IT256 Mini Project - PageRank Algorithm

Vinayaka S N(211AI040) and Vivek Vittal Biragoni(211AI041).

15 March 2023

Abstract: The significance of a web page is a personal matter and depends on the interests, knowledge, and attitudes of its readers. However, there are still objective ways to evaluate the importance of web pages. This study introduces PageRank, a method that objectively and mechanically rates web pages by measuring human interest and attention towards them. The study compares PageRank to a theoretical random web surfer, presents an efficient way to calculate PageRank for a large number of pages, and demonstrates how PageRank can be used for search and user navigation.

Introduction:
The World Wide Web presents several challenges for information retrieval due to its vast size and heterogeneity. There are currently over 150 million web pages, and this number doubles in less than a year. Moreover, the web pages cover a wide range of topics, from personal anecdotes to academic journals. Search engines also face inexperienced users and pages designed to manipulate search engine rankings. However, unlike traditional document collections, the Web provides additional information such as link structure and link text. This paper proposes using the link structure of the Web to create a global ranking of every web page called PageRank. By considering the importance of each page, PageRank can help search engines and users navigate the World Wide Web's vast and diverse content.

Definition of PageRank:
The PageRank algorithm is a method for objectively and mechanically rating web pages based on their importance and the attention they receive from human users. It takes advantage of the link structure of the web to produce a global ranking of every web page.

The ranking, called PageRank, is computed using a recursive equation that is iteratively applied until it converges. However, there is a problem with this simplified ranking function in the form of rank sinks, where pages that link only to each other trap the rank without distributing it. To overcome this problem, a rank source is introduced, which corresponds to a decay factor that balances the equation. The PageRank algorithm can be represented in matrix notation as R0 = c(A + E - 1)R0, where E is a vector over the web pages that corresponds to a source of rank.

An outline of the algorithm used(PageRank algo):

Gather web page data: Collect data on the web pages to be ranked, including their links to other pages.

Build the link matrix: Use the collected data to build a matrix where each row and column represents a web page, and the elements of the matrix represent the links between pages.

Calculate the PageRank vector: Use an iterative algorithm to calculate the PageRank vector, which assigns a rank to each web page based on its importance, as determined by the links between pages.

I*nitialize the PageRank vector*: Start by assigning an initial rank to each page in the matrix. This could be a uniform rank, or a rank based on some other metric.

Apply the PageRank algorithm: Use the iterative PageRank algorithm to update the rank of each page in the matrix based on the links to and from other pages.

Converge on a steady-state solution: Keep applying the PageRank algorithm until the ranks of the pages converge to a steady-state solution. This means that the ranks are no longer changing significantly with each iteration.

Normalize the PageRank vector: Normalize the PageRank vector so that the sum of all the ranks is equal to 1.

Use the PageRank vector: Once the PageRank vector has been calculated, it can be used to rank the web pages in order of importance, and to provide search results based on those rankings. There would be many variations to this proposed outline.

Relation between the project chosen and "Applied linear algebra"

The PageRank algorithm is based on the concept of Markov Chains, which are a type of mathematical model that describes a system that transitions from one state to another according to a set of probabilities. In the case of the PageRank algorithm, the states are the web pages, and the probabilities represent the likelihood of transitioning from one page to another through hyperlinks.

To compute the PageRank of a web page, we need to solve a system of linear equations, where each equation corresponds to a web page and its PageRank score. These equations are based on the hyperlink structure of the web, and they relate the PageRank scores of the pages to the scores of the pages that link to them.

To do this, we can represent the hyperlink structure of the web as a matrix, called the hyperlink matrix or the transition matrix. In this matrix, each row represents a page, and each column represents a hyperlink. If there is a hyperlink from page i to page j, then the (i,j) entry of the matrix is 1/N, where N is the number of hyperlinks on page i. If there is no hyperlink from page i to page j, then the (i,j) entry of the matrix is 0.

The PageRank scores of the pages can then be found by solving the system of equations R = cM R, where R is a vector containing the PageRank scores of each page, M is the hyperlink matrix, and c is a damping factor used to avoid problems with sinks and sources. This is an eigenvalue problem, and the solution corresponds to the eigenvector of the matrix M with the largest eigenvalue.

In summary, the PageRank algorithm is essentially an application of linear algebra, specifically matrix algebra and eigenvector analysis, to the problem of ranking web pages based on their hyperlink structure. By representing the web as a matrix and using linear algebra techniques to solve the resulting system of equations, we can efficiently compute the PageRank scores of millions of web pages and provide users with relevant and useful search results.

Conclusion:
In conclusion, PageRank is an algorithm used to rank web pages based on their importance and relevance to a user's search query. It uses applied linear algebra to calculate the page ranks, which involves solving a large system of linear equations. By iterating the computation until convergence, the algorithm is able to produce an accurate and reliable ranking of web pages.

PageRank has numerous applications, including search engine optimization (SEO), web page ranking, and spam detection. It is widely used by search engines like Google, which use it to rank search results and improve the relevance of their search results for users. The algorithm has also been used in other domains such as social networks, recommendation systems, and citation analysis.

One important feature of the PageRank algorithm is its ability to deal with rank sinks, which are pages that accumulate rank but never distribute any rank. The introduction of a rank source, represented by the vector E, helps to overcome this problem and ensures that the PageRank values are well-defined and unique.

In summary, the PageRank algorithm is a powerful tool that has revolutionized the way we search for and access information on the internet. Its applications are diverse and far-reaching, and it has become an indispensable tool for web developers, marketers, and data scientists alike.

Latex Synopsis link:
https://www.overleaf.com/project/6411f6a59363419062e8d098