Chinese University Ranking Based on PageRank and HITS

Guiyu Cao and Yi-Su Lo

Department of Mathematics
The Hong Kong University of Science and Technology

Problem and data

- The following dataset contains 76 Chinese (mainland) University Weblink during 12/2001-1/2002, where **rank_cn** is the research ranking of universities in that year, **univ_cn** contains the webpages of universities, and **W_cn** is the link matrix whose (i,j)-th element gives the number of links from university i to j.

  https://math.stanford.edu/~yuany/course/data/univ_cn.mat

- In this mini-project, we try to analyze Chinese University Ranking by the following steps

(a) Compute PageRank with Google's hyperparameter $\alpha = 0.85$;
(b) Compute HITS authority and hub ranking;
(c) Compare these rankings against the research ranking;
(d) Compute extended PageRank with various hyperparameters $\alpha \in (0,1)$, investigate its effect on ranking.

Methods

Google's PageRank

- PageRank is an algorithm used by Google search to rank websites in this search engine result.
- For a directed weighted graph $G = (V, E, W)$ whose weight matrix decodes the webpage link structure

$$w_{ij} = \begin{cases} \sharp\text{link} : i \to j, \ (i,j) \in E, \\ 0, \ otherwise. \end{cases}$$

  Define an out-degree vector $d_i^o = \sum_{j=1}^n w_{ij}$, which measures the number of out-links from i. A diagonal matrix $D = diag(d_i)$ and a row Markov matrix $P_1 = D^{-1}W$, assumed that all nodes have non-empty out-degree.
- To obtain a unique stationary distribution, a primitive Markov Matrix follows the trick

$$P_\alpha = \alpha P_1 + (1-\alpha)F,$$
$$F = \frac{1}{n}\mathbf{1}\cdot\mathbf{1}^T,$$

  where F is a random surfer model, i.e. one can jump to any other webpage uniformly.
- With $1 > \alpha > 0$, the existence of random surfer model makes P a positive matrix, so

$$\exists!\ \pi \ \text{ s.t. } \ P_\alpha^T \pi = \pi.$$

  Google choose $\alpha = 0.85$ and in this case π gives PageRank scores.

Hyperlink-Induced Topic Search(HITS)

- HITS also known as hubs and authorities is a link analysis algorithm that rates Web pages.
- Define in-degree $d_k^i = \sum_{j=1}^n w_{jk}$. High out-degree webpages can be regarded as *hubs*, as they provide more links to others. On the other hand, high in-degree webpages are regarded as *authorities*, as they are cited by others intensively.
- HITS-authority. This use primary **right** singular vector of W as scores to give ranking. Since

$$L_a = W^T W,$$
$$L_a(i,j) = \sum_k W_{ki} W_{kj},$$

  the higher value of $L_a(i,j)$, the more references received on the pair of nodes, as $\sum_k \sharp\{i \leftarrow k \to j\}$.
- HITS-hub. This use primary **left** singular vector of W as scores to give ranking. Since

$$L_h = WW^T,$$
$$L_h(i,j) = \sum_k W_{ik} W_{jk},$$

  the higher value of $L_h(i,j)$, the more be hit from both i and j, as $\sum_k \sharp\{i \to k \leftarrow j\}$.

Spearman's rank and Kendall's rank correlation coefficient

- Spearman's rank and Kendall's rank correlation coefficient are used to measure association between two measured variables.
- Spearman's ρ. For a sample of size n, the n raw scores X_i, Y_i are converted to ranks rgX_i, rgY_i, and

$$\rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}},$$

  where $cov(rg_X, rg_Y)$ is covariance of the rank variables, σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.
- Kendall's τ. Any pair of observations (x_i, y_i) and (x_j, y_j), where $i \neq j$, are said to be concordant if both $x_i > x_j$ and $y_i > y_j$; or both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant if $x_i > x_j$ and $y_i < y_j$; or $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.
  The Kendall τ coefficient is defined as

$$\tau = \frac{n_{con} - n_{dis}}{n(n-1)/2},$$

  where n_{con} represent number of concordant pairs and n_{dis} represents number of discordant pairs.

Experiments and Results

Ranking by PageRank with hyperparameter $\alpha = 0.85$

Ranking	Authority ranking	Hub ranking	Page ranking
1	**THU**	**PKU**	**THU**
2	**PKU**	**USTC**	**PKU**
3	**UESTC**	**ZSU**	**SJTU**
4	**SJTU**	**NJAU**	**NJU**
5	**NJU**	**SJTU**	**UESTC**
6	**ZSU**	**THU**	**SCUT**
7	**Fudan**	**WHU**	**ZSU**
8	**SCUT**	**TJU**	**DLUT**
9	**SEU**	**SEU**	**Fudan**
10	**HUST**	**SDU**	**SEU**

Table 1: TOP 10 on Authority ranking, Hub ranking and PageRank ranking by PageRank algorithm.

- Authority ranking, hub ranking and page ranking by PageRank are different, as they focus on different target.
- Top 2 authority ranking is same as page ranking, and THU & PKU are behave well among these three types of ranking.

Ranking by HITS

Ranking	Authority ranking	Hub ranking	Research ranking
1	**THU**	**PKU**	**PKU**
2	**PKU**	**USTC**	**THU**
3	**UESTC**	**ZSU**	**Fudan**
4	**SJTU**	**SJTU**	**NJU**
5	**NJU**	**ZJU**	**ZJU**
6	**Fudan**	**SEU**	**USTC**
7	**ZSU**	**NJAU**	**SJTU**
8	**SCUT**	**WHU**	**BUAA**
9	**ZJU**	**TJU**	**NANKAI**
10	**GZSUMS**	**THU**	**TJU**

Table 2: TOP 10 on Authority ranking and Hub ranking by HITS algorithm, and the research ranking.

- Authority ranking, hub ranking and page ranking by HITS are different, as they focus on different target.
- Top 3 authority ranking and hub ranking by HITS are same as PageRank algorithm.

Spearman's and Kendall's correlation coefficient

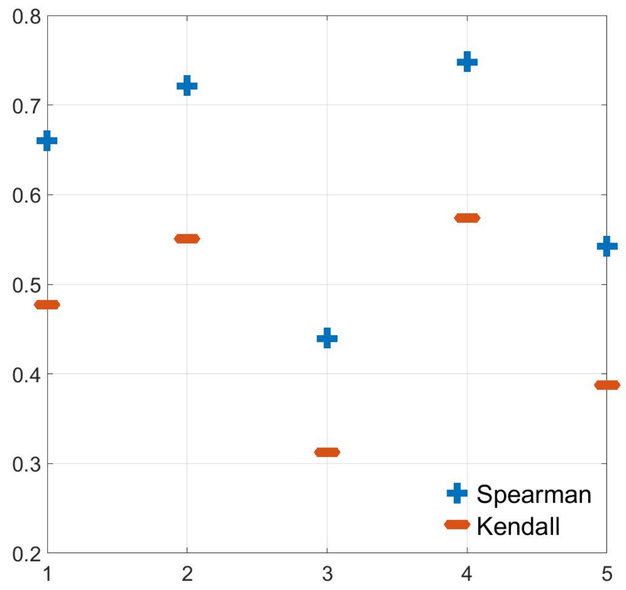

Figure 1: Rankings against the research ranking by Spearman's ρ and Kendall's τ. **1** represents Page ranking, **2,4** represent authority rank and hub rank respectively by PageRank algorithm; **3,5** represent authority rank and hub rank respectively by HITS algorithm.

- In all cases, coefficients of Spearman's ρ are higher than Kendall's τ.
- For these two Spearman's ρ and Kendall's τ, Page ranking always lies between authority ranking and hub ranking, respectively. Which shows Page ranking is a pretty good trade-off between authority ranking and hub ranking.

Extended PageRank

- As we enlarge the hyperparameter $\alpha \in [0,1)$, Kendall's and Spearman's correlation coefficient change.

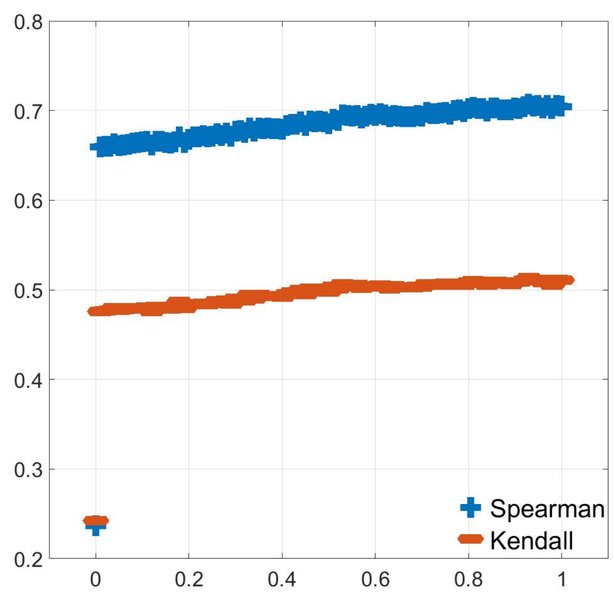

Figure 2: The effect of various hyperparameter α.

- When $\alpha = 0$, link matrix is not a positive matrix, as 0 appears in diagonal position. For nonnegative matrix, the primary vector is not vector, so using primary vector to gice PageRank score is not proper as coefficients ($\alpha = 0$) jumps into less 0.3 comparing with coefficients ($\alpha > 0$).
- As enlarge α, both Kendall's and Spearman's correlation coefficients enlarge which become more and more close to research rank, and $\alpha = 0.85$ is good enough as its one amazing trick.

Conclusions

In this project, we compute Page ranking, authority ranking and hub ranking by PageRank and HITS algorithm. As compare these rankings against research ranking by Kendall's and Spearman's correlation coefficient and investigate the effect of hyperparameters, we find $\alpha = 0.85$ chosen by Google is really one amazing trick.