Análisis de Vínculos

Alondra Berzunza

Asignación de Relevancia (Page Rank)

Asignación de Relevancia (Page Rank)

También conocido como “Google Page Rank” se trata de una puntuación de 1 al 10 que Google le asigna a cada página web.

Desarrollado por Larry Page, cofundador de Google y Serguéi Brin en 1999.

Esta calificación influye bastante en el posicionamiento de cada sitio web dado que es uno de los factores que configuran el algoritmo en el motor de búsqueda.

Asignación de Relevancia (Page Rank)

Anteriormente, el Page Rank era público, por lo que todos los dueños de un sitio web sólo tenían que acceder a la barra de herramientas para conocer su calificación. Sin embargo, esta calificación se dividía en “Pública” y “Real”.

Asignación de Relevancia (Page Rank)

El Page Rank público era el que todos podían ver con un simple botón; sin embargo, Google sólo actualizaba la información dos veces al año, por lo que tampoco era muy confiable.

El Page Rank real es aquél que no hemos visto, la valoración en tiempo real de Google a nuestra página. La calificación pública es tan solo un aproximado de la real, la cual no veremos nunca, pero podemos darnos una idea de cúal podría ser dependiendo de nuestro tráfico, conversiones, etc.

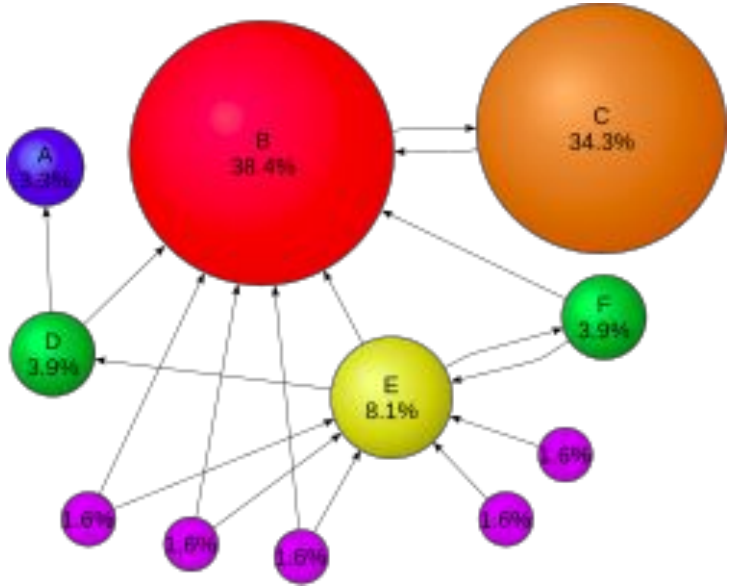
Asignación de Relevancia (Page Rank)

Entre los factores que toma en cuenta para aumentar o disminuir la calificación de tu sitio web están:

- Las visitas totales de la página.
- La calidad.
- El valor de los contenidos.
- La velocidad de carga (la calificación aumenta si la página es AMP)
- Qué tan frecuentes son las actualizaciones.
- El diseño web

Asignación de Relevancia (Page Rank)

Google ordena los resultados de la búsqueda utilizando su propio algoritmo PageRank. A cada página web se le asigna un número en función del número de enlaces de otras páginas que la apuntan, el valor de esas páginas y otros criterios no públicos.



Asignación de Relevancia (Page Rank)

Vínculos como votos: una página es más importante si tiene más vínculos

- Vínculos entrantes: los que vienen de otras páginas.
- Vínculos salientes: los que van a otras páginas.

Encontrar la relevancia es un problema recursivo

- Cada voto de un vínculo entrante es proporcional a la relevancia de la página de la que viene.
- Si la página j con relevancia r j tiene n nodos salientes, cada vínculo obtiene r j / n votos.
- La relevancia de la página j es la suma de los votos de los vínculos entrantes.

Asignación de Relevancia (Page Rank)

El algoritmo inicial del PageRank lo podemos encontrar en el documento original donde sus creadores presentaron el prototipo de Google: “The Anatomy of a Large-Scale Hypertextual Web Search Engine":

PR(A): es el PageRank de la página A.

d es un factor de amortiguación que tiene un valor entre 0 y 1.

PR(i): son los valores de PageRank que tienen cada una de las páginas i que enlazan a A.

C(i) es el número total de enlaces salientes de la página i (sean o no hacia A).

Asignación de Relevancia (Page Rank)

Esta fórmula tiene su fundamento matemático en los grafos dirigidos (cada web es un nodo y los hipervínculos los caminos dirigidos), sus matrices de adyacencia y en las cadenas de Markov asociadas al proceso de búsqueda.

Representa la probabilidad de que un navegante continúe pulsando links al navegar por Internet en vez de escribir una url directamente en la barra de direcciones o pulsar uno de sus marcadores y es un valor establecido por Google.

La introducción del factor de amortiguación en la fórmula resta algo de peso a todas las páginas de Internet y consigue que las páginas que no tienen enlaces a ninguna otra página no salgan especialmente beneficiadas.

FORMULACIÓN DEL FLUJO

Asignación de Relevancia (Page Rank)

Matriz de adyacencia estocástica M

- La i-ésima página tiene d i vínculos a otras páginas
- Si i → j entonces _____, en caso contrario M j,i = 0
- M es una matriz columna estocástica: cada columna suma a 1

Vector de relevancia r

- ri es la relevancia de la i-ésima página

Ecuaciones de flujo

Asignación de Relevancia (Page Rank)

Forma matricial de ecuaciones de flujo

$$r = M \cdot r$$

El vector de relevancia r es un eigenvector de la matriz de adyacencia M

- Debido a que M es una matriz estocástica, su primer eigenvector tiene un eigenvalor asociado de $\lambda = 1$
- r es un vector estocástico y las columnas de M suman, por lo que $M \cdot r \leq 1$

Podemos calcular las relevancias de las páginas si encontramos el primer eigenvector de la matriz M.

INTERPRETACIÓN BASADO EN CAMINATAS ALEATORIAS

Considera un navegador que visita vínculos aleatoriamente

- En el paso t se encuentra en la página i
- En el paso t + 1 elige de forma aleatorio uniforme uno de los vínculos salientes de la página i
- Visita la página j correspondiente al vínculo elegido
- El proceso se repite indefinidamente

$p^{(t)}$ es un vector cuyos elementos representan la probabilidad de que el navegador se encuentre en la página i en el paso t

Es una distribución de probabilidad sobre todas las páginas.

CAMINATA ALEATORIA: DISTRIBUCIÓN ESTACIONARIA

En t + 1 se elige un vínculo de forma aleatoria uniforme

$$p^{(t+1)} = M \cdot p^{(t)}$$

$p^{(t)}$ es la distribución estacionaria si

$$p^{(t+1)} = M \cdot p^{(t)} = p^{(t)}$$

El vector r corresponde a la distribución estacionaria **p** de la caminata aleatoria

- Esta distribución es única sin importar qué probabilidad inicial $p^{(0)}$ se elija

TELETRANSPORTACIÓN ALEATORIA

Elige un vínculo de forma aleatoria con probabilidad β o salta a una página aleatoria con probabilidad 1 – β.

En callejones sin salida: se salta a una página aleatoria.

CÓMPUTO DE PAGERANK PARA DATOS MASIVOS

M es una matriz usualmente dispersa: solo se requiere almacenar en memoria una fracción de elementos.

- **A** es una matriz densa: se requiere almacenar en memoria n^2 elementos.
- Si tuviéramos 100 millones de páginas y usáramos 4 bytes por cada elemento, necesitaríamos $40^{(16)} \approx 40$ petabytes.

PAGERANK COMO CADENA DE MARKOV

PageRank se puede ver como una cadena de Markov.

- Las páginas son el conjunto de estados de la cadena.

Para cualquier π inicial, el método de las potencias convergerá a su distribución estacionaria si M es:

- Estocástica: sus columnas suman 1.
- Aperiodica: no existe una $k > 1$ tal que el intervalo entre dos visitas a un estado es siempre un múltiplo de k.
- Irreducible: desde cualquier estado hay una probabilidad no cero de llegar a cualquier otro estado.