
PageRank Algorithm

David Schneck and Isaac Tuckey
Mathematical Foundations of ML and AI
University of Central Florida

1 Abstract

As the World Wide Web grew, there also grew a necessity for search engines that could index and allow a user to search for information on the web relevant to their queries. Larry Page and Sergey Brin developed a way to rank the relative *importance* of the many existing web pages and called it PageRank. [1] It utilized a “web crawler” to retrieve and process information about the hyperlinks that connect web pages. It then used that information to rank pages by importance since a page that has many other popular pages directing users to it is also likely a reliable or interesting page.

The core math concepts behind the PageRank algorithm are Markov chains, eigenvectors, and probability distributions. A Markov chain is defined by an initial distribution and a transition matrix. We will rely on the Frobenius-Perron theorem: proving that a matrix with real square entries has a unique largest real eigenvalue and that the corresponding eigenvector can be chosen to have strictly positive components. The power method iteration algorithm will be used to compute the eigenvalue with the largest absolute value along with its corresponding eigenvector.

To test our understanding of the PageRank algorithm we will implement a version of it in Python. It will be tested using local directories of sample HTML pages that link to each other collected from various sources. We will adapt or re-implement a crawler that can parse the pages creating a Python dictionary of the pages and their corresponding links. This will then be used to perform the PageRank algorithm and test on increasingly large data sets to determine how well our implementation performs. Time permitting, we will find alternatives to the PageRank algorithm and measure their performance to explore exactly why PageRank gave Google the competitive edge others lacked.

References

- [1] Sergey Brin and Lawrence Page. “The anatomy of a large-scale hypertextual web search engine”. In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117.