Google PageRank

James Della-Giustina, Craig Riley, Kayla Scott, Karishma Ale Magar

October 20, 2021

Introduction

What is Google PageRank?

First and foremost, you can't see PageRank when you search, but behind the scenes PageRank is a signal system for ranking web pages. It combines the mathematics of graph theory, probability, and linear algebra to obtain a list of ranks for a given set of webpages.

How is it Helpful?

PageRank is essentially an iterative searching algorithm that measures the relative importance of a given set of webpages based on weights assigned to various links.

Where is it Now?

Initially PageRank relied on individuals posting links on websites to help determine which sites offered content of value and now there are over 200 Google ranking factors and signals today.

History & Context

Google's founders Larry Page and Sergey Brin formed Google behind Brin's idea that information on the web could be ranked based upon a page's link popularity, that the more links point to a page, the higher it ranks [1]. On September 1, 1998 the first PageRank patent was filed making it the backbone behind Google search results. The Google PageRank algorithm was a turning point for Google and for the way in which search engines operated.[5]

Graph Theory

- A **Directed Graph** G is a triple consisting of a vertex set $V(G)$, a directed edge set $E(G)$, and an adjacency relation that associates each edge with two vertices [3].
    - The elements of $V(G)$ are called **vertices** of G.
    - The elements of $E(G)$ are called **directed edges** of G.
    - A vertex v is **adjacent** to vertex u if they are joined by an edge.
    - Each edge in $E(G)$ provides an **adjacency relation** between two vertices in $V(G)$.
- The **in/out-degree** of a vertex v is the number of directed edges going in to or out of a vertex respectively.
- A **simple** graph is a graph where every edge shares two distinct vertices (no loops), and every pair of vertices shares at *most* one edge.

Graph Theory

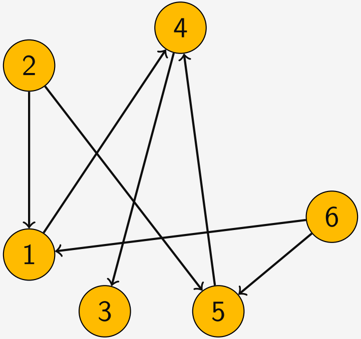

Figure 1: A simple directed graph G with 6 vertices and 7 edges.

Probability

Definition
Classical definition of probability - Let S be the sample space of an experiment. If S has N points that are equally likely to occur, then for any event A of S:

$$P(A) = \frac{N(A)}{N}$$

where $N(A)$ is the number of points of A [2].

Using this definition, we initialize the page rank of each page to be $\frac{1}{N}$ where N denotes all web-pages currently being considered on the internet.

Damping Factor d

We introduce a damping factor $d = .85$ to account for the fact that a user will not continue to click links forever and will at some point stop. However the damping factor also takes care of a variety of other problems:

- Sinks/Dangling Nodes- A page that contains no outlinks.
- Spider Traps - A set of pages that contains no outbound links from that particular set. [4].
- The possibility that a user may go from one page to another where there exists no link between the two, known as 'teleporting'.
- The equilibrium value of $\overrightarrow{0}$ that we know exists from our predator-predator model.
- An equilibrium value where one page has a rank of 1 and every other page has a rank of 0.

Linear Systems/Algebra

We can construct a matrix to represent the adjacency of pages $p_1, p_2, \ldots, p_N$ where each column represents the outbound links and each row represents the inbound links. This is a **stochastic** matrix, in which the sum of every column is equal to 1 which is also a **Markov** matrix.

Definition
Markov Matrix $\mathcal{M}$ - All $m_{ij} \geq 0$ and each column sum is 1 with a largest eigenvalue $\lambda = 1$. If $m_{ij} > 0$, the columns of $\mathcal{M}^k$ approach the steady-state eigenvector $\mathcal{M}s = s > 0$ [6].

Constructing the Adjacency Matrix A

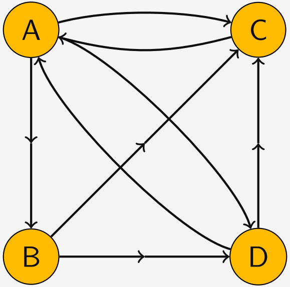

Figure 2: Digraph corresponding to 4 websites.

Constructing the Adjacency Matrix A

Our initial adjacency matrix A and PageRank vector R is given by:

$$A = \begin{pmatrix} 0 & 0 & 1 & .5 \\ .33 & 0 & 0 & 0 \\ .33 & .5 & 0 & .5 \\ .33 & .5 & 0 & 0 \end{pmatrix}; R = \begin{pmatrix} .25 \\ .25 \\ .25 \\ .25 \end{pmatrix}$$

Implementing the Damping Factor

We need to introduce this damping factor d. A typical choice is $d = .85$, and so we set up our system as:

$$A' = dA + \frac{1-d}{N}\,[1] \tag{1}$$

$$R_{t+1} = A' R_t \tag{2}$$

where R_t represents our PageRank vector, A' is our adjusted adjacency matrix, and [1] represents a matrix solely populated by 1's.

Computation

Now we iterate through Equation 2 until $|R_{i+1} - R_i|$ is 'small'. Note that we use the notation R_i for the i^{th} iteration. Our adjusted adjacency matrix:

$$A' = \begin{pmatrix} 0.0375 & 0.0375 & 0.8875 & 0.4625 \\ 0.3208 & 0.0375 & 0.0375 & 0.0375 \\ 0.3208 & 0.4625 & 0.0375 & 0.4625 \\ 0.3208 & 0.4625 & 0.0375 & 0.0375 \end{pmatrix} R_1 = \begin{pmatrix} .25 \\ .25 \\ .25 \\ .25 \end{pmatrix}$$

Computation

Our sequence of new page rank vectors R_i is as follows:

$$R_2 = \begin{pmatrix} 0.2185 \\ 0.1370 \\ 0.3398 \\ 0.3047 \end{pmatrix} R_3 = \begin{pmatrix} 0.4558 \\ 0.0994 \\ 0.2871 \\ 0.1576 \end{pmatrix} R_4 = \begin{pmatrix} 0.3486 \\ 0.1667 \\ 0.2759 \\ 0.2089 \end{pmatrix} R_5 = \begin{pmatrix} 0.3608 \\ 0.1363 \\ 0.2959 \\ 0.2071 \end{pmatrix} R_6 = \begin{pmatrix} 0.3770 \\ 0.1397 \\ 0.2856 \\ 0.1976 \end{pmatrix}$$

$$\mathrm{R}_7 = \begin{pmatrix} 0.3643 \\ 0.1443 \\ 0.2877 \\ 0.2037 \end{pmatrix} R_8 = \begin{pmatrix} 0.3686 \\ 0.1407 \\ 0.2886 \\ 0.2021 \end{pmatrix} R_9 = \begin{pmatrix} 0.3687 \\ 0.1419 \\ 0.2876 \\ 0.2017 \end{pmatrix} R_{10} = \begin{pmatrix} 0.3677 \\ 0.1420 \\ 0.2880 \\ 0.2023 \end{pmatrix} R_{11} = \begin{pmatrix} 0.3683 \\ 0.1417 \\ 0.2880 \\ 0.2020 \end{pmatrix}$$

Results

Therefore we can see that after roughly 10 iterations we have arrived at a final page ranking of websites A, B, C, and D as:

$$R_{11} = \begin{pmatrix} 0.3683 \\ 0.1417 \\ 0.2880 \\ 0.2020 \end{pmatrix}$$

Improving a Site’s PageRank

- High quality websites.
- Link partnerships with existing more established websites.
- Thorough knowledge and understanding of your marketplace and industry.
- Avoiding dishonest and illegal methods such as ’cloaking’.

Conclusion

- PageRank is a global ranking of all web pages based on their locations in the web graph structure.
- PageRank uses the structure of the web graph, Markov chains and linear algebra to make it one of the most effective and popular search engines on the internet.
- PageRank tends to favor older pages because newer pages simply do not have as many links.
- PageRank still matters for ranking today since it helps the search engine determine the most relevant result for particular query.

Acknowledgements

We want to thank Dr. Nganguia and Dr. Gluck for taking the time to walk through and reason out portions of the math with us.

References

[1] Amrani Amine. *PageRank Algorithm, Fully Explained*. 2020. URL: https://towardsdatascience.com/pagerank-algorithm-fully-explained-dc794184b4af.

[2] Saeed Ghahramani. *Fundamentals of Probability, with Stochastic Processes, 3rd Edition*. 3rd ed. Prentice Hall, 2004. ISBN: 0131453408,9780131453401.

[3] Yellen J. Gross J.L. and Zhang P. (eds.) *Handbook of Graph Theory*. 2ed. CRC, 2014. ISBN: 9781439880180.

[4] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of Massive Datasets*. 2nd ed. Cambridge University Press, 2014. DOI: 10.1017/CBO9781139924801.

[5] Larry Page et al. *The PageRank Citation Ranking: Bringing Order to the Web*. 1998.

[6] Gilbert Strang. *Linear Algebra and Its Applications*. 4th. Brooks Cole, 2005. ISBN: 0030105676,9780030105678.

Questions?