PageRank vs HITS

CSIC 5011: Topological and Geometric Data Reduction and Visualization Fall 2017

Ye Rougang, Tan Chunxi, Han Ruijian

November 12, 2017

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Abstract

In this project, we learned and used PageRank and HITS algorithm to get rankings from link matrix of webpages of 76 Chinese universities. As expected, the ranking from this perspective is shown to be distinct from the research ranking with Spearman and Kendall correlations. We took various alpha values in the PageRank to see the effect of random teleporting and compared HITS applied on the link matrix and adjacency matrix. Overall, ranking using webpage links offers another interesting viewpoint of ranking items.

Main Objectives

1. Find PageRank with Google's hyperparameter $\alpha = 0.85$.
2. Find HITS authority and hub ranking through adjacency matrix and link matrices.
3. Compare these rankings against the research ranking.
4. Find extended PageRank with various hyperparameters $\alpha \in (0, 1)$ investigate its effect on ranking.

Methods and Results

The result from PageRank is denoted as rank 1, while HITS through adjacency matrix is rank 2 and link matrix is rank 3. Some results are shown in the table 1. Since HITS gives us two results: one is authority, the other is hub, we use the former number to stand for authority ranking while later is hub ranking.

University	Rank 1	Rank 2	Rank 3
PKU	2	(70,7)	(2,1)
TSU	1	(1,16)	(1,10)
NJU	4	(61,15)	(13,17)
FDU	9	(29,34)	(76,16)
ZJU	11	(2,44)	(6,5)
USTC	12	(62,22)	(71,2)

Table 1: Weblink rankings

However, it seems the rankings with different approaches are distinct. In order to show that, we compare the difference between our results and the research ranking by using rank estimators such as Spearmans ρ and Kendalls τ .
Since we have two rankings with HITS, we use rank_21 and rank_22 to stand for authority and hub in rank 2 and do the same thing on rank 3. We also use rr to stand for research ranking. Then, we make comparisons among them. Main result has been showed in the table 2. The first column is the comparing rankings while the second and the third is Spearmans ρ and Kendalls τ. From the table, the research ranking has negative relationship with PageRank and HITS ranking, and two approaches of HITS are also different. It means links of webpages does not play positive effect on the research ranking to some extent.
In addition, we explore the effect of α in the PageRank. Here we also show the rank of PKU, TSU, NJU, FDU, ZJU and USTC under different α.

Comparison	Spearmans ρ	Kendalls τ
rr vs rank_1	-0.709	-0.527
rr vs rank_21	-0.210	-0.149
rr vs rank_22	-0.065	-0.049
rank_ 21 vs rank_31	-0.048	-0.039
rank_22 vs rank_32	0.749	0.574

Table 2: Comparison

The results has been showed in the figure 1. From the previous knowledge, we know that α gives an random effect on the dataset. Random effect is decreasing as α approaches to 1.
From the figure, the ranking does not change too much under different α. As we known, the less α means more randomness in the model. When α is small, the results is unreliable. But large α may also make model sensitive to the outliers. So choosing suitable α is very important.

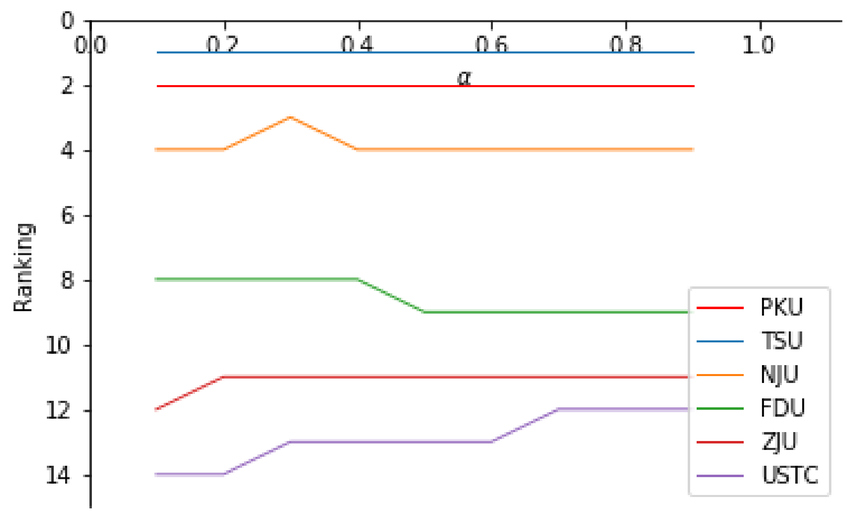

Figure 1: Rank under different α

Conclusions

In summary, one can see from the correlation values that the research ranking in the data is totally different from the web rankings here. As usually links of webpages of universities have no relation to the criteria of the traditional research ranking. The HITS algorithm with two various matrices seems to show opposite correlations. And we prefer to use the adjacency matrix to compute HITS ranking which keeps the same as the original algorithm.

References

[1] Jon M Kleinberg. Hubs, authorities, and communities. *ACM computing surveys (CSUR)*, 31(4es):5, 1999.

[2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Acknowledgements

Many thanks to our instructor Mr. Yao, who gives us the chance to deal with weblink rank dataset and the chance of this cooperation.