

Motores de búsqueda Web con hipertextos

Algoritmo PageRank

Álgebra Superior I, Enero 2021

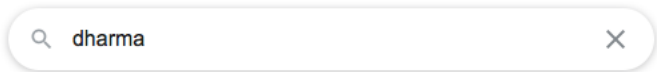




Buscar con Google

Me siento con suerte

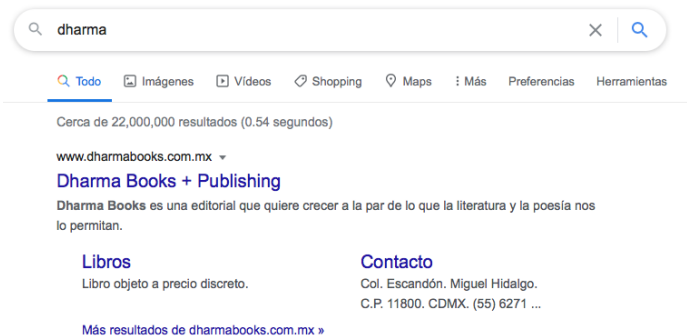
Imagina que entras al buscador Google y escribes la palabra "dharma" y unos segundos después obtendrás una lista con las páginas relevantes para tu búsqueda.



Ya que no tienes tiempo de pasearte por varias de estas páginas, esperas que el buscador ponga en primer lugar la mejor información, para que tu puedas verla primero.

Aquí tenemos dos nociones clave: **información importante** y **orden de importancia**.

Desde otro punto de vista, puedes pensar que administras un sitio web y deseas que tu sitio tenga muchas visitas. Te gustaría que tu sitio aparezca en los primeros lugares de una búsqueda en Google para generar **popularidad** y que la gente hable de ti.



A screenshot of a Google search interface. The search bar contains the word "dharma". Below the search bar, there are navigation tabs: "Todo" (selected), "Imágenes", "Vídeos", "Shopping", "Maps", "Más", "Preferencias", and "Herramientas". Below the tabs, it says "Cerca de 22,000,000 resultados (0.54 segundos)". The first search result is for "www.dharmabooks.com.mx" with a dropdown arrow. The title is "Dharma Books + Publishing" in blue. The description reads: "Dharma Books es una editorial que quiere crecer a la par de lo que la literatura y la poesía nos lo permitan." Below the description, there are two columns of links: "Libros" with the text "Libro objeto a precio discreto." and "Contacto" with the text "Col. Escandón. Miguel Hidalgo. C.P. 11800. CDMX. (55) 6271 ...". At the bottom of the search results, there is a link "Más resultados de dharmabooks.com.mx »".

Tenemos un problema

- Los usuarios quieren obtener la información [los sitios web] más relevante primero.
- Los sitios web quieren aparecer primero para obtener más visitas sin importar su **nivel de importancia**

A principios de los noventa los sitios que aparecían primero eran los que tenían más ocurrencias de las palabras clave que el usuario ingresaba al buscador. No fue un método muy efectivos para hallar **información relevante en los primeros puestos**.

El algoritmo PageRank, desarrollado en 1998, fue uno de los algoritmos más revolucionarios para calcular la relevancia de sitios web.

La idea es la siguiente

La importancia de un sitio web puede ser juzgada por la cantidad de referencias que hacen de él (mediante hipervínculos) en otras páginas (en su red).

PageRank (premisa)

- 1 Si una página i incluye un hipervínculo a una página j , significa que j es considerada importante para i .
- 2 Si hay varias páginas que incluyen un hipervínculo a una página j , significa que es comúnmente aceptado que j es importante.
- 3 Por otro lado, j podría sólo ser vinculada por una página k . Sin embargo, si el sitio k es autoritario, entonces "le transfiere" su importancia a j .

Lo fundamental es que

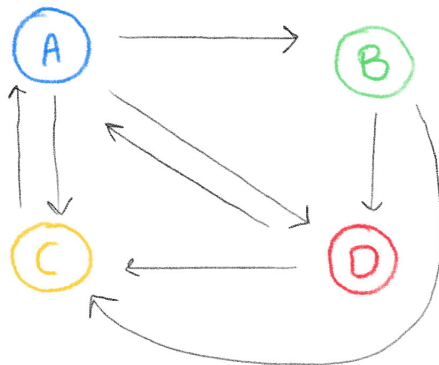
Podemos asignar un rango de importancia a un sitio, basado en el rango de las páginas que apuntan a él.

Consideremos el siguiente escenario. Hay 4 sitios de web en la misma red, digamos A , B , C y D y tienen la siguientes características:

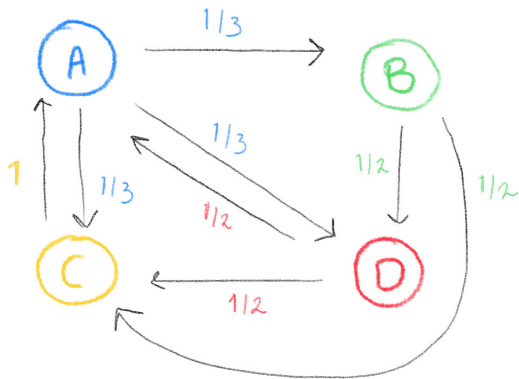
- 1 A hace referencia [tiene hipervínculos hacia] a B , C y D ,
- 2 B hace referencia C y D ,
- 3 C hace referencia sólo al sitio A ,
- 4 D hace referencia a A y a C .

PageRank: Modelo matemático

Traducimos la información al siguiente símbolo. Sólo nos importan las relaciones entre los sitios.



Ahora nos fijamos en en la importancia que "heredan" a otros sitios. Por ejemplo, el nodo A tiene tres flechas salientes, eso significa que pasa $\frac{1}{3}$ de su importancia a los nodos B , C y D .



Este diagrama tiene una matriz asociada, una matriz de popularidad

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

Que se obtiene de la siguiente manera:

$$c(X, Y) := \begin{cases} 0 & \text{si no hay flecha que sale de X a Y,} \\ \text{coeficiente de la flecha} & \text{si hay flecha que sale de X a Y} \end{cases}$$

	A	B	C	D
A	c(A,A)	c(B,A)	c(C,A)	c(D,A)
B	c(A,B)	c(B,B)	c(C,B)	c(D,B)
C	c(A,C)	c(B,C)	c(C,C)	c(D,C)
D	c(A,D)	c(B,D)	c(C,D)	c(D,D)

Denotemos por x_1 , x_2 , x_3 y x_4 la importancia de las páginas A , B , C , y D respectivamente. Para hallar esta importancia hay que resolver el sistema

$$\begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}. \quad (1)$$

Denotemos por

$$X := \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

Notemos que el sistema (1) es $AX = X$. Además,

$$AX = X \quad \text{si y sólo si} \quad (A - \text{Id}_{3 \times 3})X = 0.$$

Luego entonces, el sistema (1) es equivalente a

$$\begin{bmatrix} -1 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & -1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & -1 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (2)$$

Entonces, para determinar la popularidad de los sitios, basta resolver el sistema (2), el cual es un sistema homogéneo.

La forma escalonada reducida de la matriz asociada al sistema (2) es

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

Esto nos dice que

$$x_1 = 2x_4,$$

$$x_2 = \frac{2}{3}x_4,$$

$$x_3 = \frac{3}{2}x_4.$$

Por lo tanto, la solución X , no está determinada específicamente, pues vemos que es de la forma

$$X = \lambda \begin{bmatrix} 12 \\ 4 \\ 9 \\ 6 \end{bmatrix}, \quad \lambda \in \mathbb{R}.$$

Ahora bien, para cualquier $\lambda \in \mathbb{R}$, X es el vector de rango. Sin embargo, por razones técnicas se suele elegir el X donde la suma de todas sus coeficientes es 1, este es

$$X = \frac{1}{31} \begin{bmatrix} 12 \\ 4 \\ 9 \\ 6 \end{bmatrix} \sim \begin{bmatrix} 0,38 \\ 0,12 \\ 0,29 \\ 0,19 \end{bmatrix}$$

Vemos que

$$x_1 > x_3 > x_4 > x_2.$$

Esto significa que el buscador ordenará los sitios de la siguiente manera:

- 1 sitio A
- 2 sitio C
- 3 sitio D
- 4 sitio B

Ejercicio

Aplice el algoritmo de Page para clasificar los siguientes sitios de acuerdo con su popularidad: El sitio A refiere a los sitios B y C , el sitio B refiere a los sitios A y C , el sitio C sólo refiere al sitio A .

- [1] Page, Lawrence; Brin Sergey: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30 (1998) 107- 117
- [2] Tanase, Raluca; Radu, Remus: The Mathematics of Web Search <http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/index.html>.

Fin

... anterior respecto a 00 .

nte $\mu > 0$ tal que la forma bilineal $B[\cdot, \cdot]$ asociada a
ma de Lax-Milgram si $c(x) > -\mu$ para todo $x \in U$.

$u \in H_0^2(U)$ es una solución débil de la ecuación
ontera de Dirichlet

$$U, \quad u = \partial_\nu u = 0 \quad \text{sobre } \partial U \quad (1)$$

$f v \, dx$ para todo $v \in H_0^2(U)$.

1) tiene una única solución débil $u \in H_0^2(U)$.

na función $u \in H^1(U)$ es una solución débil del

Gracias