Tribhuvan University

Institute of Science and Technology

Seminar Report

On

"PageRank algorithm using different damping factor"

Submitted to

Central Department of Computer Science and Information Technology
Tribhuvan University, Kirtipur
Kathmandu, Nepal

Submitted by

Utsav Baral
Roll no. 630/077

In partial fulfillment of the requirement for Master's Degree in Computer Science and Information technology (M.Sc. CSIT), 1st semester

Tribhuvan University

Institute of Science and Technology

Supervisor's Recommendation

I hereby recommend that this seminar report, prepared under the supervision of Mr. Arjun Singh Saud entitled "**PageRank algorithm using different damping factor**" be accepted as fulfillment in partial requirement of the degree of Masters of Science in Computer Science and Information Technology.

………………………………...
Asst. Prof. Arjun Singh Saud
(Supervisor)
Central Department of Computer Science
and
Information Technology

LETTER OF APPROVAL

This is to certify that the seminar report prepared by **Mr. Utsav Baral** entitled "**PageRank algorithm using different damping factor**" in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science and Information Technology has been well studied. In our opinion, it is satisfactory in the scope and quality as a project for the required degree.

Evaluation Committee

……………………………….........	………………………………..
Asst. Prof. Sarbin Sayami	Asst. Prof. Arjun Singh Saud
(H. O. D)	(Supervisor)
Central Department of Computer Science	Central Department of Computer Science
and Information Technology	and Information Technology

……………………………………

(Internal)

Acknowledgement

I would like to express my sincere gratitude to **Asst. Prof. Mr. Arjun Singh Saud**, for his valuable guidance in carrying out this work under his supervision. I am grateful for his excellent guidance, trust and correction of my seminar work.

With immense pleasure, I would like to thank **Asst. Prof. Mr. Sarbin Sayami**, Head of Central Department of Computer Science and Information Technology, Tribhuvan University for his encouragement in completion of the seminar.

At last, but not least, with immense pleasure, I submit by deepest gratitude to the Central Department of Computer Science and Information Technology, Tribhuvan University, and all the faculty members of CDCSIT for providing the platform to explore the knowledge of interest.

Utsav Baral (630/077)

Abstract

Ranking a page in search engine or internet is one of the most intense works. PageRank algorithm is depended on the link analysis in which ranking of web page is decided based on outbound links, inbounds links and damping factor. In the modern computer society, ranking of web page with higher relevance is one of the most important factors to increase one's market or business.

In this seminar report, a comparison has been done between the web pages when the damping factor varies. A total of seven different damping factors has been implemented within the constant links between web pages and determined different page ranks between them. The value of damping factor being 0.85 means 85% of random web surfers follows link in between the web pages and remaining 15% of random web surfers teleport to another page randomly. There is high page rank of the web page which have highest inbound link and low page rank for the web pages with lower inbound links.

Keywords: PageRank algorithm, outbound links, inbound links, damping factor

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Overview

PageRank is one of the methods to determine a page's relevance or importance. The Page Rank algorithm was used and enhanced by Lawrence Page and Sergey Brin [1]. It works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. PageRank algorithm is depended on the link analysis in which ranking of web page is decided based on outbound links and inbounds links.

Furthermore, there are many PageRank methods such as Google PageRank, Aitken's PageRank, Power Method, Adaptive PageRank etc. which are helpful in determining the rank of the page.

Damping factor d or α is a probability value that lies between 0 and 1. It defines how much time a random web surfer follows the links from the same page. Its standard value is considered as 0.85 according to S. Brin and L. Page [1].

The main aim of this seminar report is to demonstrate how change in damping factor affects the ranking or relevance of the page.

Here PageRank algorithm has been used for determining the rank of the page based on different damping factor.

1.2 Some Important Definitions

1.2.1 Outbound Links

Outbound links refers to the links from the given page to the pages in the same site or other sites.

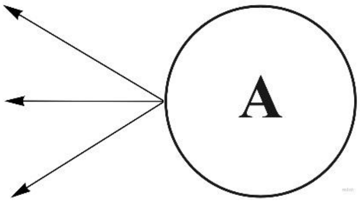


Figure 1: Outbound Links

1.1.1. Inbound Links

Inbound links refers to the links into the given page from outside the other pages or other sites.

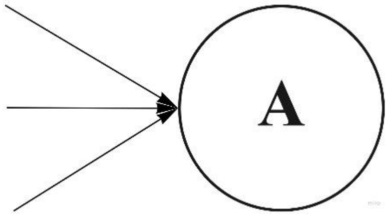


Figure 2: Inbound Links

1.1.2. Damping Factor

The damping factor parameter state that how much time random web surfer follow hyperlink structure than teleporting. The value of α or d is exceptionally empirical and in current scenario α = 0.85 is considered as suggested by Brin and Page [1].

1.1 Problem Statement

The purpose of this seminar report is to demonstrate how damping factor affect the relevance of the page. In today's world web surfer can easily teleport to a different page if there lacks any relevant information rather than following the hyperlink in the same page so different damping factor along with 0.85 is chosen.

1.2 Objective

The main objective of this seminar report is to use different damping factor and determine the effect it has on the rank of the page.

1.3 PageRank Algorithm Formula

The PageRank of page *A* is given as:

$$PR(A) = \frac{1-d}{N} + \ d * (PR(T1)/C(T1)\ + \cdots . + PR(Tn)/C(Tn))$$

Where,

Let *A* be the page whose page rank is *PR(A),*

N is the total number of pages,

Let $PR(T_1)$ is the page rank of pages T_1 which link to page *A*,

$C(T_1)$ is the number of outbound links from page T_1, *d* is the damping factor that is assumed to be between 0 and 1 usually 0.85 but in this seminar report different damping factors will be used. It follows direct links,

(1-d) is the probability of teleporting to a different page. It follows non-direct links.

To calculate page rank of any page it is required to know the page rank of each page that point to it and number of outbound links from each of those pages. Since PageRank is an iterative algorithm which means each time, calculation is carried out, closer estimate to the final value is being obtained. So, one important thing to remember is that each value is calculated and iterated until the results starts to converge.

Chapter 2: Literature Review

Various researches have been done on the Link analysis using different methods and algorithms. PageRank algorithm is one of the most used algorithms which has been used by the American multinational technology company, Google.

S. Brin and L. Page on their paper [1], has proposed that PageRank can be thought of as a model of user behavior. By assuming that there is a random user or surfer that keeps on following the link from the pages but eventually gets bored and starts on another page. And, the damping factor *d* is the probability of the random user following the links thus making *1-d* the probability of that random user teleporting to a different page. By adding the damping factor *d* to a single page or a group of pages, allows to make it nearly impossible to deliberately mislead the system in order to get a high ranking.

In paper [2], the authors claim that web surfer these days gets worn out too easily on the web because of non-availability of the relevant information and can easily teleport to new web pages rather than following the hyperlinks' structure. So, choosing the value of damping factor other than 0.85 is relevant. Furthermore, the authors have given an experimental analysis of PageRank computation for different value of the damping factor and have observed that for the value of *d=0.7*, PageRank method takes fewer numbers of iterations (25-30%) to converge than *d=0.85*.

In paper [3], the authors have discussed about the change in behavior of PageRank with respect to change in *d* and was found to be useful in link-spam detection. In most cases, suggestion of *d=0.85* given by S. Brin and L. Page [1] is used. Some studies performed by the authors indicated that for real-world graphs, values of *d* close to 1 do not provide a meaningful ranking.

Chapter 3: Methodology

3.1 Algorithm for PageRank

Step 1: Assign each page with an initial rank of $1/N$, N being the total number of pages.

Step 2: Calculate page rank of each page. Example, $PR(A) = (1-d)/N + d(PR(T1)/C(T1) + \cdots . + PR(Tn)/C(Tn))$ calculates page rank of page A.

Step 3: Once page rank of all the pages is obtained, calculate other iteration using the new obtained value of Page ranks.

Step 4: Continue iterating until the page ranks stabilize.

3.2 Example

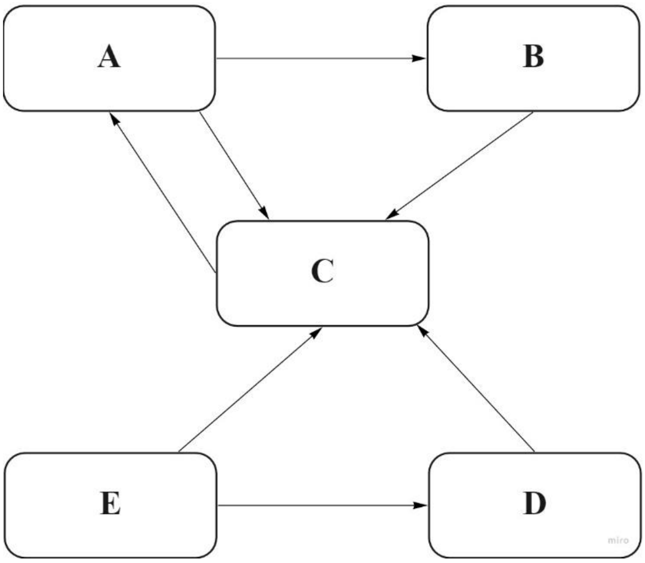


Figure 3: Set of Pages

Let us consider a network of pages which has outbound and inbound links. Each page has at least one outbound link i.e., *C(A) = 2, C(B) = 1, C(C) = 1, C(D) = 1, C(E) = 2*. Initially page rank of each page is given by *1/N* i.e., total number of pages *(N)* = 5. So, page rank of each page by initially is taken as *PR(A) = PR(B) = PR(C) = PR(D) = PR(E) = 1/5*. Now, page rank is calculated which is shown below.

PR(A) = (1-d)/N + d(PR(C)/C(C))*

= (1-0.85)/5 + 0.85(0.2/1)*

= 0.2

PR(B) = (1-d)/N + d(PR(A)/C(A))

= (1-0.85)/5 + 0.85(0.2/2)

= 0.115

Similarly,

PR(C) = 0.54

PR(D) = 0.115

PR(E) = 0.03

Multiple iteration is performed based on the result obtained in previous iterations. Once results start to converge to one value for multiple iterations then the obtained value can be taken as the result. In above example, the result obtained after 30 iterations are;

PA(A) = 0.35846798

PA(B) = 0.18234897

PA(C) = 0.38643305

PA(D) = 0.04275

PA(E) = 0.03

Now, let’s carry out calculations using different damping factor. For this seminar report damping factor values to be used are, *d=0, 0.3, 0.5, 0.7, 0.9, 1*

Following are the results obtained for different damping factors after 30 iterations.

Table 1: Page rank of pages

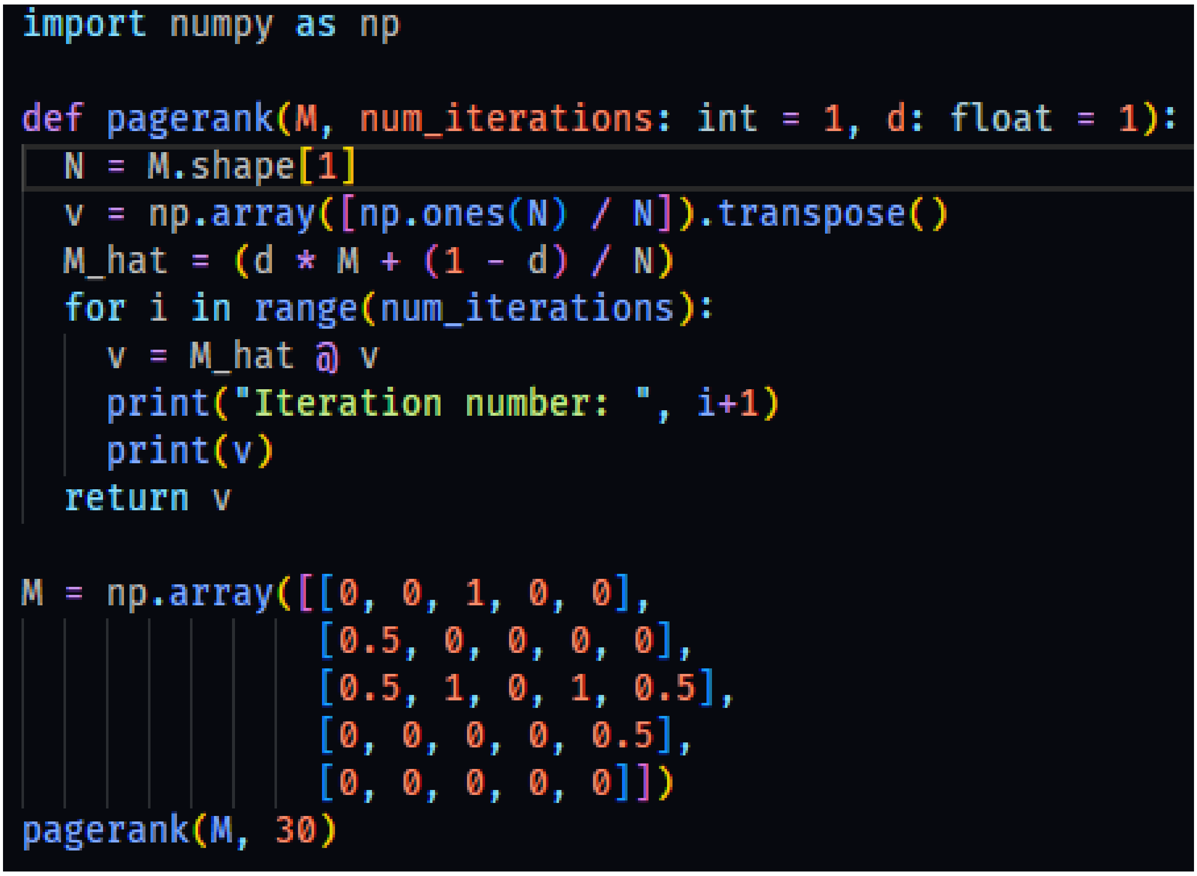
Pages *d*	A	B	C	D	E
0	0.2	0.2	0.2	0.2	0.2
0.3	0.22877323	0.17431599	0.29591078	0.161	0.14
0.5	0.26923077	0.16730769	0.33846154	0.125	0.1
0.7	0.31840617	0.17144216	0.36915167	0.081	0.06
0.85	0.35846798	0.18234897	0.38643305	0.04275	0.03
0.9	0.3721904	0.18748615	0.39132345	0.029	0.02
1	0.39998779	0.2000061	0.4000061	0	0

Chapter 4: Implementation

The implementation include in this seminar report is carried out in the python programming, following python libraries are used for implementation purpose.

NumPy: NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices. The operations used in NumPy are as follows:

- **matrix.shape[1]:** The shape property is usually used to get the current shape of an array. In this report, the 'shape [1]' property is used to get the current size of the matrix which is 5.
- **np.array():** This method is used to create array.
- **np.ones():** Returns a new array of given shape and type, filled with ones. So, in this report 'np.ones(5)' returns, [1, 1, 1, 1, 1].
- **np.transpose():** This function is used to reverse the axes of array or in mathematical terms, transposes a matrix.
- **@:** This operator is used for multiplication of two matrices.



```python
import numpy as np

def pagerank(M, num_iterations: int = 1, d: float = 1):
  N = M.shape[1]
  v = np.array([np.ones(N) / N]).transpose()
  M_hat = (d * M + (1 - d) / N)
  for i in range(num_iterations):
    v = M_hat @ v
    print("Iteration number: ", i+1)
    print(v)
  return v

M = np.array([[0, 0, 1, 0, 0],
              [0.5, 0, 0, 0, 0],
              [0.5, 1, 0, 1, 0.5],
              [0, 0, 0, 0, 0.5],
              [0, 0, 0, 0, 0]])
pagerank(M, 30)
```

Figure 4: Code Snippet

Chapter 5: Result and findings

This seminar report demonstrates how damping factor is a critical factor in changing a webpage's ranking in PageRank algorithm whose value can be set in the range [0, 1]. Result indicates following findings based on damping factor *d*.

- All webpages have equal page rank when the damping factor is set to 0. In other words, there is no change in the page rank of the pages if the random surfer does not follow the hyperlinks.
- The difference in page rank increases as the damping factor increases.
- Web pages that do not have any inbound links are likely to end up in the sink when the damping factor is 1. Which means that the random surfer would end up not visiting all the pages if the random surfer follows the hyperlink that is why the page rank of Page D and E is 0 (i.e., no inbound links).

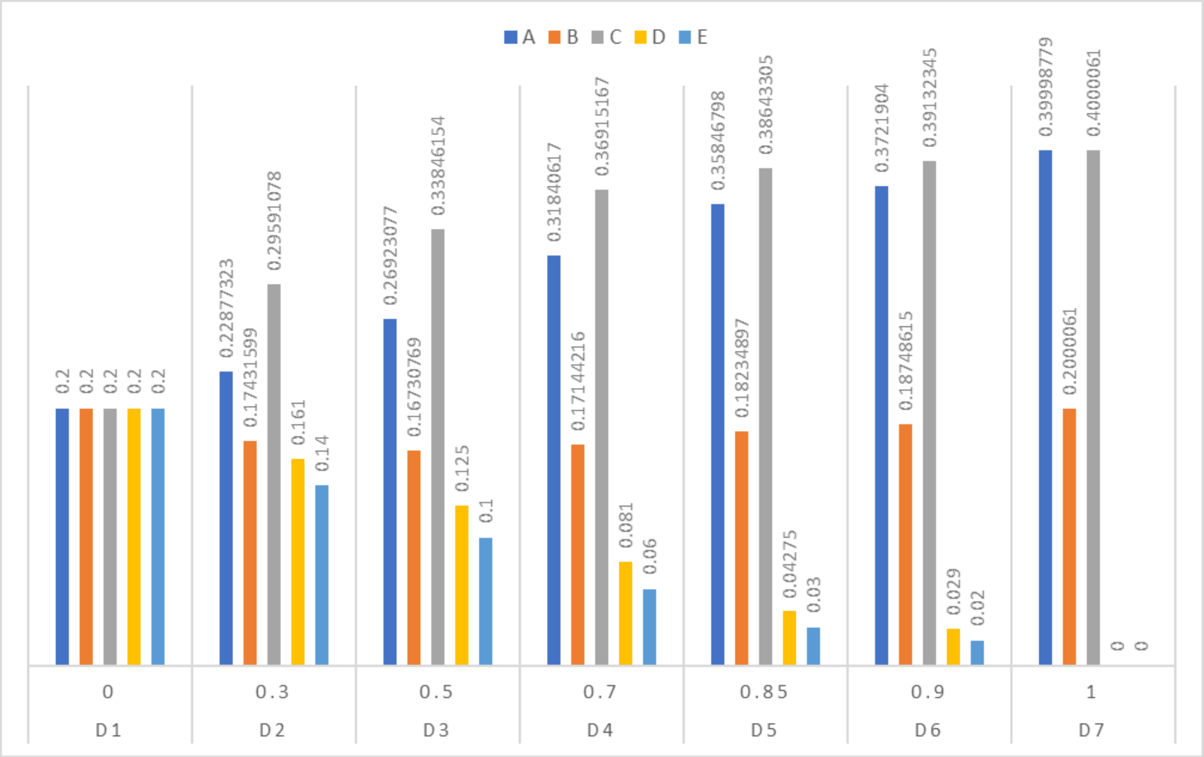


Figure 5: Results produced using different damping factors

Chapter 6: Conclusion

This seminar report mainly presents how damping factor affects the page rank of web pages, even though the link between web pages is kept constant. Here, the rank of web pages with no inbound links decreases when the damping factor is increases. The damping factor ranges from [0, 1], where the value of damping factor being 0.85 means 85% of random web surfers follows link in between the web pages and remaining 15% of random web surfers teleport to another page randomly. In this report, there is comparison done between five web pages whose inbound and outbound links are kept constant and only damping factor is changed. There is high page rank of the web page which have highest inbound link and low page rank for the web pages with lower inbound links. Also, page without any inbound link tends to go to sink when damping factor gets higher.

References

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems,* vol. 30, no. 1-7, pp. 107-117, 1998.

[2] A. K. Srivastava, R. Garg and P. K. Mishra, "Discussion on Damping Factor Value in PageRank Computation," *Modern Education and Computer Science,* vol. 9, no. 3, pp. 19-28, 2017.

[3] H.-H. Fu, D. K. J. Lin and H.-T. Tsai, "Damping factor in Google page ranking," *Applied Stochastic Models In Business and Industry,* no. 22, pp. 431-444, 2006.