GOOGLING IT: HOW GOOGLE RANKS SEARCH RESULTS

Courtney R. Gibbons

October 17, 2017

Hamilton

Definition (Relevance)

(noun): the quality or state of being closely connected or appropriate: “this film has contemporary relevance” or “the quantity, quality, and relevance of links count towards your rating” or “the Web does allow us to produce more articles of relevance to our readers.” *(Paraphrased from the OED)*

Google’s pagerank algorithm calculates the ***relevance*** of a website based on the connectedness of the internet and how that page is connected within the internet.

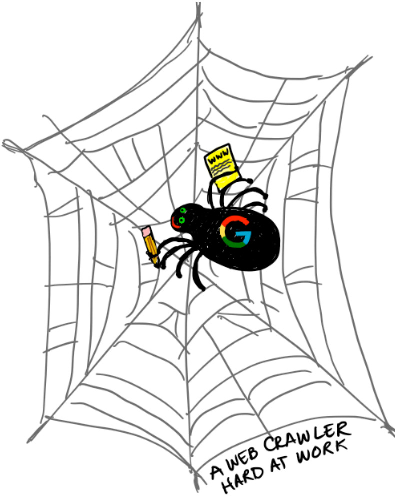
I want to tell you a little bit about how that works mathematically.

BUT FIRST!

The internet has 1.27 **billion** active webpages and almost 300 billion archived webpages (via `www.internetlivestats.com`)!

Homework. Develop a model that fits the yearly data for 1991 (the start of the world wide web, with 1 webpage) through 2017.

Google understands how the web is connected (and knows what's on each webpage) using web crawlers.



WHAT IS PAGERANK, AND WHAT DOES IT DO?

First, the pagerank takes the form of a vector

$$\mathbf{z} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{1,270,000,000} \end{bmatrix}$$

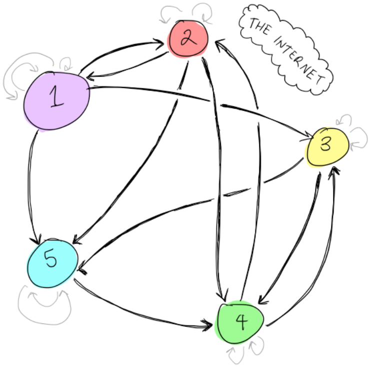
where p_n is the probability that a web surfer would go to page number n on the internet.

Example. A web surfer wants to find information about "Courtney Gibbons Math Professor"

Using webcrawlers and pagerank, Google:

1. Looks through its web cache for the text "Courtney Gibbons Math Professor"
2. Displays the pages with this text in order of their rankings in $\mathbf{z}$: the biggest p_n goes at the top, etc. (Well, under the advertisements.)

EXAMPLE: A VERY SMALL INTERNET



THE INTERNET :

		the page...				
		1	2	3	4	5
links to...	1	0	*	0	0	0
	2	*	0	0	*	0
	3	*	0	0	*	0
	4	0	*	*	0	*
	5	*	*	*	0	0

EXAMPLE: A VERY SMALL INTERNET

We have just constructed a ***weighted adjacency matrix***,

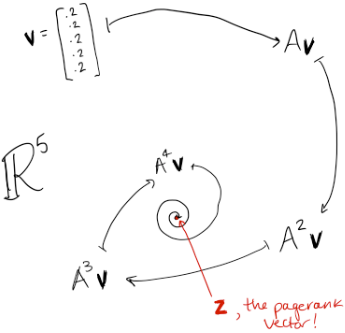
$$A = \begin{bmatrix} 0 & 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 1/2 & 0 \\ 0 & 1/3 & 1/2 & 0 & 1 \\ 1/3 & 1/3 & 1/2 & 0 & 0 \end{bmatrix}.$$

This matrix has a special property: it is ***column stochastic*** (every column adds up to 1).

A DASH OF MATHEMATICS: DYNAMICAL SYSTEMS

$$A : \mathbb{R}^5 \to \mathbb{R}^5$$
$$\mathbf{v} \mapsto A\mathbf{v}$$

For example, $A \begin{bmatrix} .2 \\ .2 \\ .2 \\ .2 \\ .2 \end{bmatrix} = \begin{bmatrix} .067 \\ .167 \\ .167 \\ .367 \\ .233 \end{bmatrix} \in \mathbb{R}^5.$



$\mathbf{v} = \begin{bmatrix} .2 \\ .2 \\ .2 \\ .2 \\ .2 \end{bmatrix}$
$A\mathbf{v}$
$\mathbb{R}^5$
$A^4\mathbf{v}$
$A^2\mathbf{v}$
$A^3\mathbf{v}$
$\mathbf{z}$, the pagerank vector!

$$A \begin{bmatrix} .067 \\ .167 \\ .167 \\ .367 \\ .233 \end{bmatrix} = A^2 \begin{bmatrix} .2 \\ .2 \\ .2 \\ .2 \\ .2 \end{bmatrix} = \begin{bmatrix} .056 \\ .206 \\ .206 \\ .372 \\ .161 \end{bmatrix}$$

Repeated applications of A ...

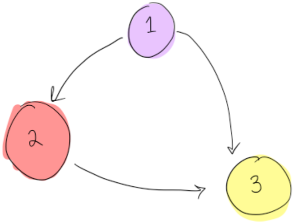
(M2)

We seem to be stabilizing to

$$\mathbf{z} \approx \begin{bmatrix} .066 \\ .198 \\ .198 \\ .352 \\ .187 \end{bmatrix}.$$

WHAT COULD GO WRONG?

What if $\mathbf{z} = \mathbf{0}$?



$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & 0 & 0 \\ 2 & \frac{1}{2} & 0 & 0 \\ 3 & \frac{1}{2} & 1 & 0 \end{array}$$

What's up? Well, for one, A isn't column stochastic anymore (even worse, A has a column of zeros).

This internet has weighted adjacency matrix $A = \begin{bmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{bmatrix}$.

In this case, $A^3 \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = \mathbf{0}$.

WHAT COULD GO WRONG?

We can fix this, however, by replacing the column of zeros:

$$A = \begin{bmatrix} 0 & 0 & 1/3 \\ 1/2 & 0 & 1/3 \\ 1/2 & 1 & 1/3 \end{bmatrix}.$$

(Now we're pretending Website 3 links everywhere instead of nowhere.)

Big Idea 1. We have just taken the internet matrix and modified it to be column stochastic so that our calculations are more meaningful in the real world. (M2)

Iteration is nice, but is there a shortcut for finding $\mathbf{z}$?

The pagerank vector $\mathbf{z}$ should satisfy the equation $A\mathbf{z} = \mathbf{z}$.

For those in the know, that means that 1 should be an eigenvalue of A and $\mathbf{z}$ should be its only associated eigenvector.

In terms of vector spaces, an eigenvector gives the direction of a 1-dimensional subspace of $\mathbb{R}^n$. We can make it ***normal*** by scaling it so its entries sum to 1. (M2)

For 5-site internet A, there's one normal eigenvector associated to 1:

$$\mathbf{z} \approx \begin{bmatrix} .066 \\ .198 \\ .198 \\ .352 \\ .187 \end{bmatrix}.$$

The eigenvalues of a matrix are related to the equation $A\mathbf{x} = \lambda\mathbf{x}$ (they're the zeros of the determinant of $A - \lambda I_n$).

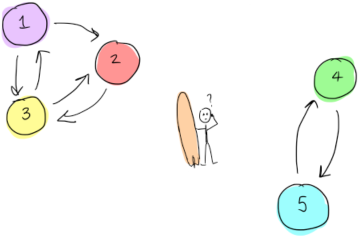
An eigenvector associated to an eigenvalue λ is a solution to the equation $A\mathbf{x} = \lambda\mathbf{x}$.

WHAT COULD GO WRONG THIS TIME?

Sometimes 1 isn't an eigenvalue:

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{bmatrix}; \ \det(A - \lambda I_3) = (-\lambda)^3 \implies \lambda = 0.$$

Okay, we already fixed that matrix. What else?



$$\begin{array}{c|ccc:cc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 2 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 3 & \frac{1}{2} & 1 & 0 & 0 & 0 \\ \hdashline 4 & 0 & 0 & 0 & 0 & 1 \\ 5 & 0 & 0 & 0 & 1 & 0 \end{array}$$

Sometimes 1 has too many associated eigenvectors.

Indeed, the matrix on the previous slide has two little submatrices,

$$A_{top} = \begin{bmatrix} 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \quad \text{and} \quad A_{bottom} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Each submatrix contributes an eigenvalue of 1, and they each have a different normal eigenvector. (M2)

Big Idea 2. If those zeros were instead tiny positive numbers, we would have a **positive** matrix, which would preclude this mini-matrix nonsense.

THE BJILLION DOLLAR IDEA

Define the $n \times n$ matrix $B = \frac{1}{n} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$.

Let p be a probability ($p \sim .15$).

The matrix $M = (1 - p)A + pB$ is now a ***positive, column stochastic matrix***.

NO MATH TALK IS COMPLETE WITHOUT A THEOREM OR TWO

Theorem (Perron-Frobenius)

If M is a positive, column stochastic matrix then the following good things happen:

1. *The matrix M has 1 as an eigenvalue of multiplicity one.*
2. *The largest eigenvalue of M is 1; all other eigenvalues have magnitude less than 1.*
3. *The normal eigenvector associated to the eigenvalue 1 has nonnegative entries.*

Theorem (Power Method Convergence)

Let M be a positive, column stochastic $n \times n$ matrix. Let $\mathbf{z}$ be its normal eigenvector corresponding to the eigenvalue 1. Let $\mathbf{v}$ be the vector with all entries equal to $1/n$. Then the sequence $(M^k\mathbf{v})_{k=1}^{\infty}$ converges to $\mathbf{z}$.

CALCULATING PAGERANK FOR REAL

The pagerank vector **z** exists.

We can approximate **z** very quickly using the Power Method (a riff on the Dynamical Systems approach) for enormous positive, column stochastic matrices.

♫ INTERMISSION ♫

SEO: SEARCH ENGINE OPTIMIZATION

Google's search result ranking algorithm is proprietary – I haven't told you the whole story because I don't know the whole story.

Why hide the algorithm?

- It's the money maker.
- It would be too easy to game.

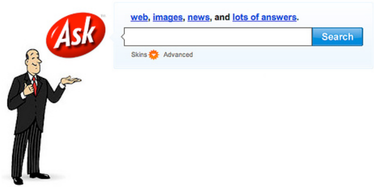
Exercise. With the people at your table, try to improve the pagerank of Website 1. You can make whatever modifications to the internet that you like **except** deleting edges that already exist.

Ways to become relevant:

- Make a bunch of new websites that link to your website only.
- Insert links back to your website in other websites (in the comments section, in an advertisement, via a sponsored article, etc.)

Good News! Google gets wise to these attempts to game and modifies the algorithm to adapt to attempts to game the rankings.

ARE THERE ALTERNATIVES?



Remember `Ask.com`? The algorithm behind that search engine used a different mathematical technique to assign ***authority*** rankings.

HITS (aka Hubs and Authorities) was developed by Jon Kleinberg at Cornell.

It also works by modifying the weighted adjacency matrix of the internet in order to calculate a ranking vector (but in a different way).

News! Google has adopted some of these ideas (e.g., giving more weight to `.edu` addresses).

More News! Google is also trying to measure ***trustworthiness*** by identifying “the correctness of factual information provided by the source” (using the internet itself to decide what a “fact” is!) [3]

WHAT ABOUT OTHER ALTERNATIVES?

There's always the card catalog.

Differences between Google (and other search engines) and a card catalog:

- Card catalogs are sorted by several criteria, including author, title, subject, publication date. They do not rank the cards within the card catalog. (That's the user's job.)
- Cards (virtual or physical) are created by human beings.
- The scholarly items in a card catalog are largely peer-reviewed or otherwise a product of an authority.
- Google uses the internet to understand how to rank pages on the internet. The card catalog uses external understanding to sort information; it isn't self-referential in the same way that Google is.

THINGS TO THINK ABOUT WHEN USING A SEARCH ENGINE

1. How can one improve the ranking of any website?
   - Is this ranking gameable?
   - Can “irrelevant” (resp. “unauthoritative”, “untrustworthy”) results be made to appear “relevant” (resp. “authoritative”, “trustworthy”)?
2. What does the rank actually measure?
   - Am I in danger of conflating “relevant” with “authoritative” with “trustworthy” results?
   - Am I missing out on important work that does not exist on the internet?
   - Are unpopular truths being suppressed by the ranking?/Are popular falsehoods being overrepresented?
   - Does the ranking lead me to believe that any “opinion” has equal (scholarly) worth?
3. Can I think critically about the value of the top hits?

A PARABLE

Math is a rare field: every result can be verified.

Vladimir Voevodsky (1966–2017) was a Fields Medalist in 2002 (for developing *homotopy theory for algebraic varieties* and for building *motivic cohomology*).

> *"**A technical argument by a trusted author, which is hard to check and looks similar to arguments known to be correct, is hardly ever checked in detail**.... Mathematical research currently relies on a complex system of mutual trust based on reputations. By the time Simpson's paper appeared, both Kapranov and I had strong reputations. Simpson's paper created doubts in our result, which led to it being unused by other researchers, but **no one came forward and challenged us on it**."*

–From *Univalent Foundations* (slideshow), Vladimir Voevodsky
Institute for Advanced Study, March 26, 2014
`http://www.math.ias.edu/vladimir/files/2014_IAS.pdf`

THE MORAL OF THE STORY

THANKS!

References.

1. Sergey Brin, Lawrence Page. The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, Vol. 33, pp. 107-17 (1998).
http://infolab.stanford.edu/pub/papers/google.pdf

2. Kurt Bryan, Tanya Leise. The $25,000,000,000 Eigenvector: The Linear Algebra behind Google, SIAM Review, Vol. 48, No. 3. (2006).
http://www.siam.org/journals/sirev/48-3/62328.html

3. Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, Wei Zhang. Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources, arXiv:1502.03519v1 https://arxiv.org/abs/1502.03519v1

4. Jon M. Kleignberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM (JACM), Vol. 46, No. 5 (1999). Pages 604-632
https://www.cs.cornell.edu/home/kleinber/auth.pdf

5. Raluca Tanase, Remus Radu. The Mathematics of Web Search (Cornell J-Term Course Notes).
http://www.math.cornell.edu/∼mec/Winter2009/RalucaRemus/

6. Internet Live Stats. http://www.internetlivestats.com/