Google PageRank (and Hub-and-Authority) Algorithm

Ngoc Duong

December 2018

1 Introduction

The PageRank Algorithm was invented by Larry Page and Sergey Brin as graduate students at Stanford, and it became a Google trademark in 1998. The PageRank algorithm utilizes the idea that the importance of a certain web page can be determined by looking at the importance of the pages that link to it. If there are many pages that link to a webpage i, then i might be important, if webpage i is backed by another webpage with high 'importance weight', it might also be relevant to the search. This method employs concepts in graph theory, where webpages are represented by nodes and links between them are represented by edges.

2 Relevant concepts

Definition 1. Random walk is a process on graphs in which we move from one node to another along the edges in a random fashion, meaning each neighbor node has the same probability of being selected and the subsequent selections of these neighbor nodes are independent of the previous selections.

Definition 2. Dangling node is a node with no edges. In random walks, once we encounter a dangling node while surfing, i.e. a page with no links, we will continue to a random page by pretending there are links to all the other pages, and so probability of selecting any of the other pages is equal.

Definition 3. The "indegree" of node i is the number of edges for which i is a head (i.e. inbound links)
The "outdegree" of node i is the number of edges for which i is a tail (i.e. outbound links)
Notice that there can be no more than two edges between any two nodes.

Definition 4. We say that two nodes i and j of a directed graph are adjacent if there is an edge between i and j.
An adjacency matrix A of a directed graph with n nodes is constructed as as follows:
If there is an edge from node i to node j, then we put 1 as the entry on row i, column j of the matrix A.
Ex: $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$ Visualizing the graph represented by this adjacency matrix:
There is an edge from node 1 and 2 to node 3 (node 2 indegree = 2), an edge from node 4 to node 1, and an edge from node 3 to node 4.

Definition 5. A matrix A $\in R^{n,n}$is said to be reducible if there exists a permutation matrix P such that $PAP^T = \begin{bmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} \\ 0 & \mathrm{A}_{22} \end{bmatrix}$, where $A_{11} \in R^{r,r}$, $A_{22} \in \mathrm{R}^{n-r,n-r}$ and $A_{12} \in \mathrm{R}^{r,n-r}$, 0 $<r<n$.
A matrix is irreducible if it is not reducible

Definition 6. A matrix is called primitive if there is a positive integer k such that A^k is a positive matrix. An irreducible nonnegative matrix A is said to be primitive, if the only

eigenvalue of A of absolute value (A) is (A).
k is the number of edges that need to be walked along to go from point i to j (assuming there is a path). A^k is matrix A multiplied with itself k times

Definition 7. A stochastic matrix satisfies two following properties:
(i) All entries are non-negative
(ii) All entries in each column add up to one

3 Setting up

We will assign to each web page P a measure of its importance I(P), called the page's PageRank. If one of the n links coming from page i points to page j, then page i will pass on 1/n of its importance to page j. More specifically, There are 4 edges (links) coming from node i, one of which connects with node j, then node j will receive the importance of node i.

P has some special properties:
(i) Its entries are all non-negative
(ii) The sum of the entries in a column is one unless the page corresponding to that column has no links (the presence of dangling nodes)

We can derive a matrix A that takes into account these dangling nodes by assigning probability 1/(n-1) as entries of all columns that correspond to a dangling node (where n is total number of pages/nodes) This allows for a modification to the H matrix (by adding H and A). We define $S = A + H$

Coming back to measure of importance I(P), the metrics for importance of page i includes the probability of visiting page i at any given time and the frequency of visiting page i over time. The random walk can start at any random page in the web, then a random walker follows an outgoing link (guided by S) with p = α and chooses another random page with p = 1- α.
The PageRank Equation is a random walk on a graph with stochastic matrix
$R = \alpha\ S + \dfrac{1-\alpha}{\text{n}} 1$

where 1 is the matrix with all rows equal to 1. R is stochastic because it is a combination of S and 1/n, both of which are stochastic

Notice when $\alpha = 1$, then $G = S$ (the original matrix of the web). Also the rate of convergence of the power method depends on the modulus of the second eigenvalue Λ_2 (this will be re-visited later), and it has been proven that $\Lambda_2 = \alpha$, so the larger the α the slower the convergence rate is. Larry Page and Sergei Brin chose this value to be 0.85 (*)

The Power method

Easy case: all the eigenvalues are distinct and have magnitude smaller than 1 except the first (largest) one How does it work? We begin by choosing a vector Io as a candidate for I and then producing a sequence of vectors I^k by $I^{k+1} = H\ I^k$

Under the best of circumstances (primitive and irreducible original adjacency matrix), the other eigenvalues of S will have a magnitude smaller than one

$1 = |\Lambda_1| > |\Lambda_2| > |\Lambda_3| > \cdots > |\Lambda_n|$

We will also assume that there is a basis $v_1,\ v_2,\ \cdots,\ v_j$ of eigenvectors for S with corresponding eigenvalues $|\Lambda_j|$

(A glimpse at the result) Given a nice primitive stochastic matrix A, we have its spectral radius = 1, and it follows that the powers of A converge to a positive matrix of dim nx1, and that the rate of convergence is based on the magnitude of the eigenvalues other than 1, specifically $|\Lambda_2|$.

4 Process and Results

We want to show that: The sequence I^1, I^2,$\cdots$,I^k will converge to some stationary vector I and the convergence is independent of the base vector I_o

(Before the math) From the probabilistic/stochastic process perspective, we can imagine a group of people setting out into a maze– which is of an extremely big dimension–at the same starting point. Recall that random walks means these people just lose all their previous memories as soon as they take a turn, and due to this huge amount of inherent randomness in the process, at n (some large number) turns later, they will just forget their original starting point and we can see/claim that the eventual distribution of these maze-walkers throughout the maze is independent on where they started. The significance of a random walk is that it can also guarantee irreducibility of the matrix S, and controls/guarantees convergence of I. In addition, the probability that an intersection i (in this case of a maze, is a node) is visited after a turn is equal to Ax, and the probability of that intersection being visited after k turns is A^kx. The sequence Ax, A^2x, A^3x, $\cdots$,A^{k}x, $\cdots$ converges to a single stochastic column vector v, with entries equal to probabilities of each intersection being visited.

Proof: By the Perron-Frobenius Theorem (which will be mentioned later), we can recall $1 = |\Lambda_1| > |\Lambda_2| > |\Lambda_3| > \cdots > |\Lambda_n|$

We will also assume that there is a basis v_1, v_2, $\cdots$, v_j of eigenvectors for S with corresponding eigenvalues $|\Lambda_j|$

We have $I^o = a_1v_1 + a_2v_2 + \cdots + a_nv_n$

$I^1 = a_1v_1 + a_2\Lambda_2\ v_2 + \cdots + a_n\Lambda_n\ v_n$

$I^2 = a_1v_1 + a_2\Lambda_2^2\ v_2 + \cdots + a_n\Lambda_n^2\ v_n$

$\ldots$

$I^k = a_1v_1 + a_2\Lambda_2^k\ v_2 + \cdots + a_n\Lambda_n^k\ v_n$

By the Perron-Frobenius Theorem (and assuming the favorable structure/nature of the matrix at hand), we see that since largest eigenvalue $|\Lambda_1| = 1$, as k gets infinitely large, $|\Lambda_1|$ dominates to reserve the output – largest eigenvector v. At the same time, the next entity in line involving taking power of $|\Lambda_2|$ shows that the convergence rate depends on this second largest eigenvalue $|\Lambda_2|$ (this also serves to corroborate the previous note (*) of Page and Brin's choice of the Damping Factor α as 0.85).

In addition, working with primitive (and thus irreducible) matrices also avoids the situation when a graph/web contains a smaller web within it, and thus all the importance weights will eventually be assigned to these sub-web pages (so that only the weights corresponding to these pages add up to one), even though other pages outside this sub-web also have links going to and from them. To guarantee that $|\Lambda_2| <1$, we need the matrix S to be primitive. When it is not primitive, there can be more than 1 eigenvalue whose magnitude ≥ 1, and those eigenvalues are not algebraically distinct, in which case it may be helpful to determine the Jordan Canonical Form. However, the best policy then is to find/modify the original matrix to a matrix that is primitive (and irreducible).

Computing I

Recall the modified stochastic matrix S (from the adjacency matrix and taking into account dangling nodes) has the form

$S = A + H$

after applying the PageRank Equation and random process, we have

$R\ I^k = \alpha S I^k + \dfrac{1-\alpha}{\text{n}} 1\text{I}^k$

$R\ I^k = \alpha A I^k + \alpha H I^k + \dfrac{1-\alpha}{\text{n}} 1\text{I}^k$

The result of interest can be obtained by evaluating each components of the sum on the RHS

Another way to think about the convergence

Let D be a diagonal matrix, whose diagonal entries are the eigenvalues i D = $\begin{bmatrix} \lambda_1(=1) & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$

Let v_1, v_2, $\cdots$,v$_n$ be corresponding (column) eigenvectors
Let P = $\begin{bmatrix} \mathrm{v}_1 & v_2 & \cdots & v_n \end{bmatrix}$. Since the eigenvectors are linearly independent, P is invertible. Let v_o be the baseline pagerank vector, like I_o Given M is the original (very large) matrix, it can be diagonalized into PDP^{-1} Diagonalization of M gives a nice diagonal matrix D so we can iterate k times for convergence.

We have $M^k v_0 = (PD^k P^1)v_0$, and since all λ_i's except λ_1 is 1, D^k converges to $\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix}$
and thus the original pagerank vector v_0 will converge to $\begin{bmatrix} \mathrm{v}_1 & 0 & \cdots & 0 \end{bmatrix} P^{-1} v_0$, which is a scalar multiple of v_1

Here again we see v_1 dominates in determining the converged stochastic column vector.

Results

The final result is the estimated PageRank stationary vector I (after convergence) – the stationary distribution of the Google matrix, a stochastic and irreducible matrix of dim nx1 listing the importance of each page (out of n pages)

5 Theorem

The Perron-Frobenius Theorem

The Perron-Frobenius theorem gives the underlying conditions for the nature/structure of the matrix and eigenvalues so that they work out well in the convergence problem.

The Perron-Frobenius theorem says that

If A is a primitive matrix, then:
(i) The largest eigenvalue $|\lambda_{max}|$ of A is positive and of multiplicity 1
(ii) $|\lambda_{max}|$ is larger than the absolute values of any other eigenvalues of A
(iii) The corresponding eigenvector of $|\lambda_{max}|$ has all entries >0

Proof

(i) Defining g(v): For any nonzero v 0, let g(v) be the largest $|\lambda|$ such that $Av \geq |\lambda| v$
Let α = maxg(v) : v $\geq$ 0, v $\neq$ 0

It can be proven that for any nonzero vector $v \geq 0$, there is a vector $w = v/||v|| \geq 0$ with norm 1 such that g(w) = g($w/||w||$) = g(v), then there is a vector $v \geq 0$ with norm 1 such that g(v) = α, or α (= $|\lambda_{max}|$) >0 is an eigenvalue of A and by definition is the largest.

Let f(Λ) = det(ΛI − A) be the characteristic polynomial of A We have f (Λ) = (Λ − $\alpha)^m q(\Lambda)$ where q is a nonzero polynomial whose roots have absolute value less than α, and $m \in N$ is the multiplicity of α

$f'(\alpha)$= $m(\alpha - \alpha)^{m1}$ q(α) + (α - α)^{m}q$'$(α) = 0

It can actually be proven that $f'(\alpha)$ is actually >0 (proof is longer so I am not including) =><=

Thus no other λ can be equal to $\alpha = \Lambda_m ax$ itself, which implies multiplicity 1 of this largest eigenvalue

(ii) Let u be a eigenvector that corresponds to the eigenvalue Λ. We have $Au = \Lambda u$. Then $|u| \geq 0$ is nonzero.

Furthermore, $A|u| \geq |Au| = |\lambda u| = |\lambda||u|$. Therefore, $|\lambda| \lesssim g(|u|) \lesssim \alpha$ (**).

Suppose $|\lambda| = \alpha$, then $A|u| \geq |Au| = \alpha|u| = A|u|$

Therefore, $A|u| = |Au|$

Also $u = f(\phi)w$ (expressing u in terms of some other vector $w \geq 0$)

Hence, $Aw = \lambda w$, so it must be that $\Lambda \geq 0$ and $\Lambda = \alpha$. This contradicts with (**)

We obtain $|\lambda_i| < \alpha$

(iii) We have $Av > 0$

then $A(Av) = A(v) = \alpha(Av)$

then $g(Av) = \alpha$

Let $w = Av/||Av||$, then $w > 0$, $||w|| = 1$

We also have $g(w) = g(Av/||Av||) = g(Av) = \alpha$

But v is the unique vector satisfying these properties, so $v = w > 0$.

6 The Hub-and-Authority Algorithm

The Hub-and-Authority Algorithm was developed by Jon Kleinberg, Professor of Computer Science at Cornell University. Also referred to as HITS (hyperlink-induced topic search), this algorithm employs a good amount of concepts used in the PageRank Algorithm. There are, however, a couple additional concepts.

Authority: As a query is inputted, an authority is a webpage that contains valuable information on the subject (determined by the query).

Hubs are webpages that are helpful in finding the authority (through endorsement, providing useful links towards authoritative pages, etc.). Hubs help with shaping the path to the important and relevant page in the "right direction".

For each page, a hub weight and an authority weight are assigned, and these weights are evaluated recursively.

Process and Results Unlike the PageRank algorithm, which let a random walker surfs around the entire world wide web, the HITS algorithm takes a subset of webpages, called **root R**, ideally having many quality hubs and authorities among them. This subset might be very disconnected (nodes far away from each other with few/no links) and so PageRank is not effective here.

From this root sub-graph, we can extend to a **seed S** sub-graph by branching out through links to and fro to take in new nodes (including hubs and authorities). It might also be helpful to note that authorities of interest might take inbound links from many hubs in this S sub-graph, and vice versa, high quality hubs usually point to the homogeneous/same authorities.

Recursively defining authority score a_i and hub score h_i for a page i:

$$\sum_{j \in S, i \leftrightarrow j} a_j = h_i$$

$$\sum_{j \in S, j \leftrightarrow i} a_i = h_j$$

A be the adjacency matrix of sub-graph S. We have the relationships:

$a = A^T h$ and $h = Aa$ After k iterations, $a_k = (A^T A)a_{k-1}$ and $h_k = (AA^T)h_{k-1}$, so a becomes the fixed-point of $A^T A$ and h is the fixed point of AA^T

Apply the power iteration method, by the Perron-Frobenius Theorem (assuming matrices $A^T A$ and AA^T are primitive), the mechanism from here works the same as in PageRank method. We care about the assumption that the matrices AA^T and $A^T A$ are real and symmetric, because we can work nicely with only real eigenvalues.

Eventually, we still get the largest eigenvalue $lambda_1$ dominates and that the convergence rate depends on $lambda_2$

7 Hub-and-Authority vs PageRank

HITS algorithm and PageRank algorithm both function with the same spirit. They both utilize the link structure and the graph representation of the webpage in order to evaluate the importance and relevance of the pages. The difference is that unlike the PageRank algorithm, which applies probabilistic modeling/Markov Chain on the entirety of the web graph, HITS only works on a small sub-graph (the seed SQ) from the web graph, and this sub-graph (nodes as authorities and hubs) change according to different query search. In short, HITS is query dependent, which means ranking a subset of pages extracted for a specific query, whereas PageRank is query independent, which means ranking the whole graph in advance

8 End notes

I wanted to but failed to be more comprehensive and deep due to the constraints in time, energy, and background knowledge in graph theory/numerical methods. The mathematical underpinnings of concepts and mechanisms of the real thing are extensive and deep at the same time. Thus, naturally there must be holes/misinformation in my understanding and my work above. *To classmates* If you are interested, more readings on your part will probably help connect the dots better.

[1] Tanase, R., Radu, R., "HITS Algorithm - Hubs and Authorities on the Internet", retrieved at http://pi.math.cornell.edu/ mec/Winter2009/RalucaRemus/Lecture4/lecture4.html
[2] Austin, D., 2008. "How Google Finds Your Needle in the Web's Haystack", retrieved at https://math.dartmouth.edu/archive/m22s14/publichtml/PageRank1.pdf
[3] Shum, K., 2013. "Notes on PageRank Algorithm", retrieved at http://home.ie.cuhk.edu.hk/ wkshum/papers/pagerank.pdf
[4] Tanase, R., Radu, R., "PageRank Algorithm - The Mathematics of Google Search", retrieved at http://pi.math.cornell.edu/ mec/Winter2009/RalucaRemus/Lecture3/lecture3.html
[5] Shahar, D., 2014. "Proof of the Perron-Frobenius Theorem", retrieved athttps://math.arizona.edu/ dshahar/PerronFrobeniusTheorem.pdf