The PageRank Algorithm: Simplified Algorithm, Complete Algorithm, Damping Factor and Sink Handling

Manav Prabhakar
December 16, 2020

Abstract

The PageRank algorithm was developed by Larry Page and S. Brin. The primary objective of the algorithm was to give a way in which the pages can be returned to the user in case of a query made by the user on the search engine. Two approaches have been discussed here, the first is a simplified one, the second is the complete algorithm with a damping factor and an approach for handling the sinks.

I. Introduction

The page rank for a page represents its importance. The value of the page rank lies between 0 and 1. The importance of a website is determined by the number of websites that link to the website in consideration. In other words, consider a website A (W_A). Let there be some other websites W_B, W_C, W_D etc.

Let all these websites have a link to W_A, then W_A will be the most important website. The page rank of the website ranges between 0 and 1. Each page is given some amount of initial rank.

II. The PageRank Algorithm

i. Simplified PageRank Algorithm

Consider a set of pages U = {U_1, U_2, …U_N}. Now, to determine the page rank of a page, say U_j, we have

$$PageRank(U_j) = \sum_{k=1}^{k} \frac{PageRank(U_k)}{Number\ of\ outgoing\ links\ from\ U_k} \ \forall\ k :\ U_k \in inb(U_j)$$

$$Where\ inb(U_i)\ means\ the\ set\ of\ pages\ that\ have\ a\ link\ towards\ U_j$$

This is the up-dation step for the simplified algorithm.

For initialization,

- Iteration 0: All ranks are initialized to 1/Total number of Pages.
- Iteration 1: The new page rank is found by updating the previous values with the above given updation rule.

Drawbacks:

The simplified algorithm does not work when there are nodes with no incoming edges. All these nodes will give the rank to A, resulting in a zero rank for them. This wasn't intended.

ii. Complete PageRank Algorithm

For tackling the drawbacks of the simplified page rank algorithm, some modifications are done to bring out the complete PageRank Algorithm.
There is a damping factor (d) and total number of pages (N) which is included to modify the algorithm.

Mathematically, the modified algorithm can be written as

$$PageRank(U_j) = \frac{1-d}{N} + d.\sum_{k=1}^{k} \frac{PageRank(U_k)}{Outgoing\ links\ from\ U_k}\ \forall\ k:\ U_k \in inb(U_j)$$

The symbols have their usual meaning as described in the previous equations.

How it tackles the drawbacks of the Simplified Algorithm

The damping factor is used to distribute d fraction of the page rank to the neighbors while (1-d) page rank to everyone in the graph. This helps the pages with no incoming links to have some page rank, thus, handling what simplified algorithm was unable to do.

When to stop

The algorithm can either be stopped after a set number of iterations (say 100) or it can be stopped as soon as the page ranks converge. This means that the page rank obtained in Iteration n is same as that obtained in Iteration n-1.

iii. Sinks

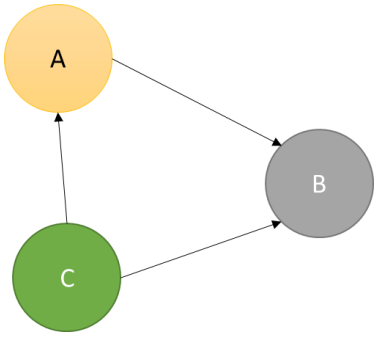

Figure 1: 3 Pages: A, B and C. B denotes a sink node, since it has no outbound edges.

A page is referred to as a sink, if there are no outgoing links from the page. The page rank can get accumulated at such pages.

There are two ways to handle this problem: -

- Adding Edges Method
- Distributing Rank Method

Adding Edges Method

In this, first the sinks are identified. Post identifying the sinks, an outgoing link is established to all other pages.

Distributing Rank Method

In this, the rank of the sink is distributed evenly among all the other pages.

Intuitively, we can estimate that both will have similar results. It is noteworthy that both functionally, indeed both yield similar results.

Also, it is important to do this computation before entering the page rank algorithm. Once, the sink has been handled, then the page rank algorithm needs to be executed.

III. Experimentation and Results

The PageRank algorithm was implemented using python. For sink handling, Adding Edges Method was used. Apart from that, both the simplified algorithm and the complete

algorithm was used. The distribution of page ranks was calculated for a large range of damping factors to determine the role of their value in page rank distribution.

```
Adjacency Matrix :

 [[0 0 0 0]
 [1 0 1 0]
 [1 0 0 0]
 [1 0 0 0]]

Initialized Rank:  [0.25 0.25 0.25 0.25]

Page Status: Denoting number of inbound edges, number of outbound edges and Sink node status

 [[3. 3. 1.]
 [0. 2. 0.]
 [1. 1. 0.]
 [0. 1. 0.]]

Adjacency Matrix after sink node handling (if any) P:

 [[0 1 1 1]
 [1 0 1 0]
 [1 0 0 0]
 [1 0 0 0]]
```

Figure 2: Test Adjacency Matrix, Page Status (column 0: Inbound Edged, column 1: Outbound Edges, column 2: Page Status – 1 if sink, else 0

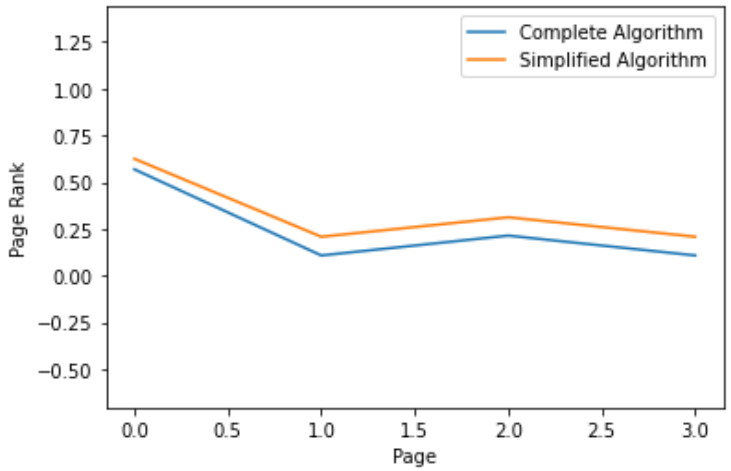

Figure 3: Comparing page ranks returned by the two algorithms after 1 iteration.

The two algorithms seem to work well even though there was a sink node, primarily because of sink handling being done before. Apart from that, the page ranks given by the simplified algorithm are not very distinct and are slightly higher when compared to the complete algorithm.

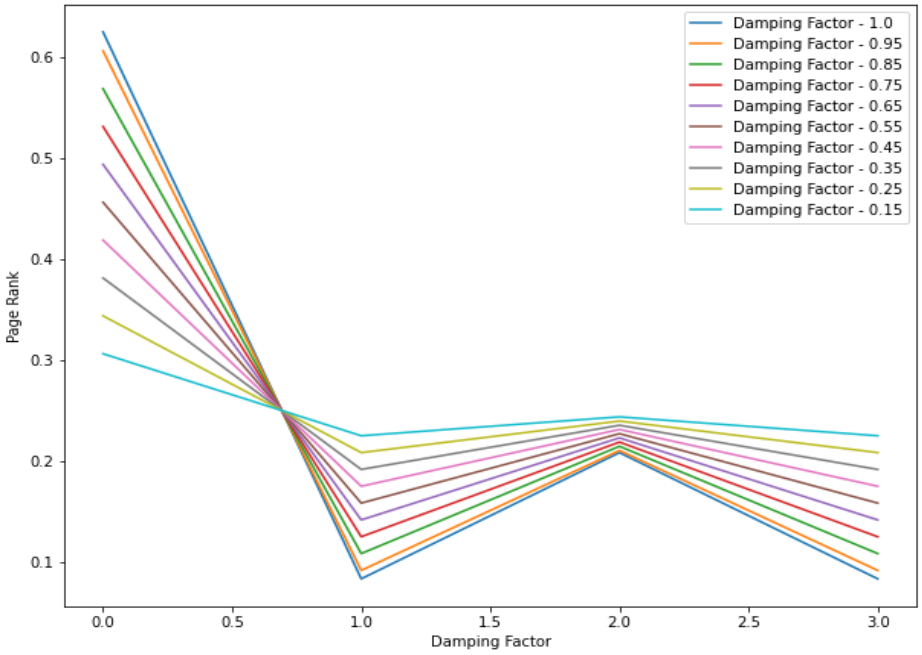

Figure 4: Comparing page ranks obtained with different damping factors

The above graph shows a comparison and trend for the value of the damping factor and the page rank obtained. Clearly, we can see that as the damping factor increases, the values of page ranks obtained resembles that of those obtained in Figure 2 for the simplified algorithm. In other words, we can say that as the damping factor moves to 1, the complete algorithm moves to simplified one which can also be verified mathematically by putting the value of d = 1 in the equation of Complete algorithm.

IV. Conclusion

The experimentation and the implementation of the page rank algorithms proved to be successful. The results obtained were similar to those expected. We also observed that as the damping factor moves to 1, the complete algorithm and the simplified algorithm gives similar results. Based on the observations, we can have a suitable damping factor for our algorithm. The damping factor suitability would though depend on the number of pages as well as it clearly affects the way page rank is affected. The Adding Edges method also worked fine without which we won't have been able to achieve the results because the number of outbound edges = 0 would imply a indefinite term in the updation rule of both the algorithms.

References

[1]https://www.quora.com/How-is-PageRank-calculated-What-was-the-initial-PageRank-How-does-the-algorithm-begin
[2] http://www.cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm
[3] https://en.wikipedia.org/wiki/PageRank