PageRank and Search Engines

Srisha Murthy Nippani

July 4, 2025

1 Theory

PageRank is an algorithm used by the Google Search Engine to rank web pages based on their ”importance” in response to a user’s search query. It is natural to represent the web as a directed graph. Each node is a webpage, and an edge from node i to node j indicates that webpage i has a forward link (i.e. hyperlink) to webpage j. It is intuitive that page i is important if it has many backlinks (i.e. there are many pages that have a forward link to i), or perhaps, page i has ”important” backlinks. This intuition lends itself to the following recursive definition:

Definition 1.1 (Simple Ranking)**.** Let i be a page. Then, let F_i be the set of pages that i points to, and B_i be the set of pages that point to i. Then, the rank of i is given by:

$$r(i) = \sum_{j \in B_i} \frac{r(j)}{|F_j|}.$$

We say that a rank vector r is an assignment between page i and its rank $r(i)$.

However, there is an issue with the ”Simple Ranking”. Suppose there exist two pages that point to each other but no other page. Then, this loop will accumulate rank. In reality, should a random web surfer get into a small loop of web pages, the surfer might get bored and jump to another page. Therefore, we introduce a damping factor α, satisfying $0 < \alpha < 1$, which denotes the probability that the surfer clicks on a hyperlink on the page. Then, $1 - \alpha$ is the probability that the surfer ”teleports” to another page. In this scenario, consider a probability distribution v where $v(i)$ denotes the probability that the surfer ”teleports” to page i. We may interpret the PageRank model as a random walker clicking or ”teleporting” between web pages. Now, we shall replace the ”Simple Ranking” with PageRank:

Definition 1.2 (PageRank)**.** Following the notation above, the PageRank of i is given by:

$$r(i) = (1 - \alpha)v(i) + \alpha \left(\sum_{j \in B_i} \frac{r(j)}{|F_j|} \right).$$

We may use the theory of stochastic processes (as discussed in lecture) to analyze PageRank. Consider the adjacency matrix W, where $W_{i,j} = 1$ if there is a forward link from page i to page j and 0 otherwise. Let D be the diagonal matrix define by $D_{i,i} = \sum_j W_{i,j}$. Let P be the transition matrix where $P_{i,j}$ is the probability transitioning from page j to page i. Then, for any two pages i, j, the transpose $(P^*)_{i,j} = \dfrac{W_{i,j}}{\sum_j W_{i,j}} = (D^{-1}W)_{i,j}$ which means that $P = W^* D^{-1}$. A key observation is that for every page i,

$$(Pr)_i = \sum_j P_{i,j} \cdot r(j) = \sum_{j \in B_i} \frac{1}{|F_j|} \cdot r(j) + \sum_{j \notin B_i} 0 \cdot r(j) = \sum_{j \in B_i} \frac{r(j)}{|F_j|},$$

which means that $r(i) = (1-\alpha)v(i) + \alpha(Pr)_i$. Therefore, we may restate the PageRank equation:

Definition 1.3 (PageRank equation restated)**.** Following the notation above, the PageRank vector is given by:

$$r = (1-\alpha)v + \alpha Pr.$$

We shall discuss a few notable results concerning PageRank that can be proven using tools from linear algebra and real analysis.

Lemma 1.4. $\|Pr\|_{\ell_1} \leq \|r\|_{\ell_1}$.

Proof. We see that

$$\begin{aligned}
\|Pr\|_{\ell_1} &= \sum_i |(Pr)_i| \\
&= \sum_i \left| \sum_j P_{i,j} \cdot r(j) \right| \\
&\leq \sum_i \sum_j |P_{i,j}||r(j)| && \text{(by triangle inequality)} \\
&= \sum_j \sum_i P_{i,j} |r(j)| \\
&= \sum_j |r(j)| \left(\sum_i P_{i,j} \right) \\
&= \sum_j |r(j)| \cdot 1 && \text{(P is column stochastic)} \\
&= \|r\|_{\ell_1}
\end{aligned}$$

as desired. □

Theorem 1.5. There exists a unique solution to the PageRank equation as stated in 1.3.

Proof. Let X^n be a set of probability distributions on $\mathbb{R}^n$, where n is of course the number of pages. By noting that $X^n = \bigcap_{i=1}^{n} \{x \in \mathbb{R}^n : x_i \geq 0\} \cap \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1 \right\}$, we see that X^n is the finite intersection of closed sets and is therefore closed. Since $\mathbb{R}^n$ is complete (with respect to the ℓ_1 norm), and X^n is a closed subset of $\mathbb{R}^n$, X^n is complete. Now, define a map

$$T_\alpha : X^n \to X^n : x \mapsto (1-\alpha)v + \alpha Px.$$

This is well-defined (i.e. $T_\alpha(X^n) \subset X^n$) because $v \in X^n$ and X^n is closed under convex combinations. Additionally, T_α is a strict contraction. Observe that for any $x, y \in X^n$,

$$\begin{aligned}
\|T_\alpha(x) - T_\alpha(y)\|_{\ell_1} &= \|(1-\alpha)v + \alpha P x - (1-\alpha)v - \alpha P y\|_{\ell_1} \\
&= \|\alpha P x - \alpha P y\|_{\ell_1} \\
&= \alpha \|P(x-y)\|_{\ell_1} \\
&\leq \alpha \|x - y\|_{\ell_1}, && \text{(by Lemma 1.4)}
\end{aligned}$$

and since $0 < \alpha < 1$, T_α is a strict contraction. By the Banach fixed-point theorem, T_α has a unique fixed point r. In particular, $r = T_\alpha(r) = (1-\alpha)v + \alpha P r$ is the unique solution to the PageRank equation. □

Fortunately, the standard proof for the Banach fixed point theorem is constructive. Therefore, we have a method of obtaining the solution r to the PageRank equation: choose any point $r_0 \in X^n$, and the sequence of functional iterates $(T_\alpha^n(r_0))$ always converges to the fixed point r.

Since r is a probability distribution, $\mathbb{1}^* r = 1$. Here, $\mathbb{1}^*$ is the row vector with all entries 1. Therefore, we may write

$$T_\alpha = (1-\alpha)v\mathbb{1}^* + \alpha P$$

and rewrite the PageRank equation (1.3) as:

$$r = T_\alpha(r).$$

Then, we say that r is a right eigenvector corresponding to the eigenvalue 1. Observe that $v\mathbb{1}^*$ is an n-by-n matrix where each column is v. Therefore, each column of T_α is a convex combination of v and a column of P, which means that T_α is column stochastic. We know from Perron-Frobenius theory that 1 is the maximal eigenvalue of T_α and r is the top eigenvector. Interestingly, in a paper titled ”A Note on the Eigenvalues of the Google Matrix” by Lars Eldén, it is proven that if $\{1, \lambda_2, \ldots, \lambda_k\}$ are the eigenvalues of P, then $\{1, \alpha\lambda_2, \ldots, \alpha\lambda_k\}$ are the eigenvalues of T_α.

The most basic implementation of PageRank is given as such:

Algorithm 1 PageRank (basic)

```
Set $\alpha$ such that $0 < \alpha < 1$                    ▷ Typically, $\alpha = 0.85$
Set $\varepsilon$ (the threshold), $r_0$ and $v$             ▷ Typically, $v(i) = \frac{1}{n}$ for all $i$
Construct $T_\alpha$
Set $r \leftarrow T_\alpha(r_0)$ and $\delta \leftarrow \|r - r_0\|_{\ell_1}$.
while $\delta > \varepsilon$ do
    Set $r_0 \leftarrow r$ and $r \leftarrow T_\alpha(r_0)$
    Set $\delta \leftarrow \|r - r_0\|_{\ell_1}$
end while
return $r$
```

Note that matrix-vector multiplication is typically $O(n^2)$. However, for sparse matrices, it is $O(M)$, where M is the number of non-zero entries. PageRank typically involves sparse matrices; each web page links to only a small fraction of the millions of available pages. Since P is sparse, the computation of Px is $O(M)$ and therefore the computation of $T(x) := (1-\alpha)v + \alpha P x$ is $O(M+n)$ since scaling a vector is $O(n)$. It follows that each iteration of the while loop is $O(M+n)$, where M

is the number of non-zero entries of P. If $M \gg n$, then this is just $O(M)$. Hence, the complexity of PageRank would be $O((M+n)\cdot k)$ where k is the number of iterations of the while loop. One might estimate a lower bound for k. By the definition of convergence, for any $\varepsilon > 0$, there exists k such that for all $j \geq k$, $\|T_\alpha^j(r_0) - r_0\|_{\ell_1} < \varepsilon$. Such a k may be estimated from the proof of Banach's fixed point theorem. Since T is a strict contraction, we see that

$$\|T_\alpha^{j+1}(r_0) - T_\alpha^j(r_0)\|_{\ell_1} \leq \alpha \|T_\alpha^j(r_0) - T_\alpha^{j-1}(r_0)\|_{\ell_1} \leq \ldots \leq \alpha^j \|T_\alpha(r_0) - r_0\|_{\ell_1} = C\alpha^j$$

for $C = \|T_\alpha(r_0) - r_0\|_{\ell_1}$. Then, for $l \geq j$, applying the triangle inequaliy yields

$$\|T_\alpha^{l+1}(r_0) - T_\alpha^j(r_0)\|_{\ell_1} \leq \sum_{i=j}^{l} \|T_\alpha^{i+1}(r_0) - T_\alpha^i(r_0)\|_{\ell_1} \leq \sum_{i=j}^{l} C\alpha^i = C\sum_{i=j}^{l} \alpha^i.$$

Letting $l \to \infty$ gives us an estimate of the error

$$\|T_\alpha^j(r_0) - r\|_{\ell_1} \leq C\sum_{i=j}^{\infty} \alpha^i = C \cdot \frac{\alpha^j}{1-\alpha}.$$

If $C \cdot \dfrac{\alpha^j}{1-\alpha} < \varepsilon$, then we need $\alpha^j < \dfrac{\varepsilon(1-\alpha)}{C}$. Taking log on both sides, we have $j \log \alpha < \log \dfrac{\varepsilon(1-\alpha)}{C}$. Since $0 < \alpha < 1$, assuming $\varepsilon < 1$, both $\log \alpha < 0$ and $\log \dfrac{\varepsilon(1-\alpha)}{C} < 0$. Therefore, $j > \dfrac{\log \frac{\varepsilon(1-\alpha)}{\|T_\alpha(r_0) - r_0\|_{\ell_1}}}{\log \alpha}$ and this provides a lower bound for the number of iterations k.