Google's PageRank Algorithm For Web-Indexing

Sanchit Jalan,Kripi Singla,R. Shaanal

June 11, 2023

Team Members

- Sanchit Jalan (2022101070)
- Kripi Singla (2022102063)
- R. Shaanal (2022102071)

Introduction

Revolutionizing how the modern world operates, the Internet is a powerful medium in which anyone around the world, regardless of location, can access endless information about any subject and communicate with one another without bounds. All that is needed is a computer and the World Wide Web. One of the greatest results of the Internet was the establishment of hyperlinks. The World Wide Web is an extensive computer network consisting of billions of web pages holding documents of information. Hyperlinks are the pathways from one web page to another, initiating the capability of communication between these pages. Interactions between documents are performed by referencing one another via links. Here lies the foundation on how the most dominant search engine, Google, does its magic using its PageRank algorithm.

Google PageRank is an algorithm developed by Larry Page and Sergey Brin, the founders of Google, to measure the importance and relevance of web pages. It was one of the key factors that contributed to Google's early success as a search engine.

The basic idea is that authoritative pages get more links. So pages with more links should rank higher in search results.

Website indexation is the process by which a search engine adds web content to its index. This is done by "crawling" web pages for keywords, metadata, and related signals that tell search engines if and where to rank content.

In essence, PageRank treats a link from one page to another as a vote of confidence or endorsement and this is how it assists in web indexing.

The algorithm works by iteratively calculating the PageRank scores for all web pages in a large graph of interconnected pages. Initially, each page is assigned an equal score. In each iteration, the PageRank score of a page is updated

based on the scores of the pages that link to it. Pages with higher scores are considered more important and influential.
One key aspect of PageRank is that not all votes (links) are equal. The algorithm considers the quality and relevance of the linking page when calculating the score. A link from a highly reputable and relevant website carries more weight than a link from a less reputable or unrelated site. Additionally, the number of outgoing links from a page also affects the weight of each link.
PageRank also takes into account the concept of "damping factor" or "random surfer model." It assumes that a user randomly clicks on links while browsing the web and incorporates this randomness into the calculation.

Related Works

Google's PageRank algorithm, initially introduced as part of the Google search engine, continues to play a significant role in today's world beyond web search. The original PageRank patent from 1998 expired in 2018 and, to the surprise of many, wasn't renewed. But that didn't mean PageRank was dead.It is still used in various ways in Search Engines,Web Analytics,Recommendation Systems,Social Networks,Citation Analysis,Fraud Detection..
Though there is no longer a toolbar that gives us a webpage's PageRank score doesn't mean it's not still used.In 2017 Google's Gary Illyes confirmed on Twitter that the algorithm still uses PageRank. The Algorithm has developed a lot from start of its establishment.It was incorporated in Google's Toolbar and many new algorithms have developed from it . Penguins Algorithm is one of the many algorithm's that has developed from PageRank. PageRank was removed from the Google's Toolbar as it was very easy to manipulate.Though there is no longer a toolbar that gives us a webpage's PageRank score doesn't mean it's not still used.In 2017 Google's Gary Illyes confirmed on Twitter that the algorithm still uses PageRank.

Overview of PageRank Algorithm

1.Initialization

Each web page is initially assigned an equal PageRank value. This value can be thought of as the probability of a random surfer landing on that page.

2.Importance of incoming links

The PageRank algorithm considers a page to be more important if it receives many incoming links from other pages. The importance of a linking page is determined by its own PageRank score. The more important the linking page, the more weight its outgoing links carry.

3.Calculation

$$PR(A) = (1-d) + d\sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)} \tag{1}$$

PR(A) is the PageRank score of page A.
d is the damping factor, typically set to 0.85.
PR(Ti) is the PageRank score of page Ti, which has a link to page A.
C(Ti) is the total number of outgoing links on page Ti.

4.Iterative calculation

The calculation is performed iteratively, with the PageRank scores being updated in each iteration. The algorithm continues until the PageRank scores converge, meaning they stabilize and stop changing significantly.

Application Of Linear Algebra and Graph Theory

The PageRank algorithm utilises various linear algebraic concepts to calculate and update the PageRank scores of web pages, namely, Matrix Representation, Eigenvector calculation, Stochastic Matrix,Transition Matrix,Random surfer Model and Damping factor. The damping factor introduces a Markov chain model.

1. Modeling the Web Graph

The web graph can be represented as a directed graph, where each web page is a node, and the hyperlinks between pages are represented as edges. This graph can be represented using an adjacency matrix.

2. Use of Stochastic Matrix

To analyze the web graph, the adjacency matrix is converted into a stochastic matrix. Each column of the matrix represents a web page, and the entries represent the probability of moving from one page to another via a hyperlink. The matrix is constructed such that the sum of each column is equal to 1, ensuring that the matrix is a Markov matrix.

3. Transition matrix

The stochastic matrix is further transformed into a transition matrix by incorporating a damping factor. The damping factor accounts for the probability that a random surfer stops following hyperlinks and randomly jumps to another page. The transition matrix incorporates this behavior and is used to model the random surfer's movement through the web graph.

4. Eigenvalues and eigenvectors

The PageRank algorithm involves finding the dominant eigenvector of the transition matrix. The dominant eigenvector corresponds to the stationary distribution of the random surfer's long-term behavior. It represents the importance or ranking of each web page.

5. Final Calculation for PageRank

The power iteration method is commonly used to find the dominant eigenvector. It involves iteratively multiplying the transition matrix by an initial vector until convergence. The resulting vector is the PageRank vector, which assigns a score to each web page based on its importance.
In practice, the power iteration method may not be efficient for large-scale web graphs. Instead, specialized algorithms like the PageRank algorithm exploit the sparsity of the matrix and use iterative methods such as the Arnoldi iteration or the power method with a shift to compute the dominant eigenvector more efficiently.

Timeline

In the next part of the project we would add mathematical equations and will also visualize how the PageRank Algorithm works for a small network of webpages..

Contributions

- Sanchit contributed towards the writing document on LaTeX.He also found how the algorithm works and the application of Linear Algebra in it .
- Kripi Singla also did research on working of algorithm and application of Linear Algebra ..
- R. Shaanal contributed towards the finding the related works and formulating the algorithm..