PageRank

Tan Yan Daniel Bae

December 12, 2018

1 Introduction

The early search engines of the Internet used text-based ranking systems that assigned relevance to webpages based on the highest number of keywords present. Although this approach intuitively makes sense, it often ranked webpages with a high volume of keywords and no other content as relevant results to a search, which is not desired behavior. A better approach to generating webpage rankings is to examine the links to that page. Under this schema, a page is highly ranked if it has many links or a highly-ranked page directed to it, encapsulating both popularity and authority into a ranking.

2 Definitions

Definition 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a countable probability space, where

1. The outcome space Ω is some countable set of webpages on the internet;
2. The set of all possible events $\mathcal{F} = \{\{\omega\} \mid \omega \in \Omega\}$ is set of all events consisting of exactly one outcome, i.e. the event that a web surfer is on a certain webpage;
3. The probability measure $\mathbb{P}$ is a function $\mathcal{F} \to [0, 1]$ denoting the probability of the event that a web surfer is on a certain webpage.

Definition 2. Let a random walk of some countable set of webpages Ω be a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$ defined as follows:

Let Ω be enumerated by the set $S = \{0, 1, 2, 3, \ldots\} \subseteq \mathbb{N}$, such that $\Omega = \{\omega_i \mid i \in S\}$.

Let the index set $T = \{0, 1, 2, 3, \ldots\} = \mathbb{N}$ denote the number of links that the web surfer has clicked on, i.e. the number of times that the web surfer has jumped from one webpage to another.

Let $\{X_t : \Omega \to S\}_{t \in T}$ be a family of S-valued random variables indexed by T such that $\mathbb{P}\{X_t = i\}$ corresponds to the probability of the event that the web surfer lands on the i-th webpage after clicking on t links or making t jumps from one webpage to another.

Then $\{X_t : \Omega \to S\}_{t \in T}$ precisely describes a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$. Furthermore, we assume that a random walk of Ω satisfies the Markov property.

Definition 3. (The Markov property) Let a stochastic process on some countable probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with countable state space S be described by $\{X_t : \Omega \to S\}_{t \in T}$, where the index set T is exactly $\{0, 1, 2, 3, \ldots\} = \mathbb{N}$.

Then this stochastic process is said to be memoryless, a.k.a. possesses the Markov property, a.k.a. is a Markov process, if and only if for all $t \in T$ and $x_0, x_1, \ldots, x_t \in S$:

$$\mathbb{P}\{X_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_0 = x_0\} = \mathbb{P}\{X_t = x_t \mid X_{t-1} = x_{t-1}\}$$

which implies

$$\mathbb{P}\{X_t = x_t\} = \sum_{x_{t-1} \in S} \mathbb{P}\{X_t = x_t \mid X_{t-1} = x_{t-1}\} \mathbb{P}\{X_{t-1} = x_{t-1}\}$$

For a random walk of some countable set of webpages Ω, this property implies that the probability that the web surfer lands on a certain webpage after clicking on t links is solely dependent on which webpage the web surfer was on before clicking the t-th link.

Definition 4. Let a random walk of some finite set of n webpages Ω be a Markov process on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ be described by $\{X_t : \Omega \to S\}_{t \in T}$, where $S = \{1, 2, \ldots, n\} = [n]$ and $T = \{0, 1, 2, 3, \ldots\} = \mathbb{N}$.

Then, for each $t \in T$, let $\mathbf{v}_t$ be a $n \times 1$ vector where $v_i = \mathbb{P}\{X_t = i\}$ for $i \in S$. Observe that $\mathbf{v}_t$ corresponds to the probability distribution of which webpage the web surfer is on after clicking t links. We say that $\mathbf{v}_t$ is a probablistic vector describing the state at step t.

Furthermore, let P be a $n \times n$ matrix where $p_{ij} = \mathbb{P}\{X_t = i \mid X_{t-1} = j\}$ for $i, j \in S$ and any arbitrary $t \in T$. Observe that by Definition 3, $\mathbf{v}_t = P\mathbf{v}_{t-1}$ for all $t \in T$. So multiplying by P corresponds to the web surfer clicking a link. Thus we say that P is the Markov transition matrix.

Definition 5. Let A be an adjacency matrix that represents a directed graph consisted of some finite set of n webpages enumerated 1 to n, with the k-th webpage represented as node k, and links between two webpages as edges, such that $A_{ij} = 1$ if and only if there exists an edge from node j to node i, i.e. the j-th webpage has a link to the i-th webpage, and $A_{ij} = 0$ otherwise.

Definition 6. Let $L(j)$ denote the number of outgoing links from the j-th webpage. Given a $n \times n$ adjacency matrix A representing a directed graph of n webpages, we have $L(j) = \sum_{i=1}^{n} A_{ij}$. If there exists a node k such that $L(k) = 0$, i.e. the k-th webpage has no outgoing links to other webpages, we say that node k is a dangling node.

Definition 7. Let the Markov transition matrix P of a random walk of some finite set of n webpages be constructed as follows: given a $n \times n$ adjacency matrix A representing a directed graph of the

n webpages, for each entry p_{ij} in P, $i, j \in [n]$,

$$p_{ij} = \begin{cases} \dfrac{1}{L(j)} & \text{if } A_{ij} = 1 \\ \dfrac{1}{n} & \text{if } A_{ij} = 0 \text{ and } L(j) = 0 \\ 0 & \text{otherwise} \end{cases}$$

Thus

Definition 8. A matrix is said to be column-stochastic if and only if each of its entries are between 0 and 1 inclusive, and each of its columns sum up to 1. Similarly, a vector is said to be a stochastic or probablistic vector if and only if its components are between 0 and 1 inclusive and sum up to 1. Note that by definition, the Markov transition matrix P is column-stochastic, and each $\mathbf{v}_t$ is a probablistic vector.

Definition 9. An Markov transition matrix is said to be irreducible if ... A positive Markov transition matrix with all positive entries is necessarily irreducible Note that the Markov transition matrix P as constructed above is not necessarily irreducible.

Definition 10. Let M denote the "PageRank Matrix" or "Google Matrix" defined by Page and Brin. Define

where $\mathbf{1}_{m \times n}$ denotes ... and d denotes the "damping factor", representing the probability of a web surfer randomly traveling from one page to any other page. $0 \leq d < 1$.

Definition 11. Let λ denote an eigenvalue of the square matrix A and let $\mathbf{v}$ denote its corresponding eigenvector, such that $A\mathbf{v} = \lambda\mathbf{v}$. A probablistic eigenvector is...

3 Main Ideas

Let x be the PageRank vector, where the i^{th} entry of x is the ranking of the i^{th} webpage. We initialize x as a vector with all entries equal to $\frac{1}{n}$ and then iteratively perform t random walks through n webpages. Since P is the Markov transition matrix, the product of P^t and x represents the ranks of each page after t traversals. We note that after infinitely many random walks, the probability that the web surfer is on a given page is proportional to the number of in-bound links to that page. Therefore, we want the PageRank vector to be the solution to $\lim\limits_{t \to \infty} A^t x$

However, we note that there are two special cases that can lead to x not converging properly. Consider the case when a random web surfer reaches a webpage by directly typing in its hyperlink. This would be like "jumping" from a node i in G to any other node j in G, regardless of if the edge (i, j) exists in G. If we assume that each webpage is equally likely to be "jumped" to by the random web surfer, we obtain a new $n \times n$ probabilstic matrix, which we will denote as B, where all of the elements are $\frac{1}{n}$ to represent the probability of jumping from a webpage to another webpage.

Since we know d is the probability that the surfer will "jump", we can now define the PageRank matrix M to be $(1-d)P + dB$.

In order to prove the existance of the PageRank vector, we first rely on the Perron-Frobenius Theorem.

Theorem 1. (Perron-Frobenius Theorem) Consider a $n \times n$ positive column-stochastic matrix M. Let $\lambda_1, \lambda_2, \cdots, \lambda_n$ be the eigenvalues of M sorted in decreasing order of magnitude. Then 1 is an eigenvalue with multiplicity 1 such that $\lambda_1 = 1$ and there exists a corresponding probabilistic eigenvector.

Since the PageRank matrix M is constructed to be a postitive column-stochastic matrix, by Perron-Frobenius theorem we know that there exists a probabilistic eigenvector $\mathbf{v}^*$ corresponding to the eigenvalue 1.

The next theorem uses the fact that all of the eigenvectors λ_i for $1 < i \leq n$ of M are less than 1 to show that the PageRank vector is equal to $\mathbf{v}^*$.

Theorem 2. (Power Method Convergence Theorem) Let M be a $n \times n$ positive column-stochastic matrix, with $\mathbf{v}^*$ denosting the probabilstic vector corresponding to 1. Let z be the column vector with all entries equal to $\frac{1}{n}$. Then the sequence $z, Mz, \cdots, M^k z$ as $k \to \infty$ converges to $\mathbf{v}^*$.

```
def build_prob_matrix(adj_list):
# number of nodes
n = len(adj_list)
P = np.zeros((n,n), dtype=float)
for (j, connected_nodes) in enumerate(adj_list):
    if connected_nodes: # non-empty
        P[connected_nodes, j] = 1.0 / len(connected_nodes)
    else: # dangling node
        P[:, j] = 1.0 / n
return P
```

4 Conclusion

References

[1]