Markov Chain Interpretation of Google Page Rank

Jia Li

December 1, 2005

Suppose there are N Web pages in total. Let the page rank of page i, $i = 1, ..., N$ be $PR(i)$. The page ranks are determined by the following linear equations:

$$PR(i) = (1 - d) + \sum_{j:\ i \text{ linked to } j} PR(j) \frac{d}{C(j)} \ , \ i = 1, ..., N$$

where $C(j)$ is the total number of links contained in page j.

Model the network of Web pages by a Markov chain. Regard the network as a huge finite state machine, where every state is a page. The conclusion we will arrive at is that the page ranks are proportional to the stationary probabilities of the states in the Markov chain. That is, if you wander around the Web pages randomly according to this Markov chain, after a long time, the probability of visiting any page at any time converges, and this probability is not affected by how you start your navigation. The higher the probability is, the higher the rank of the page will be. The scaling factor between the page rank and the probability is N/d, where $0 < d < 1$ is a chosen constant related to how likely you will restart your navigation by not following links in pages.

Set up the Markov chain as follows.

1. Every page is a state, $i = 1, ..., N$.
2. Add an imaginary state, referred to as the *Restart* page, and label it as state 0.
3. The transition probabilities between the states are defined as follows. Note that the transition probability $p_{j,i}$ is the probability of entering state i given the current state is j. Valid transition probabilities have to satisfy $\sum_i p_{j,i} = 1$ for any j.
   (a) Every state i, $0 \leq i \leq N$ has probability of $1-d$ transiting to the restart state 0, that is, $p_{i,0} = 1 - d$ for all the states i. Heuristically, this means that no matter which page you are currently in, you always have a fixed probability of restarting instead of hopping around via links.
   (b) The probability of going from state 0 to state i, $i \neq 0$, is $p_{0,i} = d/N$. That is, from the restart state, the probability of going to any real page is equal, and the total probability of going to a real page is d (the rest probability, $1 - d$, is assigned to restart again).
   (c) The probability of going from state j, $j \neq 0$, to state i, $i \neq 0$, is

   $$p_{j,i} = \frac{d}{C(j)} \cdot I(j \text{ links to } i) \ ,$$

   where $I(\cdot)$ is the indicator function that equals 1 if the argument is true, 0 otherwise. This means that for every page, besides the probability of $1-d$ going to restart, the rest probability d is evenly divided among the $C(j)$ links contained in it. If a page i is not linked to j, $p_{j,i} = 0$.

Let the stationary probabilities (i.e., limiting probabilities) of state i be π_i. By a theorem on Markov chain (assuming the MC is irreducible and ergodic, easily satisfied by a connected finite graph without cycling patterns), these probabilities satisfy the following set of linear equations:

$$\begin{aligned} \pi_i &= \sum_{j=0}^{N} \pi_j p_{j,i} \quad i = 0, 1, ..., N \\ \sum_{i=0}^{N} \pi_i &= 1 \end{aligned}$$

Specific to the Markov chain set up above:

$$\begin{aligned} \pi_0 &= \sum_{j=0}^{N} \pi_j p_{j,0} = \sum_{j=0}^{N} \pi_j (1-d) = (1-d) \sum_{j=0}^{N} \pi_j = 1 - d \\ \pi_i &= \pi_0 p_{0,i} + \sum_{j=1}^{N} \pi_j p_{j,i} \\ &= (1-d) \cdot \frac{d}{N} + \sum_{j:\ i \text{ linked to } j} \pi_j \frac{d}{C(j)} \end{aligned}$$

Multiple the last equation by $\frac{N}{d}$:

$$\frac{N}{d} \cdot \pi_i = (1-d) + \sum_{j:\ i \text{ linked to } j} \left(\frac{N}{d} \cdot \pi_j \right) \frac{d}{C(j)}$$

Define the page rank as $PR(i) = \frac{N}{d}\pi_i$, we get:

$$PR(i) = (1-d) + \sum_{j:\ i \text{ linked to } j} PR(j) \frac{d}{C(j)} \ , \quad i = 1, 2, ..., N$$

which is precisely the page rank equation of the early Google.