Google's PageRank Algorithm For Web-Indexing

Sanchit Jalan,Kripi Singla,R. Shaanal

June 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 in Algorithm

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.Eigen Vector are calculated using Iterative methode till the vectors stabilise to some constant..

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 and Markov Chains

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 behaviour and is used to model

the random surfer's movement through the web graph.Markov chain is used in the PageRank algorithm to model the behavior of web surfers as they navigate through web pages.

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 behaviour. 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.

Linear Algebra Concepts Used

PageRank algorithm uses several Linear Algebra concepts like Markov Chains, Stochastic Matrix,Graph Theory and Adjacency Matrix and finally Eigen Values and Eigen vectors.This section of the document explains these concepts in a very concise manner.
Though Graph Theory is not a sub branch of Linear Algebra but they are interdependant on each other and in the document we are describing briefly how is it used .

Graph Theory and Adjacency Matrix

A graph is an object that consists of a non-empty set of vertices and another set of edges. In this case, we can refer to graphs as a network, vertices as nodes, and edges as links connecting the nodes.
A directed graph or digraph is a set of nodes and a collection of directed edges that each connects an ordered pair of vertices.
For any two vertices i and j of a directed graph, if there is an edge from i to j or from j and i, the two vertices are adjacent.
An adjacency matrix is an n $\times$ n matrix containing 1's in its entries on row i, column j of the matrix if there is an edge from node i to node j and 0's otherwise.
Example Illustrating Graph Theory and Adjacency Matrix:-

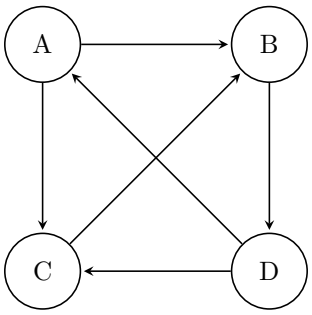


Figure 1: A Directed Graph

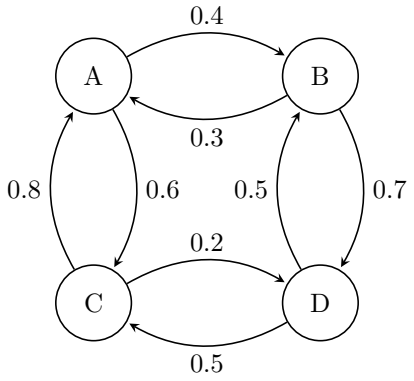
The Above graph is a Basic Digraph in where there 4 Vertices and 6 vertices.Adjacency Matrix A of the above Graph is as follows:

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

In the above matrix we can associate Row 1 as Node A , Row 2 as Node B ,Row 3 as Node C and Row 4 as Node 4.Similarly for Columns.
In the above Digraph since there is an edge between A and B so there is an entry of 1 in the A(1,2).

Markov Chains

A mathematical model that describes an experiment or measurement that is performed many times in the same way, where the outcome of a given experiment can affect the outcome of the next experiment. The process starts at an initial state, namely x0, and transitions successively from one state to another, say $x_1, x_2, \ldots, x_k$. The outcome of a given state depends only on the immediately preceding state.



Using the above diagram as an example to illustrate the concept we can say that assuming that a person is at Node A . The probability of the person going to Node B from A is 0.4 and to Node C from A is 0.6 . It does not depend which state person was earlier than Node A.

Stochastic Matrix

A transition matrix, is a special type of stochastic matrix or probability matrix which is used in Markov Chain Models, is a square matrix that represents the probabilities of transitioning between states in the Markov chain. It provides a concise and systematic way of describing the dynamics of the Markov chain.
A column-stochastic matrix is a square matrix in which all entries are greater than or equal to zero (nonnegative) and whose columns are probability vectors.
A right stochastic matrix is a square matrix of nonnegative real numbers whose rows add up to 1.
A left stochastic matrix is a square matrix of nonnegative real numbers whose columns add up to 1.

Formally, let's assume we have a Markov chain with k states. The transition matrix T is defined as an (k x k) matrix, where each element T(i, j) represents the probability of transitioning from state i to state j.

The (i, j) th element of the transition matrix T represents the probability of moving from state i to state j in a single time step.

The transition matrix satisfies the following properties:

1. Each element T(i,j) is non negative : $T(i, j) \geq 0$ for all i,j
2. The sum of probabilities in each row is equal to 1: $\sum_j T(i, j) = 1$

The state matrix at state $k + 1$ is denoted as x_{k+1}, while T represents the transition matrix. The state matrix at state k is denoted as x_k. The equation can be written as follows:

$$x_{k+1} = T \cdot x_k \tag{2}$$

To illustrate an example of Transition Matrix using the Markov Chain Model in Figure 1. The Transition Matrix of the above Markov chain model is as follows:

$$\begin{bmatrix} 0 & 0.4 & 0.6 & 0 \\ 0.3 & 0 & 0 & 0.7 \\ 0.8 & 0 & 0 & 0.2 \\ 0 & 0.5 & 0.5 & 0 \end{bmatrix}$$

In the above matrix we can associate Row 1 as Node A , Row 2 as Node B ,Row 3 as Node C and Row 4 as Node 4.

Eigen Values and Eigen Vectors

An eigenvector of a square matrix A is a nonzero vector $\vec{x}$ such that $A\vec{x} = \lambda\vec{x}$ for some scalar λ, where λ is an eigenvalue.
If A is a column-stochastic matrix, there exists an eigenvalue $\lambda = 1$.

The characteristic equation of a square matrix A is used to calculate the eigen value of a matrix and is given by:

$$\det(A - \lambda I) = 0$$

where I is the identity matrix and λ is the eigenvalue.

Given eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ for a matrix A, the corresponding eigenvectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ can be obtained by solving the equation:

$$A\vec{v}_i = \lambda_i \vec{v}_i$$

where $i = 1, 2, \ldots, n$.
The eigenvectors can be represented as:

$$\vec{v}_1 = \begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{n1} \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{n2} \end{pmatrix}, \ldots, \vec{v}_n = \begin{pmatrix} v_{1n} \\ v_{2n} \\ \vdots \\ v_{nn} \end{pmatrix}$$

Example How Eigen Values and Eigen Vectors are calculated in 2x2 Matrix:

Let A be a 2x2 Matrix whose eigen values and vectors need to be calculated..

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

Using characterstic equation of a Matrix

$$\det(A - \lambda I) = 0$$

we get the following matrix

$$A = \begin{bmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{bmatrix}$$

Solving this determinant we get (3 - λ)(3 - λ) - 1 = 0

Simplifying this equation we get λ^2 - 6λ + 8 = 0

Solving this for λ we get 2 values as : $\lambda = 4$ and $\lambda = 2$

Therefore the Eigen Values for the given Matrix is 4 and 2.

To find the corresponding Eigen Vectors from Values we can use the following method:

$$A\vec{v} = \lambda_i \vec{v}$$

where:

$$\vec{v_1} = \begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix}$$

$$\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix} = \lambda \begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix}$$

Substituting Values of λ as 4 we get the following equations:

$$\begin{bmatrix} 3v_{11} + v_{21} \\ v_{11} + 3v_{21} \end{bmatrix} = \begin{bmatrix} 4v_{11} \\ 4v_{21} \end{bmatrix}$$

Solving this we get v_{11}=v_{21}

So Eigen Vector corresponding to λ=4 is :

$$\vec{v_1} = \begin{bmatrix} x \\ x \end{bmatrix}$$

where x $\neq$ 0

State Of The Art Literature

The remarkable thing about google's page rank algorithm was its applicability in various fields such as biotechnology.
Gene Rank was a breakthrough technique developed using Google PageRank Algorithm to generate prioritised gene lists by exploiting biological background information. Gene rank is a modified version of the PageRank algorithm which uses the same maths that is applied in search engine optimization.
While going through the many variations of PageRank in the recent times we go through it's application in Too Central To fail. So, The "Too Central to Fail" (TCF) systemic risk measure using the PageRank algorithm is a method to

assess the potential impact of the failure of a specific entity within a networked system. It's a link analysis algorithm that assigns importance scores to web pages based on their connectivity and incoming links.

Gene Rank

What is Gene Rank?

Gene rank refers to a measure or score that quantifies the importance or relevance of a gene within a particular biological context. It is commonly used in bioinformatics and genomics to prioritize genes based on their potential significance or involvement in specific biological processes, diseases, or experimental conditions.

Why Gene Rank?

Interpretation of simple microarray experiments is usually based on the fold-change of gene expression between a reference and a "treated" sample where the treatment can be of many types from drug exposure to genetic variation. Interpretation of the results combines lists of differentially expressed genes with previous knowledge about their biological function. The PageRank algorithm employed by the popular search engine Google – tries to automate some of this procedure by generating prioritised gene lists by exploiting biological background information.

How is PageRank Implemented in GeneRank?

GeneRank utilizes the principles of PageRank to assess the importance or relevance of genes based on their relationships in a gene network. Here's a general overview of how PageRank is used in GeneRank:

1. Gene Network Construction: A gene network is created, representing the relationships between genes. This network can be derived from various biological data sources, such as protein-protein interaction networks, co-expression networks, or functional association databases.

2. Network Representation: The gene network is represented as a graph, where genes are nodes, and the relationships between them are represented by edges. Each gene is connected to other genes it interacts with or shares functional associations.

3. Node Importance Calculation: The PageRank algorithm is applied to the gene network graph to calculate the importance or relevance score for each gene. Initially, all genes are assigned equal importance scores.

4. Iterative Score Update: The importance scores are iteratively updated based on the principle that the importance of a gene is proportional to the importance of the genes it is connected to. This process continues until the scores converge to stable values.

5. Damping Factor: PageRank incorporates a damping factor, typically set to 0.85, to model the probability of a random jump or teleportation between

genes in the network. It prevents the algorithm from getting trapped in isolated gene clusters.

6. Final Gene Ranking: After convergence, the genes are ranked based on their importance scores. Genes with higher scores are considered more relevant or important within the gene network.

"Too central to fail" systemic risk measure

Introduction

So, The "Too Central to Fail" (TCF) systemic risk measure using the PageRank algorithm is a method to assess the potential impact of the failure of a specific entity within a networked system. It's a link analysis algorithm that assigns importance scores to web pages based on their connectivity and incoming links

How is PageRank implemented?

To apply the PageRank algorithm to measure systemic risk, we need to represent the interconnectedness of entities within a system as a network. Each entity is represented as a node, and the relationships or connections between entities are represented as edges. The strength of the connections can be based on factors such as financial interdependencies, operational dependencies, or other relevant measures.

Here are the steps to calculate the TCF systemic risk measure using the PageRank algorithm:

1. Define the network: Identify the entities within the system and establish the connections between them. This network should capture the relevant relationships and interdependencies between the entities.

2. Assign initial scores: Assign an initial PageRank score to each entity in the network. The initial scores could be equal for all entities or based on some predefined criteria.

3. Iterative computation: Use the PageRank algorithm to iteratively update the scores of each entity based on the importance of the entities that are connected to it. The iterative computation involves calculating the PageRank score for each entity by considering the scores of its neighboring entities.

4. Damping factor: Introduce a damping factor to prevent the score from flowing indefinitely within the network. Typically, a damping factor of 0.85 is used, which means that an entity has an 855. Convergence: Repeat the iterative computation step until the PageRank scores converge. This convergence indicates that the scores have stabilized and further iterations would not significantly change the scores.

6. TCF measure: Once the PageRank scores have converged, the TCF systemic risk measure can be calculated. Entities with higher PageRank scores are considered more central within the network and are perceived as having a higher systemic risk. These entities are often referred to as "too central to fail" since their failure could potentially have a significant impact on the overall system.

Other Applications

There are several other algorithms which implement PageRank which are as follows:

Protein Rank

ProteinRank. The goal of ProteinRank is similar in spirit to that of GeneRank. Given an undirected network of protein-protein interactions and human-curated functional annotations about what these proteins do, the goal is to find proteins that may share a functional annotation. Thus, the PageRank problem is, again, a localized use. The teleportation distribution is given by a random choice of nodes with a specific functional annotation. The PageRank vector reveals proteins that are highly related to those with this function, but do not themselves have that function labeled.

Item Rank

A recommender sys- tem attempts to predict what its users will do based on their past behavior. Netflix and Amazon have some of the most famous recommendation systems that predict movies and products, respectively, that their users will enjoy. Localized PageRank helps to score potential predictions in many research studies on recommender systems.

Contributions

- Sanchit contributed towards the writing document on LaTeX.Found how the algorithm works and the application of Linear Algebra in it .Also explained the concepts of Linear Algebra briefly
- Kripi Singla also did research on working of algorithm and application of Linear Algebra . Also worked on SOTA part for Gene Rank .
- R. Shaanal worked on SOTA part for Too Central To fail..

References

- Related Works and Introduction
- Algorithm Analysis
- Linear Algebra Concepts in PageRank
- Purdue University Paper for SOTA
- Gene Rank