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. A stochastic matrix, probability matrix, or Markov transition matrix is ... A column-stochastic or left-stochastic matrix is ...

Definition 2. A positive matrix is a matrix with all positive entries.

Definition 3. An Markov transition matrix is said to be irreducible if ... A positive stochastic matrix is necessarily irreducible.

Definition 4. Let G be a directed graph of the webpages, with each webpage as a node and links between the webpages as edges ... such that $G_{ij} = 1$ if ... also let n denote the number of nodes ... let $L(i)$ denote the number of outgoing edges ...

Definition 5. A dangling node is a node that has no outgoing edges

Definition 6. Let P be the probablistic matrix

Definition 7. Let d denote the "damping factor" ... representing the probability of a web surfer randomly traveling from one page to any other page. $0 \leq d < 1$.

Definition 8. Let $\mathbf{1}$

Definition 9. Let M denote the "PageRank Matrix" or "Google Matrix" defined by Page and Brin. Define

Definition 10. 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

We can design a Markov chain for the network of webpages and links, which is represented by G.

Given G, we can construct P such that P_{ij} is the probability that a surfer on page i clicks on a link to page j. If we assume that each page has an equal probability of being selected, the column vector P_i contains $L(i)$ entries with value $\frac{1}{L(i)}$ when the edge from i to j exists in G and 0's in the other entries.

In order to make P more realistically account for the probability of reaching a webpage through links, we have to handle the special cases of dangling nodes and disconnected components of G.

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 determine the ranks of the webpages, we let x be our initial PageRank vector with all entries equal to $\frac{1}{n}$.

We want to find the PageRank vector, which is the stationary distribution of the Markov chain modeled by G.

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 $\mathbf{v}^*$.

This implies that after infinitely many random walks, the web surfer is most likely to be on the webpage corresponding to the most links.

```
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]