Google PageRank as Mean Playing Time for Pinball on the Reverse Web

Desmond J. Higham[*]

Abstract

It is known that the output from Google's PageRank algorithm may be interpreted as (a) the limiting value of a linear recurrence relation that is motivated by interpreting links as votes of confidence, and (b) the invariant measure of a teleporting random walk that follows links except for occasional uniform jumps. Here, we show that, for a sufficiently frequent jump rate, the PageRank score may also be interpreted as a mean finishing time for a reverse random walk. At a general step this new process either (i) remains at the current page, (ii) moves to a page that points to the current page, or (iii) terminates. The process is analogous to a game of pinball where a ball bounces between pages before eventually dropping down the exit chute. This new interpretation of PageRank gives another view of the principle that highly ranked pages will be those that are linked into by highly ranked pages that have relatively few outgoing links.

1 The PageRank System

Google's PageRank algorithm [1] assigns an importance ranking to each known web page. This information may then be used by a search engine

[*]Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, UK. Supported by a Research Fellowship from The Royal Society of Edinburgh/Scottish Executive Education and Lifelong Learning Department.

that seeks to find relevant and important matches for a user's query. Suppose there are N web pages and $W \in \mathbb{R}^{N \times N}$ denotes the adjacency matrix for the corresponding directed graph, so that $w_{ij} = 1$ if there is a link from page i to page j and $w_{ij} = 0$ otherwise. Then the ranking r_i is assigned to page i, where $\mathbf{r} \in \mathbb{R}^N$ satisfies

$$\left(I - dW^T D^{-1}\right) \mathbf{r} = (1-d)\mathbf{e}. \qquad (1)$$

Here

- $D = \operatorname{diag}(\deg_i)$, with $\deg_i := \sum_{j=1}^N w_{ij}$, is the diagonal out-degree matrix,
- d is a parameter in $(0, 1)$,
- $I \in \mathbb{R}^{N \times N}$ is the identity matrix,
- $\mathbf{e} \in \mathbb{R}^N$ has $e_i = 1$ for $1 \leq i \leq N$.

The vector $\mathbf{r}$ may be motivated from a recursive definition of importance that regards a link from page i to page j as a vote of confidence for page j from page i [2, 3, 1]. With this viewpoint, d is a damping factor in the resulting iteration, and $\mathbf{r}$ in (1) is the limiting set of rankings. It is also possible to interpret the normalized vector $\mathbf{r}/\sum_{i=1}^N r_i$ as the invariant measure for a Markov chain by introducing the concept of a teleporting random walk [2, 3, 1]. Given that we are currently at page i, the next step of this random walk proceeds as follows:

> with probability $1 - d$ jump to a page chosen uniformly at random over the whole web,
>
> with probability $dw_{ij}/\deg_i$ jump to page j.

We note that for the system (1) to be properly defined we require $\deg_i \neq 0$ for $1 \leq i \leq N$. Analogously, for the teleporting random walk to be properly defined we require that every page has at least one outgoing link. In practice, the "dangling page" issue ($\deg_i = 0$) must be dealt with, but in this work we assume for simplicity that D^{-1} exists. We also assume that $w_{ii} = 0$; that is, self-links are ignored.

Both the “votes of confidence” and “teleporting random walk” interpretations help to characterize the information captured by $\mathbf{r}$. In the next section we show that there is an alternative formulation that gives further insight into the PageRank computation. This alternative is also based on a random walk process, but it differs in that (a) the random walk follows links in the *reverse* direction, and (b) the ranking becomes a vector of *mean hitting times*, rather than an invariant probability distribution.

2 Mean Hitting Time Formulation

The following lemma defines a Markov chain based on the link matrix W. We use $\|\cdot\|_\infty$ to denote the L_∞ norm.

Lemma 2.1. *Introduce a state space that consists of the N web pages, ordered from 1 to N, plus an extra page, labeled $N+1$, that we refer to as the* exit page*. Define the matrix $P \in \mathbb{R}^{(N+1)\times(N+1)}$ by*

$$\begin{aligned} p_{ij} &= \alpha \frac{w_{ji}}{\deg_j} + \beta \delta_{ij}, \qquad \text{for } 1 \le i,j \le N, && (2)\\ p_{i,N+1} &= 1 - \sum_{j=1}^{N} p_{ij}, \qquad \text{for } 1 \le i \le N, && (3)\\ p_{N+1,j} &= \delta_{N+1,j}, \qquad \text{for } 1 \le j \le N+1, && (4) \end{aligned}$$

where $\alpha > 0$ is a fixed parameter, $\beta := (d-\alpha)/d$ and δ_{ij} is the Kronecker delta. If

$$d \le \frac{1}{\|W^T D^{-1}\|_\infty}, \qquad (5)$$

then for all choices $0 < \alpha < d$ the matrix P is non-negative and stochastic; that is, $p_{ij} \ge 0$ for all $1 \le i,j \le N+1$ and $\sum_{j=1}^{N+1} p_{ij} = 1$ for all $1 \le i \le N+1$. Hence, P is a valid transition matrix for a Markov chain.

Proof. By construction, the required properties hold if $\sum_{j=1}^{N} p_{ij} \le 1$ for $1 \le i \le N$. In (2), using the definition of β, this gives

$$\alpha \left(\|W^T D^{-1}\|_\infty - \frac{1}{d} \right) \le 0,$$

which follows under (5). □

We remark that the final row of P is arbitrary for our purposes—the definition in (4) is made for concreteness.

The following theorem connects the Markov chain in Lemma 2.1 with PageRank.

Theorem 2.1. *Under condition (5), consider the Markov chain defined by (2)–(4) with any fixed $0 < \alpha < d$; that is, let*

$$\mathbb{P}\left(X_{n+1} = j,\ \text{given } X_n = i\right) = p_{ij}.$$

For $1 \leq i \leq N$, let z_i denote the mean hitting time for state $N+1$, starting from state i; that is,

$$z_i := \mathbb{E}(h^i), \qquad \text{where } h^i := \min\left\{n > 0 : X_n = N+1,\ \text{given } X_0 = i\right\}.$$

Then the vector $\mathbf{z} \in \mathbb{R}^N$ is linearly proportional to the PageRank vector $\mathbf{r}$ in (1), so that,

$$\frac{\mathbf{z}}{\sum_{i=1}^N z_i} = \frac{\mathbf{r}}{\sum_{i=1}^N r_i}.$$

Proof. It follows from standard Markov chain theory, see for example [4, Theorem 1.3.5] or [5, Exercise 7, Section 6.3], that $\mathbf{z}$ is the minimal non-negative solution to the linear system

$$\left(I - \widehat{P}\right)\mathbf{z} = \mathbf{e},$$

where $\widehat{P} \in \mathbb{R}^{N \times N}$ has $\widehat{p}_{ij} = p_{ij}$. From (2), $\widehat{P} = \alpha W^T D^{-1} + \beta I$, so the system may be rearranged to

$$\left((1-\beta)I - \alpha W^T D^{-1}\right)\mathbf{z} = \mathbf{e}.$$

Substituting $\alpha = (1-\beta)d$, this system may be written

$$\left(I - dW^T D^{-1}\right)\mathbf{z} = \frac{1}{1-\beta}\mathbf{e}. \tag{6}$$

Note that $\|dW^T D^{-1}\|_1 < 1$, so $I - dW^T D^{-1}$ is non-singular, see, for example, [6, Lemma 2.3.3]. So $\mathbf{z}$ is in fact the unique solution to (6). Comparing (6) with (1) we see that $\mathbf{z}$ is a multiple of $\mathbf{r}$, and the result follows. □

To interpret this result, we first note that the probabilities in (2) and (3) show that the move from page i may be summarized as follows:

with probability β remain at page i,
with probability $1 - \beta - \alpha \sum_{j=1}^{N} w_{ji}/\deg_j$ jump to exit page,
with probability $\alpha w_{ji}/\deg_j$ jump to page j.

The jump to page j in the third case is possible only if $w_{ji} \neq 0$; that is, only if page j points to page i. In this case the factor $\deg_j$ in the denominator indicates that such a jump is more likely to take place when j has relatively few outgoing links. Also, the more nodes there are pointing to page i, the larger $\sum_{j=1}^{N} w_{ji}/\deg_j$ and hence the smaller the probability $1-\beta-\alpha \sum_{j=1}^{N} w_{ji}/\deg_j$ of jumping to the exit page.

The process is generally following reverse links, but also incorporates pausing at the current page and jumping to the exit page. The mean hitting time vector $\mathbf{z}$ measures the average number of steps that we make, starting from page i, before we hit the exit page. Theorem 2.1 shows that $\mathbf{z}$ and $\mathbf{r}$ are equal, when normalized, and hence a highly PageRanked page is precisely a starting point that gives high average life expectancy under this process.

A loose analogy for this new Markov chain is a game of pinball on the web with the ball bouncing between pages, following reverse links (moving from page i to page j only if page j links to page i) before finally succumbing to the "game over" page. The ranking r_i is the average length of a game that starts at page i.

This new interpretation of PageRank gives an alternative perspective on the basic premise that highly ranked pages are well-connected. To get a high mean hitting time, that is, to have a high average longevity, page i must be linked to by "long lived" pages that do not give out their links too frivolously.

References

[1] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report,

Stanford Digital Library Technologies Project, 1998.

[2] Desmond J. Higham and Alan Taylor. The sleekest link algorithm. *The Institute of Mathematics and Its Applications (IMA) Mathematics Today*, 39:192—197, 2003.

[3] Amy N. Langville and Carl D. Meyer. Deeper inside Pagerank. *Internet Mathematics*, to appear.

[4] J. R. Norris. *Markov Chains.* Cambridge University Press, 1997.

[5] Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes.* Oxford University Press, third edition, 2001.

[6] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, third edition, 1996.