IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

# Google PageRank Web Search

**Sudhanshu Jaiswal[1], Samiksha Kulaskar[2], Dr. Vandana Mulye[3]**

[1]*MSc Blockchain Technology, MITWPU, Pune, Maharashtra*

[2]*MSc Blockchain Technology, MITWPU, Pune, Maharashtra*

[3]*Assistant Professor, Dept of Computer Science, MITWPU, Pune, Maharashtra*

**Abstract:**

Given a few number of interlinked web pages, we need to sort the pages according to relevance with a given query. The whole idea is to create a mini search engine based on Google's Page Rank Algorithm in this paper. The working of the algorithm begins with the mathematical formula where Web pages are sorted and ordered according to their importance, where there can be changes or modification in the formula on a mathematical basis.

**Keywords:** PageRank, Web, Search Engine, Information Retrieval.

## I. Introduction

PageRank is defined as one of the oldest techniques which is based on casting votes to web pages to increase the performance of retrieving the information on the Web; The method is used by Google, which is a Web based search engine developed at Stanford University. Just like a research paper is more important because it has lots of citations, a web page is also more important if lots of other web pages refer to it i.e. have outlink to the page. The importance of a scholarly article is more if it has citations from other important articles similarly a page with outlinks from other important pages has higher importance than those which does not. Further we say that if a page has ‘n’ outlinks, it distributes all of its' importance equally to all the pages it is referring to. So it is like a democratic system in which pages are voting for other pages by giving equal part of their vote to all pages which are being referred to by it. The value of vote describes the importance of a page.

## II. Related Works

There are many related works which use the techniques of link based to sort and rank web pages. The algorithms used basically are dependent on the input query and the users get ranking of web pages accordingly, i.e. The user gets the ranking of web pages is dependent on the specific query they put on the search engine. One of the most popular and oldest is Kleinberg’s Algorithm. This algorithm is compared to the SALSA Algorithm and it is the improvement of the latter and also there are various ranking algorithms which are based on analysis of link and dependent on query; where performance of the algorithms are compared and their

theoretical aspects are developed. Luca Pretto did an analysis on Google's PageRank theoretically. All these different aspects and formulations deal with the one and only theory of Google's PageRank.

## III. Literature Survey

| Sr. No | Authors | Publication Year | Title | Description |
|---|---|---|---|---|
| 1. | M. R. Henzinger | 2001 | Hyperlink analysis for the Web | Hyperlinks play a key part in how the internet works. They show relationships between sites and pages. Search engine algorithms use hyperlink analysis |
| 2. | R. Lempel, S. Moran | 2001 | SALSA: The stochastic approach for link-structure analysis | SALSA takes web analysis into uncharted realms, devising ground-breaking stochastic tools. An unprecedented leap for assessing web structures. Methodologies reimage |
| 3. | J. M. Kleinberg | 1999 | Authoritative sources in a hyperlinked environment | Authority dynamics on the web get examined. He-lping understand how web info spreads. Hype-rlinked environments were the foundation. Web information dissemination is explained. |
| 4. | S. Brin, L. Page | 1998 | The anatomy of a large scale hypertextual Web search engine | Provides essential knowledge on stochastic processes vital for medical disciplines, together with web analytics. |
| 5. | S. Karlin | 1966 | A First Course in Stochastic Processes | PageRank provides, revolutionizes web search and information retrieval methods. |
| 6. | A. Borodin, G. O. Roberts, J. S. Rosenthal, P. Tsaparas | Unknown | Finding authorities and hubs from link structures on the World Wide Web | Examines the architecture and functionality of major Internet search engines, and presents indexing and ranking problems. |
| 7. | L. Page, S. Brin, R. Motwani, T. Winograd | Unknown | The PageRank citation ranking: bringing order to the Web | Examines techniques for identifying authentic sources and hubs on the Web, advancing the understanding of Web topology and information transmission mechanisms. |
| 8. | Kurt Bryan, Tanya Leise | Unknown | The $25,000,000,000 Eigenvector: The Linear Algebra behind Google | Vectors and eigenvalues are key. Google's search depends on what these reveal about websites. Math shows search's nature: careful analysis of websites' traits like quality, relevance, connectivity. |
| 9. | David Austin, Grand Valley State Univ. | Unknown | How Google Finds Your Needle in the Web's Haystack | Google's search engine utilizes complex algorithms to retrieve and rank information from the vast expanse of the internet. This intricate process involves analyzing and sorting through an immense amount of data. |
| 10. | Amy N. Langville, Carl D. Meyer | Unknown | Google's PageRank and Beyond: The Science of Search Engine Rankings | The scientific rules that make Google's PageRank method work are explored here, giving ideas into how search engines rank pages. PageRank relies on complex mathematical algorithms to sort websites. |

## IV. Proposed Methodology

1. Making and Storing web pages
2. Search Engine
3. Adjacency Matrix formation
4. Rank Calculation (using Eigenvector analysis)
5. Sorting of web pages and showing the results

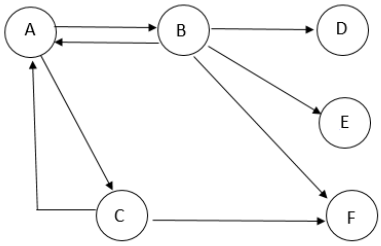


Fig.1: Web Page Graph

## V. Mathematical Aspects

Suppose a page **Pi** has inlinks from a set of pages 'S' in which a page **Pj** has **Lj** outlinks, then we can give a recursive definition of rank of page **Pi** as:

$$\mathbf{Rank(Pi) = \sum Rank(Pj) / Lj \ , Pj \in S}$$

The above problem is similar to egg and chicken problem because to find out rank of page **Pi** we need to know rank of page **Pj**, now if **Pi** and **Pj** both have outlinks to each other, it becomes recursive because again to find out rank of **Pj** we need to know rank of **Pi** which is yet to be known.

Linear Algebra allows us to deal with such problems. From the above equation we can always find out the relationship among the web pages to find out their ranks. Although these relations are recursive, we can represent them into matrix **(AX = b)** form.

A matrix is created, we call it the Hyperlink matrix, **H = [ Hij ]** where the **ith** row and **jth** column entries are :

$$\mathbf{H_{ij} = 1 / Lj \ , \ if \ Pj \in Si,}$$

$$\mathbf{H_{ij} = 0 \quad , \ otherwise}$$

This Matrix **H** has some fundamental features. They have all nonnegative entries. Also, the sum is **1** for all the entries in a column unless and until the page has no any links with respect to that column. Such kinds of matrices which have all nonnegative entries and every column has sum of **1** are called stochastic matrix or Markov matrix.
Let's take an example of graph and create an Hyperlink Matrix, **H :**

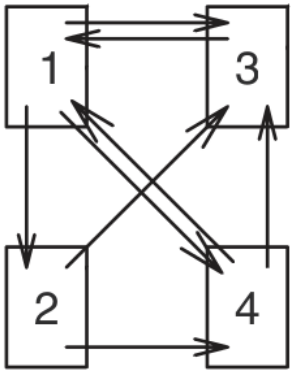


Fig.2: Graph showing connections between web pages

The Hyperlink matrix, **H** for the above graph is represented as:

$$\begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

Eigenvector is created **I = [I(Pi)]** who have PageRanks as their components, i.e. the pages having ranking or its importance. The condition which defines the PageRank may also be represented as:

$$\mathbf{I = HI}$$

We can say that the matrix **H** has eigenvector **I** with eigenvalue **1**. It can also be referred to as the stationary vector of the matrix.
As we know that any square matrix with all columns summing to 1 has an Eigenvalue as 1, hence we are sure to get an Eigenvector corresponding to Eigenvalue 1.

The stationary vector for Markov matrix generated for a graph shows the probability of being at a particular node, now we are looking at our internet as a graph and saying that the the stationary matrix for the matrix H is saying the probability of staying of an internet user at a particular web page. Higher the probability, there are high chances of visiting a web page by random users, since users are visiting a web page with high probability hence definitely the page is more important. Now our problem is limited to finding the Eigenvector corresponding to Eigenvalue 1.

## VI. Shortcomings

The above example looks pretty fine but the Internet is so vast and certainly we cannot easily apply what we have discussed so far. There are two issues with this: Webs which have non-unique rankings and also dangling nodes.

1. **Non unique Rankings:** In terms of rankings, we consider that the dimensions of $\mathbf{V_1(H)}$ (EigenSpace corresponding to Eigenvalue 1) is equal to **1**, such that there must be a unique eigenvector **X** with $\sum$ **Xi** = 1 where we can use scores for its importance. Unfortunately, this case is not true everywhere and hence the hyperlink matrix will not produce the unique rankings.

2. **Dangling Nodes:** A web or graph which has dangling nodes (node with no outlink) yields a matrix **H** where entries in one or more columns are all zero. Hence matrix **H** is sub-stochastic with respect to column, that is, sum of entries in the column are less than or equal to **1**.

## VII. Solutions

Now we have to calculate a new matrix that is both irreducible as well as primitive, we need to make changes in the way that the random surfer navigates throughout the web.

To make our modification, we will first choose a parameter between 0 and 1. If we denote A(n × n) as the matrix where all of the entries are 1, we will get the Google matrix:

$$\mathbf{G} = \alpha\mathbf{H} + (1-\alpha)(1/n)\mathbf{A}$$

Now the matrix **G** is stochastic as it is obtained by combining more than one stochastic matrix and that this matrix has a unique stationary vector **I**.

Sergey Brin and Lawrence Page, who are the developers of Page Rank, selected $\alpha = 0.85$, where $\alpha \in [0,1]$, the matrix **G** is stochastic with respect to column and $\mathbf{V_1(H)}$ is always **1**-dimensional.

## VIII. Conclusions

Here Mathematical aspects are discussed for calculating the importance of the web pages and ranking it. For implementation purposes, we can use GNU Octave and Matlab. Input would be like webpages (HTML) and query related to those pages. Adjacency matrix will be formed. Then it will find eigenvectors corresponding to dominant eigenvalues. According to eigenvector, it will calculate the rank vector. Outputs would be a list of relevant pages accordingly.

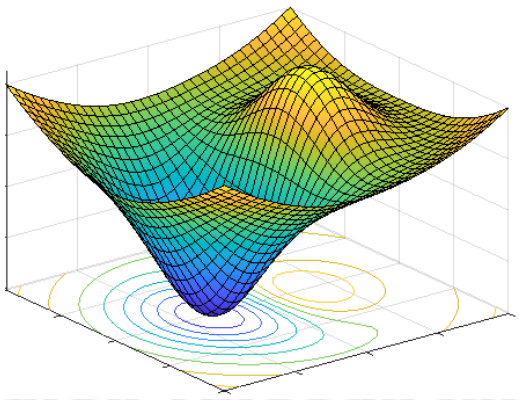

Fig.3: Matlab GUI

## IX. Future Work

In contrast to Google, with **Blockchain- Based Search Engines (BBSEs)**, no one on the web has access to browsing history or other related data as it is distributed across the network. Users will have total control of their data. The Search engine searches through the distributed ledger to display the results. The Decentralised and Distributed ledger stores details in encrypted form.


Fig.4: Decentralized Connection

## X. Acknowledgement

We are thankful to our mentor, Dr. Vandana Mulye for providing tremendous support while working on this paper. She availed all the required facilities and also helped us in how to research and find out the related works relevant to the topic.

## XI. References

[1] “The $25,000,000,000 Eigenvector: The Linear Algebra behind Google"- Kurt Bryan, Tanya Leise.
[2] “How Google Finds Your Needle in the Web's Haystack"- David Austin, Grand Valley State University.
[3] “Google's PageRank and Beyond: The Science of Search Engine Rankings"- Amy N. Langville & Carl D. Meyer.
[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, Sept. 1999.
[5] R. Lempel and S. Moran. SALSA: The stochastic approach for link-structure analysis. ACM Transactions on Information Systems, Apr. 2001.
[6] M. R. Henzinger. Hyperlink analysis for the Web. IEEE Internet Computing, 5(1), Jan.-Feb. 2001.
[7] S. Karlin. A First Course in Stochastic Processes. Academic Press, NewYork, 1966.
[8] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web.
[9] S. Brin and L. Page. The anatomy of a large scale hypertextual Web search engine. In Proceedings of the World Wide Web Conference, 1998.
[10] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the World Wide Web.