PURDUE UNIVERSITY NORTHWEST

COMPUTATIONAL MATHEMATICS [MA 57700]

FINAL PROJECT REPORT

GOOGLE'S SEARCH METHOD (PAGERANK ALGORITHM)

Submitted by

Vaishak Balachandra
MS in Computer Science
Purdue University Northwest

Submitted to

Nicolae Tarfulea, PhD
Professor of Mathematics
Department of Mathematics and Statistics
Purdue University Northwest

Table of Contents

1. Introduction

PageRank is one of the most significant algorithms in the history of the development of web search technology, developed by Google Search to evaluate the ranking of each web page within a connected network of web pages in search results. The term "web page", was coined by the co-founder Larry Page and their team. The PageRank algorithm evaluates the importance of web pages [1]. In the advancement in the recent years, many upgradations had happened in the PageRank algorithm itself, to match the demands and to smoothen the ease of users in finding the right web page they are looking for, and to increase the chances of being clicked on the right and the trusted ones [2].

1.1. Historical Background

While attending Stanford University in 1996, Larry Page and Sergey Brin created PageRank. It was a component of a study that sought to create a novel search engine. An intriguing interview with Sergey's advisor, Stanford Computer Science professor Héctor Garcia-Molina, provides insight into the development of the PageRank algorithm. Sergey devised a system for ranking webpages according to "link popularity." This implied that a page would rank higher the more links it included. Page and Brin acknowledged that Scott Hassan and Alan Steremberg were crucial to the project's success in creating Google. In addition, Page and Brin collaborated with Rajeev Motwani and Terry Winograd to publish the first PageRank article and the Google search engine prototype in 1998. They later founded Google Inc., which would grow into the massive company that powers our modern search engine. PageRank is an essential part of all Google's web-search technologies, even though it is just one of several criteria that affect Google search ranks [1].

The term "PageRank" craftily combines the concept of a web page with the name of its creator, Larry Page. A patent (U.S. patent 6,285,999) protects the PageRank technique, and it's an official trademark of Google. More interestingly, Stanford University, and not Google, is the patent owner of the same. However, Google is the sole company to be licensed under the Stanford patent. The patent was resold in 2005 to Google for an unbelievable 336 million USD, in return, Stanford received 1.8 million shares of Google [1].

PageRank was influenced by citation analysis, first conceived by Eugene Garfield at the University of Pennsylvania in the 1950s. It was also inspired by Hyper Search, by Massimo Marchiori from the University of Padua. This was around the same time PageRank was conceived [1].

1.2. Importance of PageRank Algorithm

Several sources confirm that PageRank is still very much relevant today. In 2017, Google's Gary Illyes took to Twitter to confirm that the algorithm still uses PageRank [3]. The importance of the PageRank algorithm can be summarized as follows [4]:

a. PageRank enhances the ability of search engines to return the best, quality results that prove highly relevant to users.
b. PageRank supports the ranking of results based on the importance of a site, which is biased towards the central pages that act as hubs for key information.
c. Understanding the calculation of PageRank of a website can provide insights into the basic assumptions of ranking algorithm and procedure of ranking a page.
d. The PageRank of a website is extremely important since it is one of the significant determining factors if your site will show up on the first page or the last page when people search for anything to do with your business or product.

By understanding PageRank, you will be able to make educated SEO decisions that can increase your organic visibility. For example, you can focus on creating valuable content that naturally attracts links and interlink your pages strategically to increase their relevance and authority.

1.3. Main Contributors

The PageRank algorithm can be attributed to several persons, including Larry Page, Sergey Brin, Héctor Garcia-Molina, Scott Hassan, Alan Steremberg, Rajeev Motwani, Terry Winograd, and Massimo Marchiori. Hence, the basic effort of these persons resulted in the algorithm. From these persons, Page and Brin were the ones who developed its concept and initial implementation. Their joint endeavour gave rise to an entire generation of internet searches, the very method in which ranking is determined on the WWW [1].

2. Methods

2.1. PageRank Algorithm

The PageRank algorithm gives a probability that shows how likely it is for somebody to reach a certain page by clicking on links. PageRank can be calculated for any number of documents. Some research papers assume that at the start of the process, this probability is shared equally among all documents in the collection. The PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to reflect the theoretical true value more closely. The PageRank value for each page is updated iteratively based on the formula:

$$\boldsymbol{PR(X)} = \sum \frac{\boldsymbol{PR(Y)}}{\boldsymbol{L(Y)}}$$

where,
PR(X) is the PageRank of page X.
Y represents the pages that link to X.
L(Y) is the number of outbound links from page Y.

2.1.1. Iterative PageRank Calculation

The Iterative PageRank Calculation makes up another important part of the algorithm for PageRank. Its main function is to create an improvement in the existing rank of a page by using a step-by-step and repeated process through a network. This section outlines how the algorithm performs these updates and achieves convergence over multiple iterations.

2.1.2. Overview of the Iterative Process

a. Initialization:

At the beginning of the algorithm, each page in the network is assigned an initial PageRank value. This is often set to 1/N, where N is the total number of pages, ensuring an equal distribution of rank at the start.

b. Rank Calculation:

The PageRank of each page is updated based on the ranks of the pages that link to it. The formula for this update is:

$$\boldsymbol{PR(X)} = \sum \frac{\boldsymbol{PR(Y)}}{\boldsymbol{L(Y)}}$$

c. Iterative Updates:

The algorithm iteratively updates the PageRank values for all pages in the network. In each iteration, for a page, a new PageRank value is calculated according to the ranks of its backlinks in the previous iteration. This keeps getting repeated until a stopping condition is reached.

2.1.3. Convergence and Stopping Criteria

a. Convergence:

The iterative process is designed such that it converges to a stable set of PageRank values; meaning that after some rounds, the rank values will remain steady with very slight changes in every round.

b. Stopping Criteria:

Common stopping rules include: The algorithm terminates when the difference in PageRank values between iterations is less than a predetermined threshold. Max. Iterations: This one's better—stop the process after a given number of iterations. This ensures that it is not an endless run. The PageRank algorithm works by repeating a process that imitates how users browse the web. This helps it give importance to web pages based on how they are linked to each other.

The convergence of PageRank values reflects the relative importance of each page within the overall network, ultimately leading to meaningful rankings that are utilized in various applications discussed in the following section.

2.1.4. Example Algorithm (Trial Case):

To explain the PageRank algorithm, we look at a simple network of five pages: A, B, C, D, and E. Each page is assigned a starting PageRank value of 0.2, which in most cases is 1 divided by N, where N represents the total number of pages under consideration. This is illustrated in Fig 1 [5]. Here, Page A links to both B and C. Page B links to A, C, and D. Page C links to A, D, and E. Page D links to A and E. Page E does not link with any other pages.

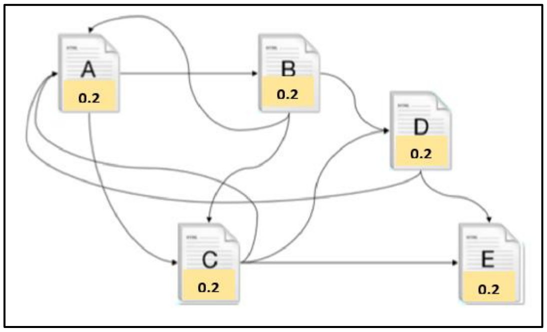

Fig 1. Initial Network of Pages and Links

The PageRank value for each page is updated iteratively based on the formula:

$$\boldsymbol{PR(X)} = \sum \frac{\boldsymbol{PR(Y)}}{\boldsymbol{L(Y)}}$$

Initial PageRank Values: Since, there are 5 pages, each page starts with an initial PageRank value of 0.2 (i.e., 1/5). i.e., PR(Y) = 0.2, where Y = {A,B,C,D,E}. The number of outbound links of each of the pages are L(Y) = {2,3,3,2,0}. The Table 1., shows the updated values of the pages after 1 iteration.

Table 1. PageRank Values After Iteration 1

Page Nodes	Iteration 0 [PR(X)]	Iteration 1 [PR(X)]
A	0.2	$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} = \frac{0.2}{3} + \frac{0.2}{3} + \frac{0.2}{2} = \frac{7}{30}$ = 0.2333
B	0.2	$PR(B) = \frac{PR(A)}{L(A)} = \frac{0.2}{2} = \frac{1}{10}$= 0.1000
C	0.2	$PR(C) = \frac{PR(A)}{L(A)} + \frac{PR(B)}{L(B)} = \frac{0.2}{2} + \frac{0.2}{3} = \frac{1}{6}$ = 0.1667
D	0.2	$PR(D) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} = \frac{0.2}{3} + \frac{0.2}{3} = \frac{2}{15}$= 0.1333
E	0.2	$PR(E) = \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} = \frac{0.2}{3} + \frac{0.2}{2} = \frac{1}{6}$ = 0.1667

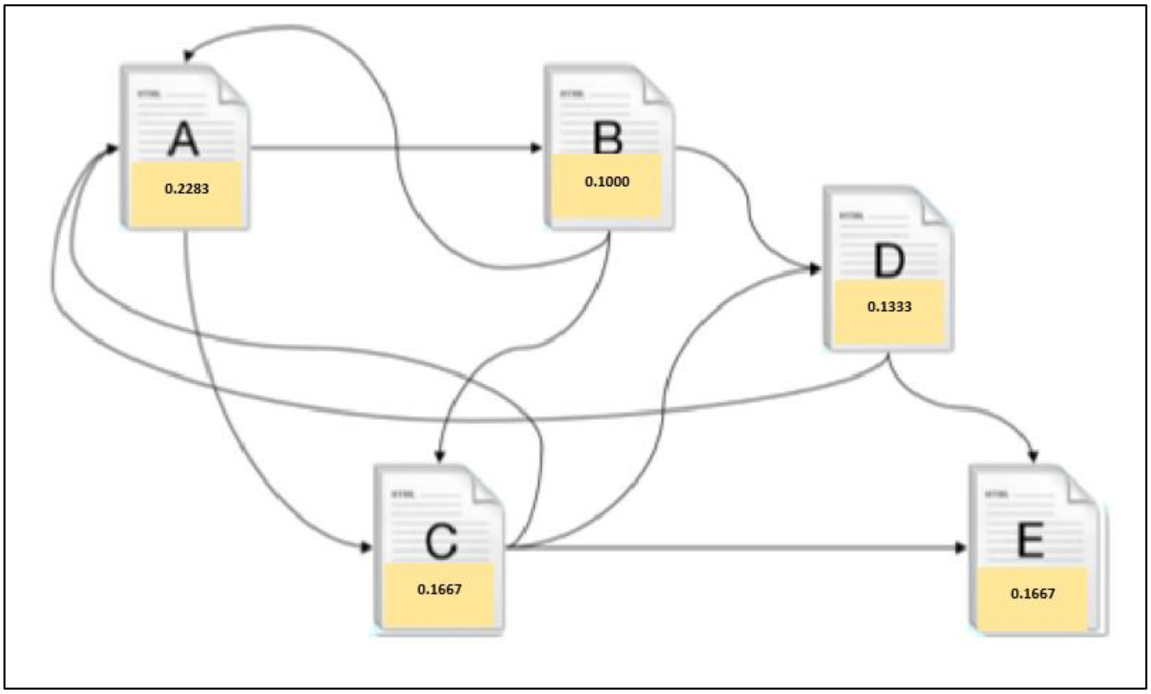

Fig 2. Network of Pages and Links in Iteration 1

MATLAB CODE:

```
% PageRank ALgorithm
num_pages  =  5;
initial_rank = 1 /num_pages;
L  = [ 0 1 1 0 0;
  1 0 1 1 0;
 1 0 0 1 1 ;
  1 0 0 0 1 ;
  0 0 0 0 0];
outbound_links = sum(L ,2 );
PR = initial_rank * ones(num_pages ,1);
new_PR = zeros(num_pages,1 );
for i = 1: num_pages
    for  j = 1:num_pages
  if   L(j , i) == 1
  new_PR (i) = new_PR(i ) + PR(j) / outbound_links(j );
  end
  end
end
disp ('PageRank values after 1 iteration:')
for i  =  1 :num_pages
   fprintf('PR(Page %c) = %.4f\n',char('A' + i -1),new_PR(i));
end
```

Output:

```
PageRank values after 1 iteration:
PR(Page A) = 0.2333
PR(Page B) = 0.1000
PR(Page C) = 0.1667
PR(Page D) = 0.1333
PR(Page E) = 0.1667
```

2.2. Damping Factor

The damping factor plays a part in the PageRank algorithm; it tells how users behave when browsing the web. Adding a chance element to the algorithm, it ensures that every page can at least get some PageRank, even if it is linked poorly. This section deals with what the damping factor is, its usual value, and why it is important in the PageRank calculation. This section will explain what the damping factor is, its usual value, and why it matters in calculating PageRank.

2.2.1. Concept of the Damping Factor

a. Modeling User Behavior:

The damping factor simulates the likelihood that a user will continue clicking on links versus returning to a random page. This assumes that while users often follow links, they can also get distracted or decide to start a new search.

b. Probabilistic Framework:

The damping factor is often written as d, and is often adjusted to about 0.85; which signifies that the probability of the user being on that page then clicking a link off that page is 85% and the probability of jumping to any other page at random is 15%.

c. Adjusted PageRank Formula

With the introduction of the damping factor, the PageRank formula is adjusted as follows:

$$\boldsymbol{PR(X) = \frac{1-d}{N} + d.\sum \frac{PR(Y)}{L(Y)}}$$

where,
PR(X) is the PageRank of page X.
d is the damping factor (commonly set to 0.85).
N is the total number of pages in the network.
Y are the pages linking to X.
L(Y) is the number of outbound links from page Y.

2.2.2. Significance of the Damping Factor

The significance of the Damping Factor are as follows:

a. **Prevents Rank Accumulation**: By introducing the chance of jumping to any page, the damping factor prevents the rank from accumulating solely among highly linked pages. This ensures that less-linked pages can still receive a portion of the PageRank.

b. **Stabilizes the Algorithm**: The damping factor contributes to the stability of the PageRank computation.

c. **Improves Relevance of Rankings**: This way, the damping factor gives more relevance to PageRank scores while considering random navigation; this would get us a better representation for user behavior and web structure ranking system.

In summary, the damping factor plays a vital role in the PageRank algorithm by simulating realistic user behavior and ensuring that all pages maintain some level of importance within the web ecosystem. Its proper implementation enhances the effectiveness of the algorithm in generating meaningful page rankings, which will be further explored in the next section on applications of PageRank.

2.2.3. Example Algorithm (Adjusted Case):

To demonstrate the modified PageRank algorithm, we first consider a very simple network: A, B, C, D, and E, where each has an initial PageRank value of 0.2—usually, it is 1/N for N pages under consideration. Initial PageRank Values: Since there are 5 pages, each page is assigned an initial PageRank value of 0.2 (which is 1 divided by 5). Thus, PR(Y) = 0.2 where Y = {A,B,C,D,E}. The number of outbound links for each page is L(Y) = {2,3,3,2,0}. Table 2 shows the updated values of the pages after 1 iteration. The Adjusted PageRank value for each page is updated using the equation 2.2.1..

$$\boldsymbol{\frac{1-d}{N} = \frac{1-0.85}{5} = 0.03}$$

Table 2. Adjusted PageRank Values After Iteration 1 (d = 85%)

Page Nodes	Iteration 0 [PR(X)]	Iteration 1' [PR(X)]'
A	0.2	$PR(A) = 0.03 + 0.85\left(\frac{7}{30}\right)$ = **0.2283**
B	0.2	$PR(B) = 0.03 + 0.85\left(\frac{1}{10}\right) = \mathbf{0.115}$
C	0.2	$PR(C) = 0.03 + 0.85\left(\frac{1}{6}\right) = \mathbf{0.1717}$
D	0.2	$PR(D) = 0.03 + 0.85\left(\frac{2}{15}\right) = \mathbf{0.1433}$
E	0.2	$PR(E) = 0.03 + 0.85\left(\frac{1}{6}\right) = \mathbf{0.1717}$

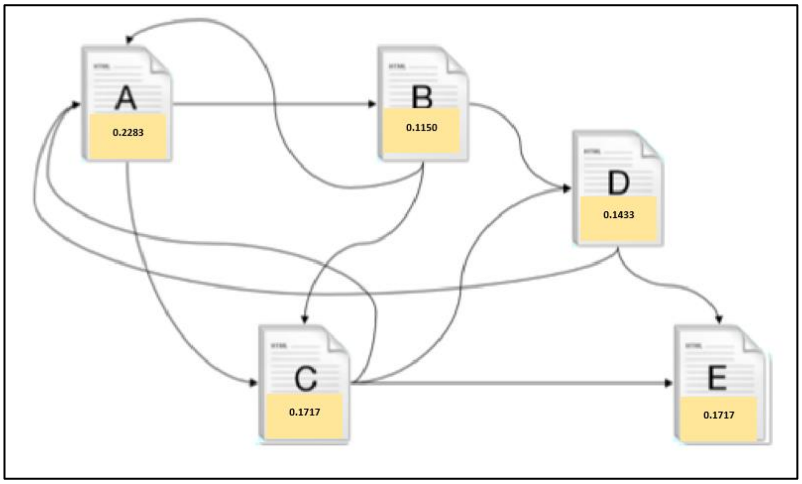

Fig 3. Network of Pages and Links in Iteration 1 (Adjusted PageRank Algorithm)

MATLAB CODE:

```
% Adjusted PageRank
d = 0.85;
pages = 5;
initial = 1/pages;
adjustment = (1 - d)/pages;
L = [0 1 1 0 0;
 1 0 1 1 0;
 1 0 0 1 1;
 1 0 0 0 1;
 0 0 0 0 0];
out_links = sum(L, 2);
PR = initial*ones(pages, 1);
new_PR = zeros(pages, 1);
for i = 1:pages
 for j =1:pages
 if L(j, i) ==1
 new_PR(i) = new_PR(i)+ d*(PR(j)/out_links(j));
 end
 end
 new_PR(i) = new_PR(i) +adjustment;
end
fprintf('Adjusted PR Values:\n');
for i =1:num_pages
 fprintf('Page %c: %.4f\n', 'A' + i - 1, new_PR(i ));
end
```

Output:

```
Adjusted PR Values:
Page A: 0.2283
Page B: 0.1150
Page C: 0.1717
Page D: 0.1433
Page E: 0.1717
```

3. Applications

3.1. Application 1: Ranking Webpages [4]

To assess a webpage's relevance, PageRank considers both the quantity and quality of incoming connections, which makes it incredibly successful at ranking webpages. A strong mechanism for identifying high-quality material is created by PageRank, which gives priority to links from authoritative pages in contrast to basic link-counting techniques. This method presents the most reliable and pertinent results, assisting consumers in navigating the enormous internet [4].

Benefit: The algorithm's ability to gauge a webpage's significance based on its link structure guarantees impartial and accurate rankings. Additionally, by considering the legitimacy of connecting pages, it helps prevent manipulation by making it more difficult for spam strategies to elevate low-quality content to a high ranking.

3.2. Application 2: Citation Networks [6]

By examining how frequently and prominently an article is cited by others, PageRank in citation networks assists in identifying important papers, authors, or journals. This is especially helpful in academic domains where identifying research that has an impact is crucial. The algorithm emphasizes foundational works in a field by adding the prestige of mentioning sources [6].

Benefit: PageRank offers a scalable and objective method of ranking articles, making it possible for researchers to find important studies fast. Additionally, it prevents biases that could result from subjective or solely manual evaluations, guaranteeing equitable acknowledgment of significant labor.

3.3. Application 3: Social Network Analysis [7]

PageRank is a valuable tool for assessing an individual's influence in social networks by examining their relationships and interactions. This algorithm is perfect for finding important influencers or leaders in a network since it ranks higher people who are well-connected and gain engagement from other influential users [7].

Benefit: PageRank makes it possible for companies and groups to target high-impact individuals with their marketing or outreach initiatives, boosting campaign efficacy. Furthermore, it offers an unambiguous, data-driven approach to comprehending social influence trends in intricate networks.

References

1. https://en.wikipedia.org/wiki/PageRank
2. https://www.positional.com/blog/pagerank
3. https://www.semrush.com/blog/pagerank/#why-pagerank-still-matters
4. https://www.educba.com/pagerank-for-website/
5. https://web.stanford.edu/class/cs54n/handouts/24-GooglePageRankAlgorithm.pdf
6. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf
7. https://cambridge-intelligence.com/social-network-analysis/