Date: 20th March 2020

A Tutorial on Page Rank Algorithm

Shubham Deshmukh

Supervised by: Dr. Othman Soufan
Department of Computer Science
ST. FRANCIS XAVIER UNIVERSITY

INTRODUCTION

WHAT IS GRAPH MINING?

Data Mining is considered as a process of extracting useful hidden information out of large databases containing huge data in an unsupervised way.

A Graph can be defined as a set of nodes or pairs which might be interconnected by edges. A Graph Mining is a process of extracting patterns or sub-graphs of some useful meaning from graphs, which determine the underlying data and can be then further used for classification or clustering. It is essentially the problem of discovering repetitive sub-graphs occurring in the primary graph. The sub-graphs can be a meaningful chunk of data from a larger graph or they can to used to remove trivial re-occurring data/graphs.

Most of the datasets are non-linear or non-flat files, means they have layers, structures, and hierarchy some real-world example of these data are Set of Computers in a Network, Social Network, Webpages on the Internet, A biological data model of a chemical, Workflow of a process, etc. There are several Graph and Web Mining Algorithms that cater to solve complex graph problems. One of them is PageRank(PR).

WHAT IS PAGERANK?

PageRank named after one of the founders of Google Larry Page is an algorithm developed by Larry Page and Sergey Brin in 1996 at Stanford University, It is used by Google search engine to rate the importance of WebPage/Website to display the best set of results in the most relevant order when a user searches something on Google Search. PageRank is a vote given by all other webpages about how important a webpage is, A link to the page counts as a vote of support.

the Internet contains billions of webpages/websites and finding the most relevant webpages was the ultimate goal of the search engines. In the initial days of the internet, the results of a user search were far from relevant results and contained a lot of unwanted pages. That became the real motivation to develop an algorithm like PageRank whose goal is to improve the quality and accuracy of results of a search.

Google applied this at their Google search engine and it laid a foundation for its toolbar PageRank, which used to publicly display webpage's score, However, the Toolbar PageRank is deprecated in the modern days, but the algorithm along with other modern techniques are being used in the backend of google search to solve the webpage's relevancy problem.

ALGORITHM IN DETAIL

PageRank is an algorithm that distributes the probability of a user clicking a random link on the internet will arrive at any specific page. A rank can be calculated by the PageRank algorithm for the collection of documents of any size. The ranking of pages is an iterative process and the scores get refined overtime while Initially the same rank/score (i.e. 1) is given to all the links.

The probability or rank is between 1 to 0, A score of 0.3 is thought to be a 30% chance of a webpage getting reached/opened.

RANDOM SURFER MODEL

The PageRank calculation uses a Random surfer model which is a model of user behavior, where a surfer clicks on links at random with no regard to content.

The PageRank of the pages is dependent upon other pages, and for calculation of PageRank of a given page, pages linking towards that page must have their PageRank Calculated.

The formula for PageRank of a Webpage A：

$$PR(A) = (1- d) + d(PR(Ti)/C(Ti)) +\ldots\ldots+ PR(Tn)/C(Tn))$$

PR(A) is the PageRank of page A.

PR(Ti) is the PageRank of pages Ti which link to A.

C(Ti) is the number of outbound links on page Ti.

d is the Damping factor which can be set between 0 and 1, (Default: 0.85)

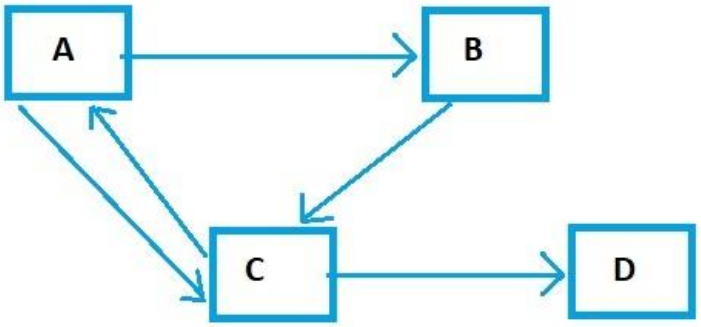
We can derive from the formula that, Inbounds link increases the page's PageRank, also when a page has no outbound link its PageRank cannot be distributed to other pages. These pages with no outbound links are also known as Dangling Links.

The PageRank has an imaginary random surfer who is clicking on links and will eventually stop clicking, the probability that the surfer will continue clicking links is d (Damping Factor).

STEPS

1) *Initialize the PageRank of all the Webpages to 1.*
2) *Calculate the PageRank of all the webpages according to the formula.*
3) *Repeat step two until we get the same PageRank in two consecutive iterations.*

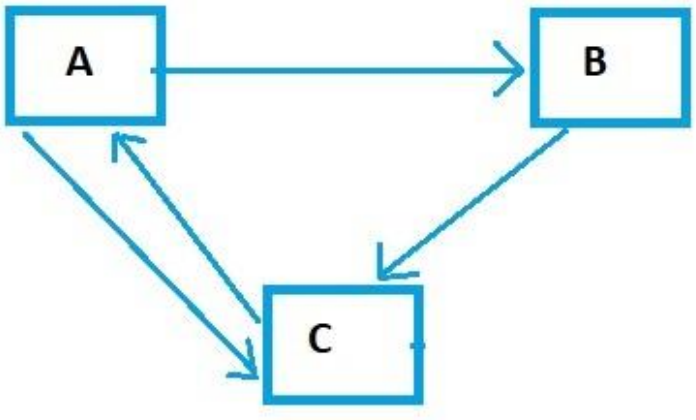
EXAMPLE



The WebPage D is a dangling link here and has no outbound links so it has no significance in the calculation of PageRank so, we can remove it from the diagram.

Also considering the Damping Factor as 0.85,

Initially, the PageRank is considered as 1.



CALCULATION:

For Iteration 1

PR for WebPage A:

$$PR(A) = 1 - d + d(PR(C)/C(C))$$

$$PR(A) = 1 - 0.85 + 0.85(1/1)$$

$$PR(A) = 0.15 + 0.85(1)$$

$$PR(A) = 1$$

PR for WebPage B:

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

$$PR(B) = 1 - 0.85 + 0.85(1/2)$$

$$PR(B) = 0.15 + 0.85(0.5)$$

$$PR(B) = 0.575$$

PR for WebPage C:

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

$$PR(C) = 1 - 0.85 + 0.85(1/2 + 0.575/1)$$

$$PR(C) = 0.15 + 0.85(1.075)$$

$$PR(C) = 1.06375$$

For Iteration 2

PR for WebPage A:

$$PR(A) = 1 - d + d(PR(C)/C(C))$$

$$PR(A) = 1 - 0.85 + 0.85(1.06375/1)$$

$$PR(A) = 0.15 + 0.9041875$$

$$PR(A) = 1.0541875$$

PR for WebPage B:

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

$$PR(B) = 1 - 0.85 + 0.85(1.0541875/2)$$

$$PR(B) = 0.15 + 0.4480296875$$

$$PR(B) = 0.5980296875$$

PR for WebPage C:

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

$$PR(C) = 1 - 0.85 + 0.85(1.0541875/2 + 0.5980296875/1)$$

$$PR(C) = 0.15 + 0.9563549219$$

$$PR(C) = 1.06354922$$

Iteration	Webpage A	Webpage B	Webpage C
0	1	1	1
1	1	0.575	1.06375
2	1.0541875	0.5980296875	1.06354922

The PageRank gets refined after several Iterations and finally gets a stable score to rely on.

PageRank can also be calculated using Matrix Multiplication, however, it's not mentioned in this tutorial.

ADVANTAGES:

- It's faster since the PageRank pre computes the score and takes lesser time.
- It's feasible as the rank is computer while Indexing and not at the time of querying.
- The calculation is not complex and it isn't expensive.

DISADVANTAGES:

- A very good new webpage won't be favored much because it will not have many links while existing older pages even though not of much importance can have a higher rank because of more links present.
- The Content of the pages doesn't matter to the algorithm since the algorithm only works with the links.

- Dangling Links can pose problems having no outbound link and result in unnecessary calculations.
- A-Rank sink problem is faced when the surfer gets in an infinite link cycle due to dangling links.
- Dead Ends can occur(page with no outbound links).
- PageRank can get affected when a group doesn't have a link to another group leaving the model into a spider trap.

IMPROVEMENTS:

- The algorithm can allocate higher rank to more significant pages(with more inbound and outbound links) rather than equally dividing.
- The algorithm can consider the contents of the page as well while calculating the PageRank.
- Steps can be introduced to switch from one network to another based on rank so that unknown or new pages can also be ranked quicker.

CONCLUSION:

The Internet or Web can be seen as a set of interconnected graph nodes (documents or pages) which are connected through links, this is also known as a web graph.

Several Link analysis algorithms are proposed one of them is the PageRank algorithm developed by Google. A page ranking algorithm, which is an application of web mining, is imperative to a success of a search engine since the internet has billions of webpages and a user will require only a few results which need needs to be the most relevant ones, and to get them, an algorithm like PageRank plays a vital role. The algorithm ranks the pages available on the Internet and shows the most relevant ones when the user queries something, based on the computed ranks, and this process makes the navigation simpler and easier for the user.

The PageRank algorithm has some pros and cons like every algorithm, however, it was one of the first to be introduced to solve the problem of the relevancy of web pages on the internet and laid the foundation for future improvement. The PageRank algorithm is not primarily used by google instead used with other algorithms to improve its efficiency in the current days.

REFERENCES:

1. http://pr.efactory.de/e-pagerank-algorithm.shtml
2. https://en.wikipedia.org/wiki/PageRank