Chapter- 9

The Math behind the PageRank

Sneha PS, Ranjini MC*
Centre for Research in Higher Mathematics, Mes Kalladi College,
Mannarkkad-678583, Kerala, India
Affiliated to University of Calicut
*Corresponding Author Ph:7034331009
E-mail : ranjini@meskc.ac.in

Abstract
PageRank, a significant algorithm in Google's search engine, measures website importance. This article explores the calculation methods and applications of PageRank algorithm. We discuss three calculation methods: iterative method, power method, and eigenvalue/eigenvector-based approaches. Applications of PageRank in information retrieval, social networks, academics are discussed. Also the advantages and disadvantages are examined.

Keywords: Internet as graph, Iterative method, Power method, Eigenvalue-Eigenvector method

Introduction
PageRank (PR), named after Larry Page, co-founder of Google, is a key algorithm used by Google Search to rank websites in search results. It measures a webpage's value by assessing links' quantity and quality. Mathematically, PageRank relies on graph theory, calculating a webpage's importance through its PageRank value. This value represents the webpage's ranking using PageRank methods.

Initially launched in 1999, Google became the top search engine, revolutionizing search results with PageRank's introduction. Google's Toolbar (2000) displayed PageRank scores for web pages. Beyond search, PageRank's applications extend to other domains.

This article provides an in-depth examination of PageRank calculation methods, applications, advantages, and disadvantages.

Ranking of Webpages

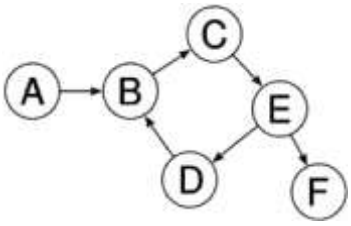
There are two popular algorithms to rank webpages:

1. HITS – Hypertext induced topic research
2. PageRank algorithm.

Internet as graph

The World Wide Web hyperlink structure forms a huge directed graph where the nodes (vertices) represent web pages and directed edges represent the hyperlinks. In the figure 1, A, B, C, D, E, F represents web pages and edges represents hyperlinks.

PageRank works by counting the number and quality of links to a page to determine an estimate of how important the website is. The assumption is that the websites receiving more links from other websites are more important.



.

Figure 1

Types of links

- Inbound link — These are links into the given page from outside, so from other pages.
- Outbound link — These are links from the given page to pages in the same site or other sites.
- Dangling link — These are links that point to any page with no outgoing links.

PageRank Algorithm

PageRank (PR) is an algorithm used by Google search to rank web pages in their search engine results. In other words PageRank is an algorithm that measures the importance of web pages by evaluating the number and quality of links leading to the page. The PageRank algorithm is based on the link between pages of the web.

1. Original summation formula for page rank

Brin and Page the inventors of PageRank began with a simple summation equation, the roots of which actually derived from bibliometrics , the analysis of the citation structure among academic papers. The PageRank of a page P_i denoted by $r(P_i)$ the sum of page ranks of all pages pointing into P_i.

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|} \quad ,$$

Where B_{P_i} is the set of pages pointing into P_i (back linking to P_i in Brin and Page's words) and $|P_j|$ is the number of out links from page P_j . Notice that the PageRank of in linking pages $r(P_j)$

in equation is tempered by the number of recommendations made by P_j, denoted $|P_j|$.

The problem with equation is that the $r(P_j)$ values, the PageRank of pages in linking to page P_i are unknown. To sidestep to this problem, Brin and Page used an iterative procedure. That is they assumed that, in the beginning, all pages have equal PageRank (of say $\frac{1}{n}$, where n is the number of pages in Google's index of the Web). Now the equation given is followed to compute $r(P_i)$ for each page P_i in the index. The equation is successively applied, substituting the values of the previous iterate into $r(P_j)$. Now introduce some more notation in order to define this iterative procedure. Let $r_{k+1}(P_i)$ be the PageRank of page P_i at iteration k+1.

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$

This process is initiated with for all pages and repeated with the hope that the PageRank scores will eventually converge to some final stable values.

Example

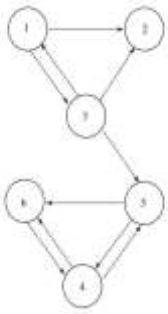


Figure 2

Webpage	Iteration 0	Iteration 1	Iteration 2	PageRank
1	$r_0(P_1) = \frac{1}{6}$	$r_1(P_1) = \frac{1}{18}$	$r_2(P_1) = \frac{1}{36}$	5
2	$r_0(P_2) = \frac{1}{6}$	$r_1(P_2) = \frac{5}{36}$	$r_2(P_2) = \frac{1}{18}$	4
3	$r_0(P_3) = \frac{1}{6}$	$r_1(P_3) = \frac{1}{12}$	$r_2(P_3) = \frac{1}{36}$	5
4	$r_0(P_4) = \frac{1}{6}$	$r_1(P_4) = \frac{1}{4}$	$r_2(P_4) = \frac{17}{72}$	1
5	$r_0(P_4) = \frac{1}{6}$	$r_1(P_5) = \frac{5}{36}$	$r_2(P_5) = \frac{11}{72}$	3
6	$r_0(P_5) = \frac{1}{6}$	$r_1(P_6) = \frac{1}{6}$	$r_2(P_6) = \frac{14}{72}$	2

In iteration 2, $\frac{1}{36}$ is the greatest value, which gives a PageRank of 5. Thus 1 and 3 are the most important websites within this network.

2. Matrix Representation

Using matrices, at each iteration, compute a PageRank vector, which uses a single $1 \times n$ vector to hold the PageRank values for all pages in the index. In order to do this, we introduce $n \times n$ matrix **H** and $1 \times n$ row vector **v**. The matrix **H** is a row normalized hyperlink matrix

with $\mathbf{H_{ij}} = \frac{1}{|P_i|}$ if there is a link from node i to node j and 0 otherwise. Consider once again the tiny web graph of Figure 2.

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

The nonzero elements of row i corresponds to the out linking pages of page *i*, whereas the nonzero elements of column I correspond to the in linking pages of page *i*. Now introduce a row vector $v_{k,}$ which is the PageRank vector at the k^{th}iteration. Generally on n^{th} iteration vector is given by $v_n = H^n v$, which is known as *Power method.* The initial vector *v* is the initial page rank assigned to every page. Here the initial vector *v* is given by the matrix $v = \begin{bmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \end{bmatrix}$

3. Computation using eigen value and eigen vector

Consider the graph of 4 Web pages. First we check the idea of how important each link is. Page 1 has a link to each of the other 3 pages, so each link receives 1/3 of Page 1's importance. Consider Page 2, it has 2 links. So Page 1 and Page 4 receive 1/2 of Page 2's importance.

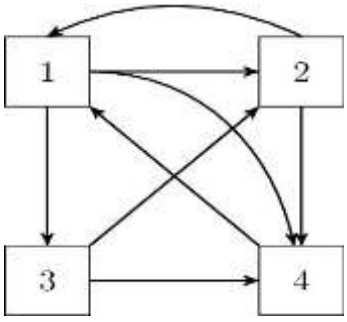


Figure 3

We see the pattern, if a node has k out links, then it passes 1/k of it's importance to each of the nodes it links to. Thus we get a system of equations:

$$x_1 = \frac{1}{2}x_2 + x_4$$
$$x_2 = \frac{1}{3}x_1 + \frac{1}{2}x_3$$
$$x_3 = \frac{1}{3}x_1$$
$$x_4 = \frac{1}{3}x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_3$$

Let's look at this matrix form, called a transition matrix (all the column sum is 1) of the above graph:

$$A = \begin{bmatrix} 0 & 1/2 & 0 & 1 \\ 1/3 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 1/2 & 0 \end{bmatrix}$$

To find PageRank vector. we have to consider eigenvectors and eigenvalues. Since the transition matrix is stochastic, by definition we know that the matrix has an eigenvalue of 1. This implies that there exists an **x** such that **Ax**= 1**x**. This x is called the eigenvector, and it is the PageRank vector. Now solve **Ax** = 1**x**, then we get the matrix:

$$\begin{bmatrix} 0 & 1/2 & 0 & 1 \\ 1/3 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

To find PageRank vector **x,** we substitute to try to get variables in one variable form. As a result we get, $x = \frac{x_1}{12}\begin{bmatrix} 15 \\ 6 \\ 4 \\ 9 \end{bmatrix}$. This means that any scalar multiple of it will be an eigenvector for matrix corresponding to the eigenvalue 1. We multiply it by the scalar 1/30 (1/sum of entries)

and we get the PageRank vector $\begin{bmatrix} 0.5 \\ 0.2 \\ 0.13 \\ 0.3 \end{bmatrix}$, Page 1 ranking highest, followed by Page 4, Page 2 and Page 3.

It would easy to solve our four equations by hand. There is also systematic method for solving such a system of simultaneous equations (the Gaussian algorithm) which can be carried out by a computer. However finding PageRank for the entire World Wide Web involves a system of more than 14 billion Page rank equations, and it is not feasible to try to solve these equations directly, even by a computer.

Google Matrices

A Google matrix is a particular stochastic matrix that is used by Google's PageRank algorithm. The matrix represents a graph with edges representing links between pages. The PageRank of each page can then be generated iteratively from the Google matrix using the power method. However, in order for the power method to converge, the matrix must be stochastic, irreducible and aperiodic. Then final Google matrix G can be expressed as:

$$\boldsymbol{G} = \alpha\, \boldsymbol{M} + (1-\alpha)\frac{1}{n} ee^T$$

Where:

- **G** is the Google matrix
- α is the damping factor, typically set between 0.8 and 0.9
- **M** is the transition matrix
- $\frac{1}{n} ee^T$ *is* the teleportation matrix, usually a matrix of equal probabilities (i.e.,the matrix $(\frac{1}{n}, \frac{1}{n}, \ldots \frac{1}{n})$).

Applications

1. **Search Engines:** It helps determine the importance of web pages based on the website's link structure. Search engines such as Google include PageRank as one of many factors to rank search results and provide users with more accurate and helpful information.

2. **Citation Analysis:** In academic research, the PageRank algorithm can be applied to analyze citation networks. The algorithm can identify influential articles or researchers in a given field by treating academic articles as nodes and citations as links. This information helps to understand the impact and importance of scientific work.
3. **Content Recommendation:** PageRank can recommend related or similar content on a website or platform. By analyzing the link structure between different pages or articles, the algorithm can identify related pages and recommend them to users as related or recommended.
4. **Fraud detection:** PageRank can be used in fraud detection systems to identify suspicious fraud patterns. Also PageRank has applications in literature, sports, biology, chemistry, neuroscience and physics.

Real life Applications

- **Sports** – The PageRank algorithm has been used to rank the performance of: teams in the National Football League (NFL) in the USA; individual soccer players; and athletes in the Diamond League.
- **Ranking tweets in twitter** - To use PageRank for ranking tweets in Twitter we can construct a synthetic graph as follows. Represent each user and each tweet by a node. Draw a directed link from user A to a user B if A follows B. Also, draw a directed edge from a user A to a tweet t if A tweets or retweets t. Now we can apply PageRank algorithm on this graph to obtain a ranking for tweets.
- **Suggesting friends over twitter:** Personalized PageRank is used by twitter to present users with other accounts they may wish to follow. told represent the wide variety of existing applications of PageRank point to a rich future for the algorithm in research contexts of all types. It seems intuitive that any problem in any field where a network comes into play might benefit from using PageRank algorithm.
- **Scientific research and academia:** PageRank has been used to quantify the scientific impact of researchers. The underlying citation and collaboration networks are used in conjunction with PageRank algorithm in order to come up with a ranking system for individual publications which propagates to individual authors.

In neuroscience, the PageRank of a neuron in a neural network has been found to correlate with its relative firing rate.

For the analysis of protein networks in biology PageRank is also a useful tool.

Scientists were able to use PageRank to help determine the position of water molecules in an ionic solution, enabling them to find the best ways to remove nuclear waste and toxic chemicals. PageRank essentially maps where toxic chemicals are likely to pool in the solution, enabling a waste cleanup team too quickly and efficiently contain and remove the toxic or radioactive contaminant.

Advantages of PageRank algorithm

- Objective and unbiased
- Quality-focused
- Resilience to manipulation
- Scalability
- Query-independent
- Foundation for other algorithms.

Disadvantages of PageRank algorithm

1. It's ordering does not favor current events. According to the algorithm, old pages typically have more votes because they have more links from other reputable pages. This means that a new page will not be as reputable until it has gained exposure and links from other pages.
2. It's use in ranking papers in the citation network, it does not account for the size of a field. The number of citations per paper in each field varies widely depending on the discipline, for example, an average paper is cited about 6 times in life sciences, 3 times in physics, and about 1 time in mathematics. The algorithm is therefore more likely to give a paper in a mathematics field a lower score than a paper in a life sciences field.
3. Rank can be raised by buying "links".

Conclusion

PageRank is a global ranking of all webpages, regardless of their content. Throughout the paper we discussed different methods in calculating rank of web pages. Graph theory and linear algebra concepts are behind the methods of PageRank calculation. We have found a number of applications and advantages of PageRank. Even though it has many advantages there is some disadvantages also. Overall findings with PageRank suggest that the structure of the Web graph is very useful for a variety of information retrieval tasks.

References

1. Amy N. Langville and Carl D. Meyer, Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University press, 2006
2. S. Brin and L. Page, The anatomy of a Large-scale Hypertextual Web Search Engine. Computer Networks, 30 (1998), pp. 107-117
3. R. S. Wills, "Google's PageRank: The math behind the search engine", Math. Intelligencer 28:4 (2006), 6-11. MR 2272767.
4. Brain moor, Mathematics Behind Google's PageRank Algorithm. Denton, Texas, 2018