CS317

Information Retrieval

Week 14

Muhammad Rafi

May 02, 2019

Link Analysis

Chapter No. 21

Today’s Agenda

- Introduction to PageRank Algorithm
- Iterative algorithm for PageRank
- Example

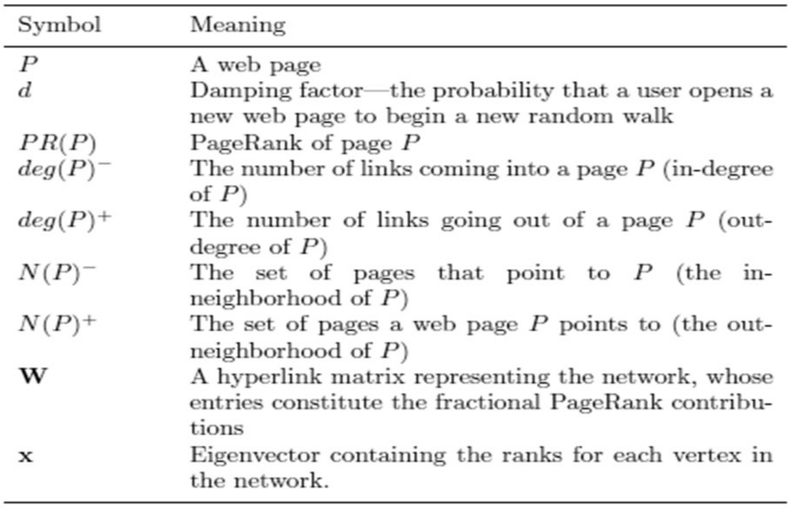
Web as a Graph

- Surfer’s Model
- Markov Chain
  - Discrete time stochastic process
  - It is characterized by an NxN probability matrix
  - Markov property
  - Left Eigen vector (v. P)
- Teleporting
- Damping factor

PageRank

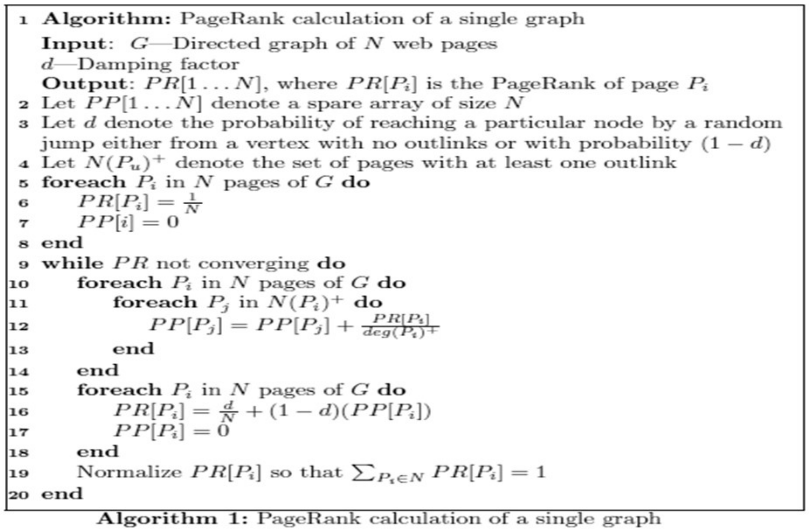
- PageRank is a link analysis algorithm.
- It is named after Larry Page and used by the Google Internet search engine.
- It is patent by Stanford.
- it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of “measuring” its relative importance within the set.

PageRank Algorithm



Symbol	Meaning
P	A web page
d	Damping factor—the probability that a user opens a new web page to begin a new random walk
$PR(P)$	PageRank of page P
$deg(P)^-$	The number of links coming into a page P (in-degree of P)
$deg(P)^+$	The number of links going out of a page P (out-degree of P)
$N(P)^-$	The set of pages that point to P (the in-neighborhood of P)
$N(P)^+$	The set of pages a web page P points to (the out-neighborhood of P)
$\mathbf{W}$	A hyperlink matrix representing the network, whose entries constitute the fractional PageRank contributions
$\mathbf{x}$	Eigenvector containing the ranks for each vertex in the network.

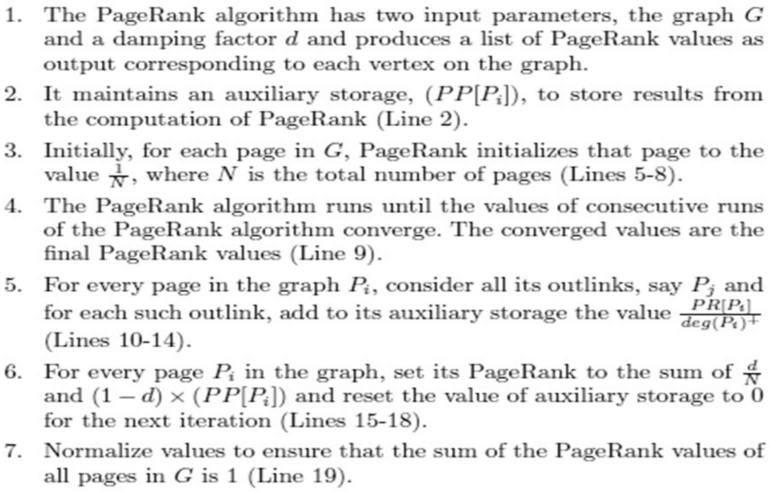
PageRank Algorithm



```
1  Algorithm: PageRank calculation of a single graph
   Input: $G$—Directed graph of $N$ web pages
   $d$—Damping factor
   Output: $PR[1 \dots N]$, where $PR[P_i]$ is the PageRank of page $P_i$
2  Let $PP[1 \dots N]$ denote a spare array of size $N$
3  Let $d$ denote the probability of reaching a particular node by a random
   jump either from a vertex with no outlinks or with probability $(1 - d)$
4  Let $N(P_u)^+$ denote the set of pages with at least one outlink
5  foreach $P_i$ in $N$ pages of $G$ do
6      $PR[P_i] = \frac{1}{N}$
7      $PP[i] = 0$
8  end
9  while $PR$ not converging do
10     foreach $P_i$ in $N$ pages of $G$ do
11         foreach $P_j$ in $N(P_i)^+$ do
12             $PP[P_j] = PP[P_j] + \frac{PR[P_i]}{deg(P_i)^+}$
13         end
14     end
15     foreach $P_i$ in $N$ pages of $G$ do
16         $PR[P_i] = \frac{d}{N} + (1 - d)(PP[P_i])$
17         $PP[P_i] = 0$
18     end
19     Normalize $PR[P_i]$ so that $\sum_{P_i \in N} PR[P_i] = 1$
20 end
```

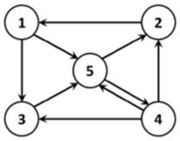
Algorithm 1: PageRank calculation of a single graph

PageRank Algorithm



1. The PageRank algorithm has two input parameters, the graph G and a damping factor d and produces a list of PageRank values as output corresponding to each vertex on the graph.
2. It maintains an auxiliary storage, $(PP[P_i])$, to store results from the computation of PageRank (Line 2).
3. Initially, for each page in G, PageRank initializes that page to the value $\frac{1}{N}$, where N is the total number of pages (Lines 5-8).
4. The PageRank algorithm runs until the values of consecutive runs of the PageRank algorithm converge. The converged values are the final PageRank values (Line 9).
5. For every page in the graph P_i, consider all its outlinks, say P_j and for each such outlink, add to its auxiliary storage the value $\frac{PR[P_i]}{deg(P_i)^+}$ (Lines 10-14).
6. For every page P_i in the graph, set its PageRank to the sum of $\frac{d}{N}$ and $(1 - d) \times (PP[P_i])$ and reset the value of auxiliary storage to 0 for the next iteration (Lines 15-18).
7. Normalize values to ensure that the sum of the PageRank values of all pages in G is 1 (Line 19).

Example



$$PR(P_u) = \sum \frac{PR(P_v)}{deg(P_v)^+}$$

$$PR(P) = (1-d) + d(\frac{PR(P_1)}{deg(P_1)^+} + \frac{PR(P_2)}{deg(P_2)^+} + \ldots + \frac{PR(P_n)}{deg(P_n)^+})$$

$$PR(1) = \frac{PR(2)}{1}$$

$$PR(2) = \frac{PR(4)}{3} + \frac{PR(5)}{2}$$

$$PR(3) = \frac{PR(1)}{2} + \frac{PR(4)}{3}$$

$$PR(4) = \frac{PR(5)}{2}$$

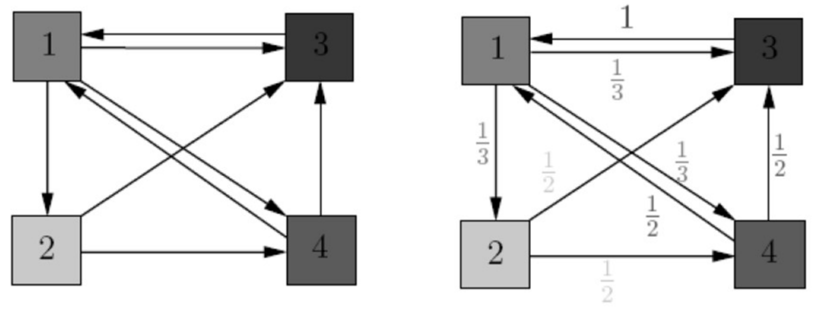
$$PR(5) = \frac{PR(1)}{2} + \frac{PR(3)}{1} + \frac{PR(4)}{3}$$

$$\mathbf{x} = \begin{bmatrix} PR(1) & PR(2) & PR(3) & PR(4) & PR(5) \end{bmatrix}^T$$

PageRank via Matrix Algebra

- We can represent a web-graph in a form of a matrix.

Example



Example

- In Matrix form

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

Example

A . V = $\begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$ $\cdot \mathbf{v} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}$

$$\mathbf{Av} = \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix}$$

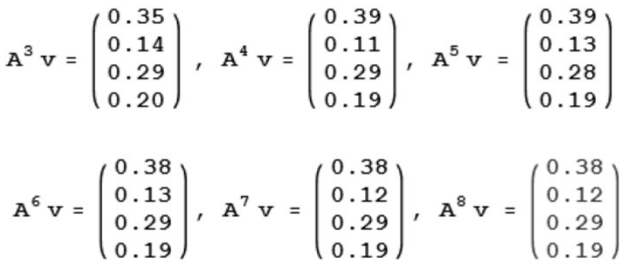
Example

A2 . V = A . A V = $\begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$ $\mathbf{Av} = \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix}$

$$\mathbf{A}^2\,\mathbf{v} = \begin{pmatrix} 0.43 \\ 0.12 \\ 0.27 \\ 0.16 \end{pmatrix}$$

$$\mathbf{A}^3\,\mathbf{v} = \begin{pmatrix} 0.35 \\ 0.14 \\ 0.29 \\ 0.20 \end{pmatrix},\quad \mathbf{A}^4\,\mathbf{v} = \begin{pmatrix} 0.39 \\ 0.11 \\ 0.29 \\ 0.19 \end{pmatrix},\quad \mathbf{A}^5\,\mathbf{v} = \begin{pmatrix} 0.39 \\ 0.13 \\ 0.28 \\ 0.19 \end{pmatrix}$$

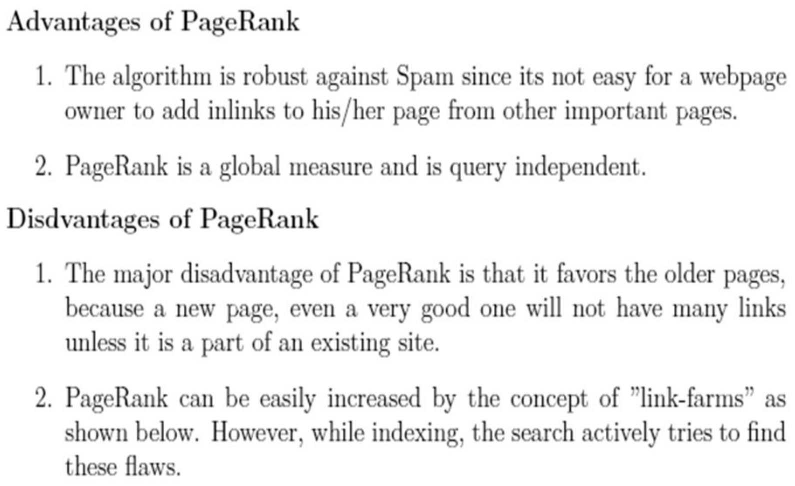
Example



$$A^3 v = \begin{pmatrix} 0.35 \\ 0.14 \\ 0.29 \\ 0.20 \end{pmatrix}, \quad A^4 v = \begin{pmatrix} 0.39 \\ 0.11 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad A^5 v = \begin{pmatrix} 0.39 \\ 0.13 \\ 0.28 \\ 0.19 \end{pmatrix}$$

$$A^6 v = \begin{pmatrix} 0.38 \\ 0.13 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad A^7 v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad A^8 v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}$$

Advantage/Disadvantage PageRank



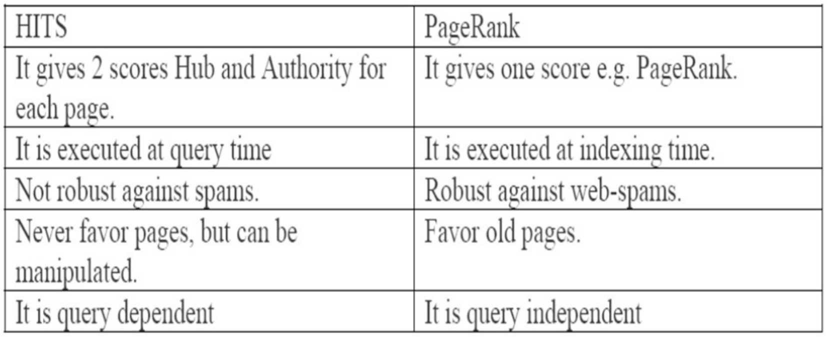
Advantages of PageRank

1. The algorithm is robust against Spam since its not easy for a webpage owner to add inlinks to his/her page from other important pages.
2. PageRank is a global measure and is query independent.

Disdvantages of PageRank

1. The major disadvantage of PageRank is that it favors the older pages, because a new page, even a very good one will not have many links unless it is a part of an existing site.
2. PageRank can be easily increased by the concept of "link-farms" as shown below. However, while indexing, the search actively tries to find these flaws.

HITS vs. PageRank



HITS	PageRank
It gives 2 scores Hub and Authority for each page.	It gives one score e.g. PageRank.
It is executed at query time	It is executed at indexing time.
Not robust against spams.	Robust against web-spams.
Never favor pages, but can be manipulated.	Favor old pages.
It is query dependent	It is query independent