google's PageRank algorithm. Beyond it, many novel methods like HITS, SALSA, TrafficRank, etc. are proposed. In this report we compare these ranking methods and the results.

1 Introduction

World wide web(WWW) could be regarded as a directed graph. Every website is a vertex and a link from website$i$ to website$j$ is an edge. How to rank them by the importance is a practical problem for search engine like Google. A famous method that Google proposed is PageRank[4], which adopted such a principle, *the "importance" of a page depends on the importance of pages pointing to it*. Though at 1998, there were only millions of websites, we have nearly two billian websites now.[1] Besides PageRank, other ranking methods like HITS, SALSA, TrafficRank, etc. were gradually proposed by scholars. In this report, we investigate these ranking methods by Chinese Mainland's Universities' data.

2 Data

The data we used is from `https://github.com/yao-lab/yao-lab.github.io/blob/master/data/univ_cn.mat`, which contains 76 universities of Chinese mainland. ResearchRank denotes the research ranking of these universities. And $W \in \Re^{76\times 76}$ is the link matrix whose (i,j)-th element gives the number of links from university i to j.

Table 1: ResearchRank

	1	1	3	4	5
ResearcRank	pku.edu.cn	tsinghua.edu.cn	fudan.edu.cn	nju.edu.cn	zju.edu.cn

3 PageRank Ranking

PageRank was the ranking method used by Google[4]. In this model, people's visits to a websites are assumed to be a Markov chain on graph $G = \{V, E, W\}$, which is a connected graph. The transition probability matrix $P = \{\mathbf{Prob}\,(x_{t+1} = j \mid x_t = i)\} \in \Re^{|V|\times|V|}(|V|$ is the number of vertice).This kind of row Markov matrix satisfies that $P_{ij} \geq 0$ and $P\mathbf{1} = \mathbf{1}$. According to Perron

[1] These data are from `https://www.internetlivestats.com/`.

Preprint. Under review.

theorem for nonnegative matrices, the max eigenvalue of P is 1 and there is a left eigen vector $\pi \geq 0$ s.t. $\pi^T P = \pi$ and $1^T \pi = \pi$. This eigen vector π is regarded as the equilibrium distribution.

In PageRank model, WWW is modeled as $G = \{V, E, W\}$. The transition according the links between two websites is reflected by

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

Additionally, the random visit to websites is modeled as $E = \frac{1}{|V|}\mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ is a vector consisting of $|V|$ ones. Considering these two kind of transitions, the final transition matrix is

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

In [4], $\alpha = 0.85$.

Applying the PageRank method to Chinese universities' data, we get a sequence of rankings PageRank_α with different α. We let $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.85, 0.9\}$ and compare these ranking results with the research ranking by Spearman's ρ and Kendall's τ. The comparison results are as Table 2. The comparison among different α is shown in Table 3 and Table 4.

Table 2: 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

Table 3: 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

Table 4: 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

Table 5: PageRank_α

	1	2	3	4	5
PageRank_0.85	tsinghua.edu.cn	pku.edu.cn	sjtu.edu.cn	nju.edu.cn	uestc.edu.cn
PageRank_0.9	tsinghua.edu.cn	pku.edu.cn	sjtu.edu.cn	nju.edu.cn	uestc.edu.cn

As we can see from these tables, the larger α, the closer PageRank and ResearchRank are. And similar brings the similar ranking. These results show that PageRank may reflect the research level of a university.

4 HITS Ranking

The HITS algorithm is widely used method for ranking retrieved webpages. This algorithm was proposed in [5]. In this article, the author called the webpage usually pointed to by a large number of hyperlinks authority. And on the other hand, a webpage that points to many authority webpages is called a hub. In context of literature citation, a hub is a review paper that cite many original paper, while an authority is an original seminal paper cited by many papers. The HITS algorithm assigns importance score to hubs and authorities, and computes them in a mutually reinforcing way: a good authority must be pointed to by several good hubs, while a good hub must point to several good authorities.

The set of webpages forms a directed graph $G = (V, E)$, where webpage p_i is a node in V and hyperlink e_{ij} is an edge in E. Based on this notation, we give the details of HITS algorithm. Firstly, each webpage is assigned two scores, authority score x_i and hub score y_i. The mutually reinforcing relationship is represented as

$$x'_i = \sum_{e_{ji}\in E} y_j, y'_i = \sum_{e_{ij}\in E} x_j, x_i = x'_i/\|x'_i\|, y_i = y'_i/\|y'_i\|, \tag{1}$$

where $\|\cdot\|$ is L_2 norm. Iteratively solving (1) and finally we will obtain stable solutions x^*_i, y^*_i. Write $L_{ij} = w_{ij}$ if $e_{ij} \in E$ and 0 otherwise. And then we obtain an adjacency matrix L. Then (1) can be written as

$$x' = L^T y, y' = Lx, x = x'/\|x\|, y = y'/\|y\|.$$

Let $x^{(t)}, y^{(t)}$ denote hub and authority score in t iteration. Then the iteration can be rewritten as

$$c_1 x^{(t+1)} = L^T L x^{(t)}, c_2 y^{(t+1)} = L L^T y^{(t)},$$

starting with $x^0 = y^0 = (1, 1, ..., 1)^T$, where c_i, $i = 1, 2$ are scaling parameter such that $\|x^{(t+1)}\| = \|y^{(t+1)}\| = 1$. By observation we know that the final solution x^*, y^* are the principal eigenvectors of $L^T L$ and LL^T, respectively. Thus we also can directly calculate the singular value decomposition of L to obtain the final solutions.

We use HITS on the dataset mentioned in Section 2. And we present the result below.

Table 6: Hub ranking and authority ranking by HITS algorithm

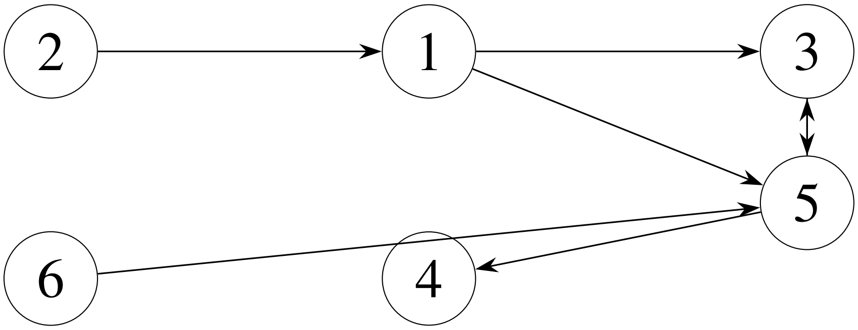
	1	2	3	4	5
Hub rank	pku.edu.cn	ustc.edu.cn	zsu.edu.cn	sjtu.edu.cn	zju.edu.cn
Authority rank	tsinghua.edu.cn	pku.edu.cn	uestc.edu.cn	sjtu.edu.cn	nju.edu.cn

Table 7: Comparison between HITS and ResearchRank

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

5 SALSA Ranking

In this section, we introduce another ranking method, named SALSA algorithm [6]. Rather than forming an adjacency matrix E, a bipartite undirected graph was built without any weight. To make this method well understand, we present a simple example.



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

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

Recall that HITS uses the adjacency matrix E to compute the authority and hub scores. On the other hand, PageRank computes a measure analogous to an authority score using a row-normalized weighted matrix. SALSA uses both row and column weighting to compute its hub and authority scores. Let L_r be be L with each nonzero row divided by its row sum and L_c be L with each nonzero column divided by its column sum. For our example,

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

Then, H, SALSA's hub matrix, consisting of the nonzero rows and columns of $L_r L_c^T$ and A is the nonzero rows and columns of $L_c^T L_r$. By simple calculation, we have

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

where the columns and rows of the above two matrices are corresponding to the hub set and authority set. Nextly, we need to find the connected component of each graph. For these small examples, we can find them by inspection. But for bigger graphs, where connectedness can not be determined by inspection, we can use graph traversal algorithm to identify them. Here both H and A contain two connected components with $H_1 = \{2\}, H_2 = \{1, 3, 5, 6\}$ and $A_1 = \{1\}, A_2 = \{3, 4, 5\}$. Now base on this, we can assign hub and authority scores on this nodes. Below we give the method for calculating authority score, and hub score is the same:

$$\text{score}_A = \frac{\text{number of nodes of connected component}}{\text{number of nodes of authority set}} \times \frac{\text{number of indegrees of nodes}}{\text{number of indegrees of connected component}}.$$

We calculate the authority score of node 3 for example. The number of connected component node 3 in is 3, and the number of authority set is 4. And the sum of indegree of the connected component is 6, while the indegree of node 3 is 2. So the authority score of node 3 is $\frac{3}{4} \times \frac{2}{6} = \frac{1}{4}$. Using this calculation method, the SALSA authority and hub scores are

$$\pi_A = (0.25, 0.25, 0.125, 0.375), \pi_H = (0.2667, 0.2, 0.1333, 0.2667, 0.1333).$$

Then the authority ranking is $(1/5\ 2\ 3/6)$ and the hub ranking is $(5\ 1/3\ 4)$., where / indicated a tie.

Now we apply SALSA to the dataset mentioned in Section 2. And the numerical result is

By inspection, we can notice that the first 5 authority rank of HITS are quiet the same as those of SALSA, while the hub rank have a little different. Now we give a closer look of the numerical results. For hub rank, the first 3 webpages are the same and the fourth and fifth rank of HITS are sjtu and zju,

Table 8: Hub ranking and authority ranking by SALSA algorithm

	1	2	3	4	5
Hub rank	pku.edu.cn	ustc.edu.cn	zsu.edu.cn	njau.edu.cn	sjtu.edu.cn
Authority rank	tsinghua.edu.cn	pku.edu.cn	uestc.edu.cn	sjtu.edu.cn	nju.edu.cn

Table 9: Comparison between SALSA and ResearchRank

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

while for SALSA, it is njau and sjtu. The total outdegree of zju and sjtu are 383, 647, respectively, while that of njau is up to 688. To this extend, the actual ranking of the webpage of njau is more likely to be higher than zju, whose ourdegree is only 383. But why zju ranks so high? In the dataset, we can notice that this webpage points nearly 1/3 of its hyperlinks to pku, whose authority rank is the first. This is one of the weakness of HITS algorithm. If one webpage uses many hyperlinks to point to some high-ranking webpages, then it will rank high. This mechanism provides a way to cheat. But for SALSA, it does not have this mutually reinforcing effect, so the ranking appears more fair. Another strength of SALSA is that, it does not need to much calculation compared to HITS.

However, compared to SALSA, the overall result for HITS are more similar with ResearchRank, since the correlation coefficient of HITS algorithm is larger. See Table 7 and Table 9. And we can see that, authority ranking is closer to ResearchRank than hub ranking.

6 Traffic Ranking and Temperature Ranking

[3] used the method of entropy maximum and conservation equations of network flow to rank websites. This method regards the WWW as an unweighted directed graph $G = \{V, E\}$. Assuming that there are Y users browsing the internet at one moment,

$$y_{ij} = \text{the number of users following link } (i,j) \text{ per unit time}, \quad (i,j) \in E.$$

The conservation equation of every vertex is

$$\sum_{j|(i,j)\in E} y_{ij} - \sum_{j|(j,i)\in E} y_{ji} = 0 \quad (i = 1, \dots, |V|).$$

Given that $Y := \sum_{(i,j)\in E} y_{ij}$, we can regard $p_{ij} := \frac{y_{ij}}{Y}$ as a distribution on edges.

Then apply the principle of maximal entropy in this ranking problem. The principle of maximal entropy aims to maximize the freedom in choosing the distribution $p_{ij}, (i,j) \in E$. Thus, this model could be formulated as a programming problem as below,

$$\begin{aligned} \max - &\sum_{(i,j)\in E} p_{ij} \log p_{ij} \\ \text{s.t.} &\sum_{(i,j)\in E} p_{ij} = 1; \\ &\sum_{i|(i,j)\in E} p_{ij} - \sum_{i|(j,i)\in E} p_{ji} = 0, \quad j \in V; \\ &p_{ij} \geq 0, \quad (i,j) \in E. \end{aligned}$$

However, our data has weighted matrix W. We can regard

$$\omega_{ij} := \frac{W_{ij}}{\sum_{(i,j)\in E} W(i,j)}$$

as a prior distribution. Thus we can replace the entropy in objective function with the cross entropy,

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

Additionally, just like the $\frac{1}{n}E$ in PageRank, we also need to give a random surfer. Thus we connect every vertex in the graph. And same with the weighted matrix excluding the diagonal elements (the diagonal elements should be zero because of the network flow model).

Supposed we have the solution of this programming problem, we can define the ranking result from the primal and dual solution. The primal solution is $p_{ij}, (i,j) \in E$. The traffic into the vertex j is

$$H_j := \sum_{i|(i,j)\in E} p_{ij}.$$

We can rank vertices by this index. This ranking result is called as TrafficRank. However, the Spearman's ρ between TrafficRank and ResearchRank is just 0.584. And Kendall's τ is just 0.420. This ranking may be not suitable for universities.

Every constraint has a dual solution, which is also the Lagrange multiplier of this constraint. [3] pointed out that Lagrange multipliers of conservation equations constraints, $\lambda_i, i \in V$, could be defined as the inverse of temperature of vertex i. This temperature is analogized from thermodynamics. Thus we can get the Temperature ranking from $1/\lambda_i$. However, this ranking result is far from the ResearchRank. The Spearman's ρ and Kendall's τ are just 0.160 and 0.116 respectively.

Table 10: TrafficRank and Temperature Rank

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

Remark:

1. [3] gives a heuristic method to solve the programming problem. However the initial point of that method is tricky. Interior point method is a suitable candidate to solve it.
2. The final model we use is the cross entropy with random surfer. We also tried the original information entropy with random surfer, information entropy without random surfer and cross entropy without random surfer. These combinations perform worse than cross entropy with random surfer, no matter in terms of optimality of solution nor the rank correlation coefficient with ResearchRank.

7 Conclusion

In this report, we compare several webpage ranking methods and give some necessary analysis on them.

8 Contribution

MA Chiyu completed section 3 and 6. WU Yuqia finished section 4 and 5.

References

[1] Yao, Yuan. A mathematical introduction to data science. preprint, 2017. `https://github.com/yao-lab/yao-lab.github.io/blob/master/book_datasci.pdf`

[2] Langville, Amy N., and Carl D. Meyer. Google's PageRank and beyond: The science of search engine rankings. Princeton university press, 2011.

[3] Tomlin, John A. "A new paradigm for ranking pages on the world wide web." Proceedings of the 12th international conference on World Wide Web. 2003.

[4] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." Computer networks and ISDN systems 30.1-7 (1998): 107-117.

[5] Kleinberg, J. M. Authoritative sources in a hyperlinked environment. In Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms (1999) pp. 668-677.

[6] Ronny Lempel and Shlomo Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In The Ninth International World Wide Web Conference, New York, 2000. ACM Press.