PageRank

Presented by
Nathan Carlisle
&
Michael McDermott

Acknowledgements

•This presentation is based on the fourth section of “A Survey of Eigenvector Methods for Web Information Retrieval”, by Amy N. Langville and Carl D. Meyer, published in SIAM Review (2005).

Review

- Right eigenvector: $Av = \lambda v$
- Left eigenvector $v^T A = \lambda v^T$
- The Power Method

$$x^{(k+1)} = A x^{(k)}$$

$$x_{k+1}^T = x_k^T A$$

The Key Ideas Behind PageRank

- The rank of a page is found by using “votes” from other pages.

- These “votes” come in the form of links.

- Links from important sites carry more weight than links from unimportant sites.

- The more sites a source links to, the less important its vote.

This idea can be expressed in the following way:

$$r(P) = \sum_{Q \in B_p} \frac{r(Q)}{|(Q)|}$$

B_p = all pages pointing to P

$|Q|$ = number of out links from Q. This acts to normalize $r(P)$ in order to scale the results.

If we turn this into an iterative method using matrices, we set:

$$\pi_j^T = (r_j(P_1), r_j(P_2), \ldots, r_j(P_n))$$

and iteratively computing $\pi_j^T = \pi_{j-1}^T \boldsymbol{P}$

where $\boldsymbol{P}$ is the matrix with $p_{ij} = \frac{1}{|P_i|}$

if P_i links to P_j, 0 otherwise.

This is nothing more than the power method which will converge to a normalized eigenvector.

Computing PageRank

In order to compute PageRank we must solve the linear system

$$\boldsymbol{\pi}^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P})=0\,,\ \text{with}\ \boldsymbol{\pi}^{\boldsymbol{T}}e=1$$

Due to the size of the Google matrix (having 8 billion pages last year), the only feasible numerical method to solve the matrix is the Power Method.

Issues

- P may not be a stochastic matrix, because it may contain all zeroes in one row/column when the a node has an outdegree of zero.

- Replacing all of the zero rows with $\boldsymbol{e}^T/n$ gives a matrix $\bar{\boldsymbol{P}}$ that is stochastic, but could still be reducible.

- The solution, then, is to adjust the matrix in some way.

The Google Solution

- First, **P** is adjusted to be $\bar{\boldsymbol{P}}$,

- This makes irreducibility likely, but not guaranteed, so the matrix is further adjusted to $\bar{\bar{\boldsymbol{P}}}$.

- If the matrix is irreducible, then the power method is guaranteed to converge to the UNIQUE dominant eigenvector.

But How?

- It is done through the use of the following formula:

$$\bar{\bar{\boldsymbol{P}}} = \alpha \bar{\boldsymbol{P}} + (1-\alpha)\boldsymbol{E}$$

Where P-bar is the previously adjusted matrix, α is a scalar between 0 and 1, and **E** is the correction matrix.

The Parameter α

- The parameter α is a scalar value where $0<\alpha<1$.
  - Assuming a single dominant eigenvalue, convergence of the Power Method depends on $\left|\frac{\lambda_2}{\lambda_1}\right|$.
  - The farther α is from 1, the faster the convergence will be but the farther $\bar{\bar{\boldsymbol{P}}}$ will be from the original **P**.
  - However, the closer α is to 0, the slower the convergence will be, but $\bar{\bar{\boldsymbol{P}}}$ will be closer to **P**.
  - Google uses $\alpha \approx 0.85$

The Parameter *E*

- **E** is the personalization matrix.

- It is used by Google to either penalize (in the case of people trying to “cheat”) or promote (in the case of sponsors) websites.

- Has the effect of creating artificial links.

Updating PageRank

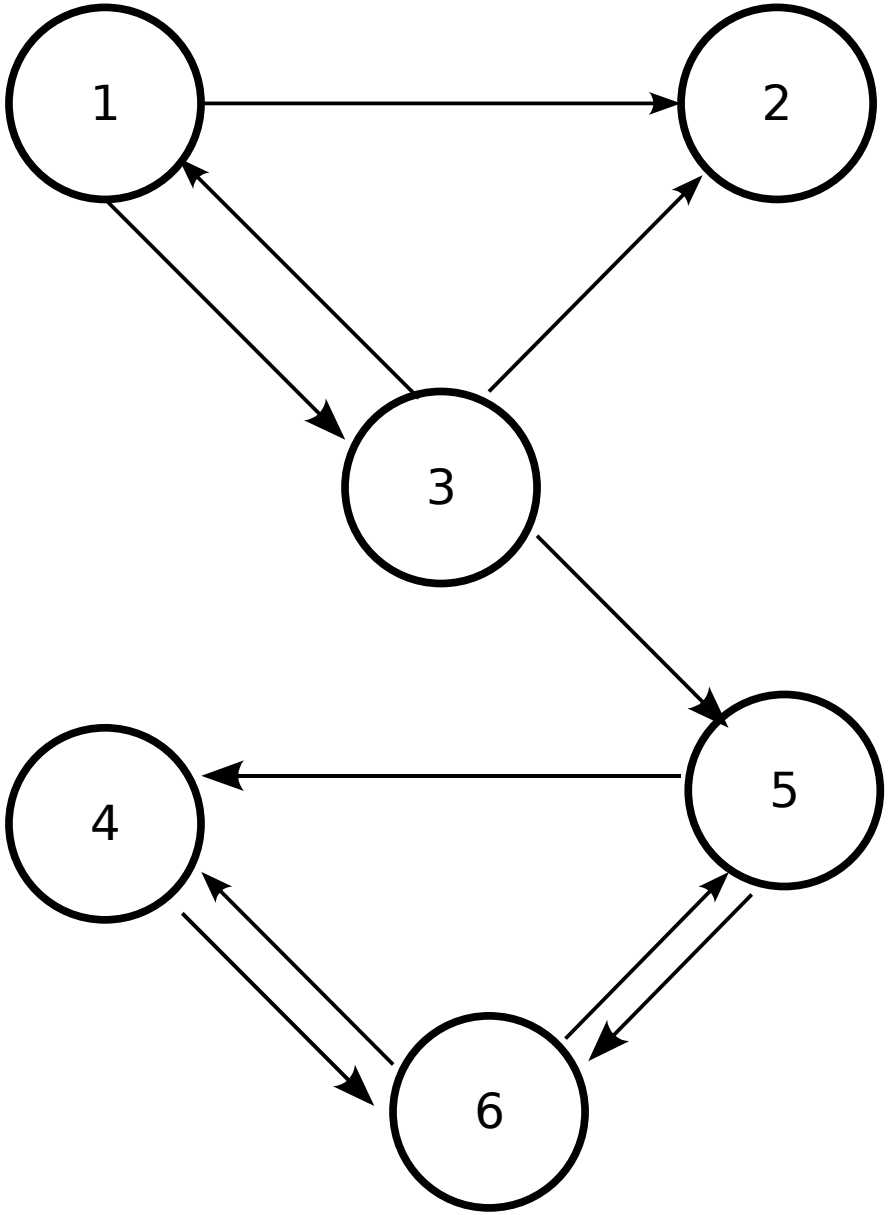
- In order to prevent the results from becoming stale, PageRank must be recomputed.

- Due to the expense of computing PageRank, the computations are done once every several weeks.

- When PageRank is recomputed, Google must start from scratch.

Example

Consider the following (small) internet:



We then create the following Google matrix **P**:

$$\boldsymbol{P}=\begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Seeing that there are all zeros in the second row, we modify **P** to get:

$$\bar{\boldsymbol{P}} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

In order to force irreducibility, we choose $\alpha = 0.9$ and $\boldsymbol{E} = \boldsymbol{e}\boldsymbol{e}^T/n$ and get:

$$\bar{\bar{\boldsymbol{P}}} = \alpha \bar{\boldsymbol{P}} + (1-\alpha)\boldsymbol{E} = \begin{pmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix}$$

This result is nothing more than the Google matrix for an Internet of six websites

From this we can find the PageRank vector, using the Power Method. These ranks are query-independent. Suppose the query is entered containing terms 1 and 2. From the term-document matrix we get:

$$\pi^T = (0.03721 \quad 0.05396 \quad 0.04151 \quad 0.4751 \quad 0.2060 \quad 0.2862)$$

These ranks hold regardless of the query, and tell us:

$$\begin{aligned} &\text{term 1} \rightarrow \text{document 1, document 4, document 6} \\ &\text{term 2} \rightarrow \text{document 1, document 3} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \end{aligned}$$

Comparing the ranks of these 4 documents, we list them in decreasing order:

- Document 4 (0.3751)
- Document 6 (0.2862)
- Document 3 (0.04151)
- Document 1 (0.03721)

Google used $\boldsymbol{E} = ee^T / n$ before (as in the example), and now it uses $\boldsymbol{E} = ev^T$

Conclusion

- Google works by using a matrix **P** of authority scores divided by hub scores.

- The matrix P is first adjusted to make it stochastic.

- The modified matrix is taken and modified to ensure irreducibility, using a factor α and a perturbation matrix **E**.

- The dominant eigenvector of this matrix is computed using the Power Method, the result of which is the PageRank vector.