Assignment 2

Rajbir Malik

2017CS10416

September 23, 2018

Report on **Personalized Page-rank**

Overview

Page Rank is an algorithm that is used to find/rank web-pages based on search parameters. It use popularity based measure, using no. of links along with their weights as a measure of finding the most suitable weight/credibility for a web-page.

Building to the Page Rank Algorithm

The page-rank algorithm is based on the idea of random walk. A graph is realized with each node having same rank/weight. Then, a random walk is started by selecting a random node. This is done recursively until the probability vector converges to a unique vector, which would be our ideal *Rank Vector.*

Adjacency Matrix The modified* adjacency matrix is a matrix where a_{ij} is given by $[\ell(p_i, p_j)]$ which is the weight of link from $Node(p_i)$ to $Node(p_j)$. Now, when starting the random walk, all Nodes are given equal probability of getting assigned and all links of a particular Node have the same weight adding up to one.

Therefore, adjacency matrix is given by

$$
\mathbf{M} = \begin{bmatrix}
\ell(p_1, p_1) & \ell(p_1, p_2) & \cdots & \ell(p_1, p_N) \\
\ell(p_2, p_1) & \ddots & & \vdots \\
\vdots & & \ell(p_i, p_j) & \\
\ell(p_N, p_1) & \cdots & & \ell(p_N, p_N)
\end{bmatrix}
$$

And, initial rank vector,

$$
\mathbf{R_0} = \begin{bmatrix}
1/N \\
1/N \\
\vdots \\
1/N
\end{bmatrix}
$$

Mathematics Involved

Now, the mathematical relation between $\mathbf{R}$ and $\mathbf{M}$ is given by,

$$
\mathbf{R_i} = \mathbf{M} * \mathbf{R_{i-1}}
$$

Continuing in this fashion, we shall end up on convergence,

$$
\mathbf{R} = \mathbf{M} * \mathbf{R}
$$

Thus, $\mathbf{R}$ must be the eigenvector of $\mathbf{M}$ with eigenvalue $\mathbf{1}$. The way we have defined $\mathbf{M}$, it is ensured to have an eigenvalue $\mathbf{1}$.

Modifications

There, still, remains a very trivial flaw in our system. The Node that doesn't get any links to itself gets void in our final approximation, which is unrealistic. In our model, loops act as *Rank-Sinks* as except for the Nodes in this loop all others will become void.

To manage this we include a probability that decides whether the random walk is to continue or not, given by **p**.

Now our modified page-rank becomes,

$$\mathbf{R_i} = \mathbf{p} * \mathbf{M} * \mathbf{R_{i-1}} + (\mathbf{1} - \mathbf{p}) * \mathbf{R_0}$$

For, sensible results **p** is around $(\mathbf{0.8} - \mathbf{0.9})$.

Personalized Page-Rank

Having, almost, a sensible Page-Rank we can move to the personalized version, that uses history and cache to get results more suitable for us. It gives more preference to web-pages we usually visit.
Here, we define **E**, which is the personalized form of $\mathbf{P_0}$, determined from a user's preference.

It follows the formulation,

$$\mathbf{R_i} = \mathbf{p} * \mathbf{M} * \mathbf{R_{i-1}} + (\mathbf{1} - \mathbf{p}) * \mathbf{E}$$

Summary

Personalized Page-Rank is an indexing system that provides rank to websites and keeps on modifying it on the basis of our history and preferences. It is harder to compute, as it takes more time. This method is suitable from a users perspective as it will provide the websites best suited to them. Their choices and requirements are met quite accurately.