Influencia de nodos en una red: PageRank

Teoría & Aplicación

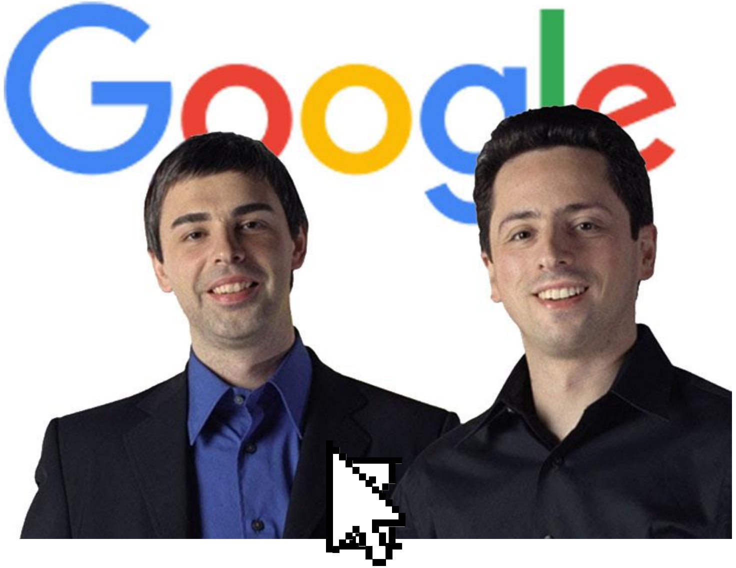
- David Halliday
- Christian Camilo Pabón
- Juan David Sánchez Murcia
- Tania Vanesa Vásquez Guevara

01

Introducción

¿Qué es PageRank?

Es un algoritmo desarrollado por Larry Page y Serguéi Brin para optimizar las búsquedas de páginas web en Google



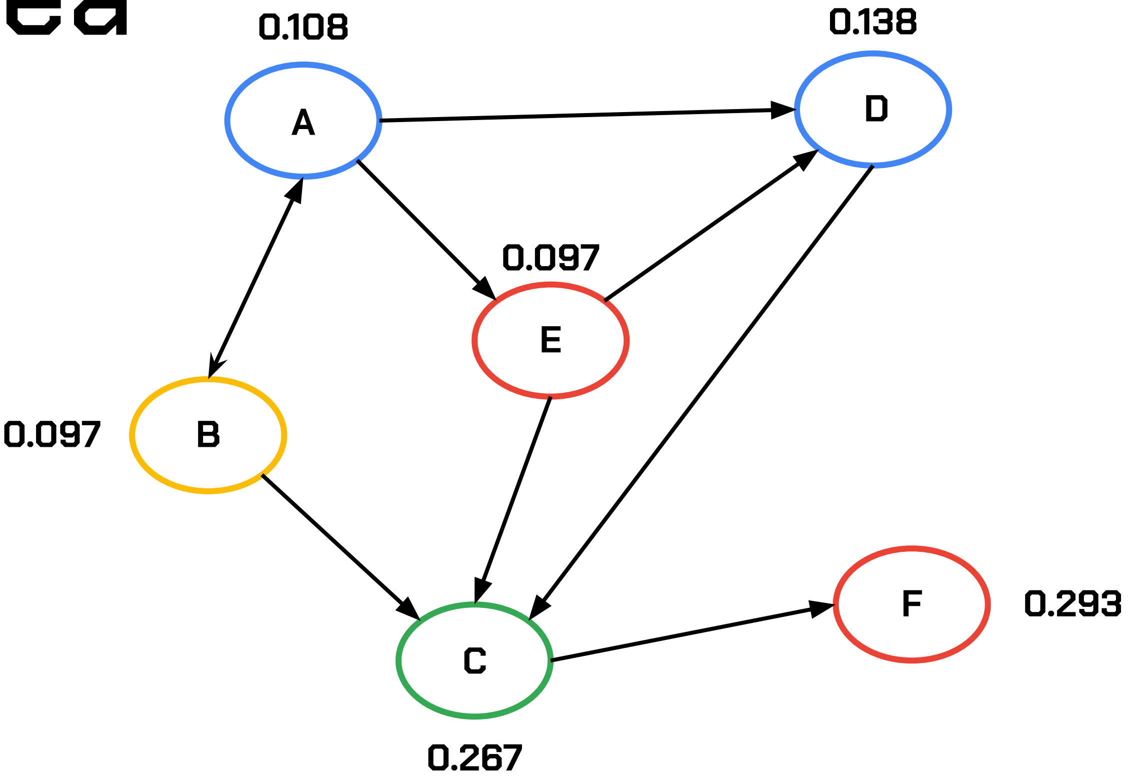
Historia

altavista™

Ask Jeeves

Google

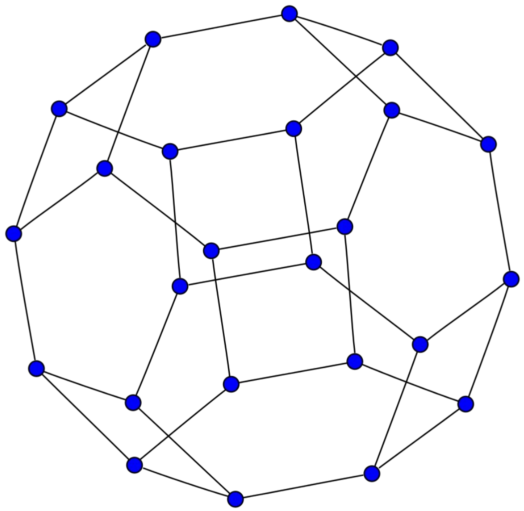
La idea



Usos

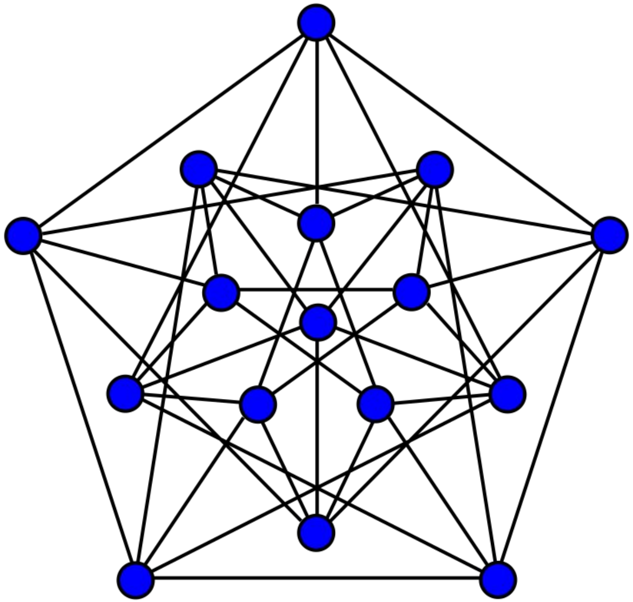
- MonitorRank: Brindar asistencia en el diagnóstico de errores en sistemas distribuidos.
- Predecir el tráfico en carreteras.
- ItemRank: Sistema de recomendación de items.
- GeneRank, ProteinRank, IsoRank.

Química, biología, neurociencia, sistemas diseñados, sistemas matemáticos, deportes, literatura, bases de datos datos, sistemas de recomendación, redes sociales, etc.



02

Grafos



Definiciones

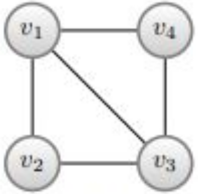
Definición 1. *Un grafo es un par $G = (V, E)$, donde $V = \{v_1, v_2, \ldots, v_n\}$ es un conjunto finito de elementos y $E = \{e_1, e_2, \ldots, e_m\} \subseteq V \times V$ es un subconjunto de pares de nodos. Si E es un subconjunto de pares ordenados de nodos, el grafo G es un grafo orientado (o grafo directo).*

Definición 2. *Si $G = (V, E)$ es un grafo directo, un camino directo en G es una secuencia de nodos $(v_0, v_1, \ldots, v_k)$ tal que $(v_i, v_{i+1}) \in E$ para cada $i = 0, 1, \ldots, k-1$. Decimos que k es la longitud del camino.*

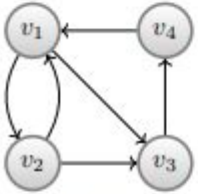
Si ninguno de los nodos, con la posible excepción de v_0 y v_k, aparece dos veces en la secuencia, el camino es simple. Si $v_0 = v_k$ el camino está cerrado. Un camino que es simple y cerrado se llama ciclo.

Definición 3. *En el mismo contexto de la Definición anterior, un camino no directo en G es una secuencia de nodos $(v_0, v_1, \ldots, v_k)$ tales que (v_i, v_{i+1}) o (v_{i+1}, v_i) pertenece a E para cada $i = 0, 1, \ldots, k-1$.*

Definición 4. *Un grafo directo es conexo si por cada par de nodos (v_i, v_j) en V existe un camino no directo que los une. Un grafo directo es fuertemente conexo si por cada par de nodos (v_i, v_j) en V existe un camino directo que los une.*



Grafo No Directo



Grafo Directo

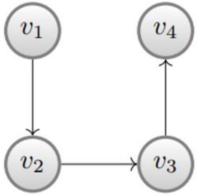


Gráfico conectado

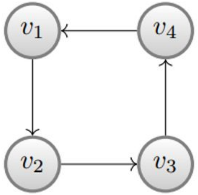


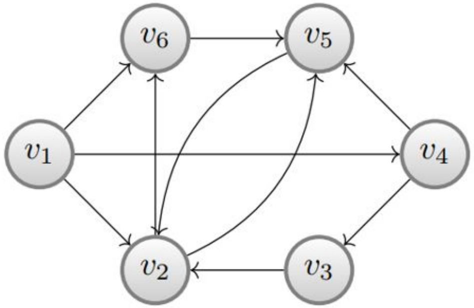
Gráfico fuertemente conectado

Matriz de Adyacencia

Supongamos que $G = (V, E)$ es un gráfico directo.

Definición 5. *Matriz de adyacencia La matriz de adyacencia A de G es una matriz cuadrada $n \times n$, donde $n = |V|$, tal que*

$$a_{ij} = \begin{cases} 1 & \textit{if } (v_i, v_j) \in E \\ 0 & \textit{if } (v_i, v_j) \notin E. \end{cases}$$



El grafo

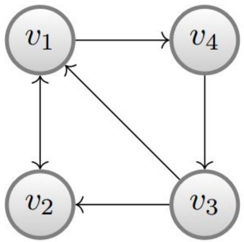
$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

La matriz de adyacencia

Matriz de Permutación

Definición 6. *Una matriz P de $n \times n$ cuyos filas se obtienen por cualquier permutación de las filas de la matriz identidad se denomina matriz de permutación.*

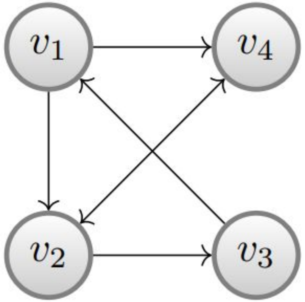
Definición 7. *Dada cualquier matriz cuadrada A de $n \times n$, el resultado $A' = PAP^T$, donde P es una matriz de permutación, se denomina permutación simétrica de A.*



$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

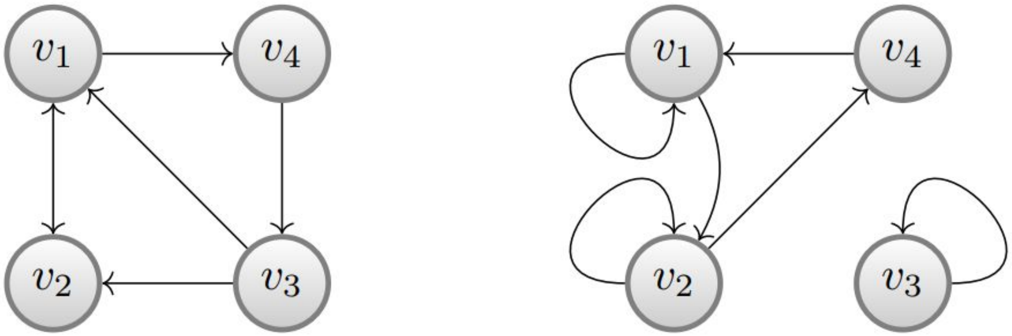
$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

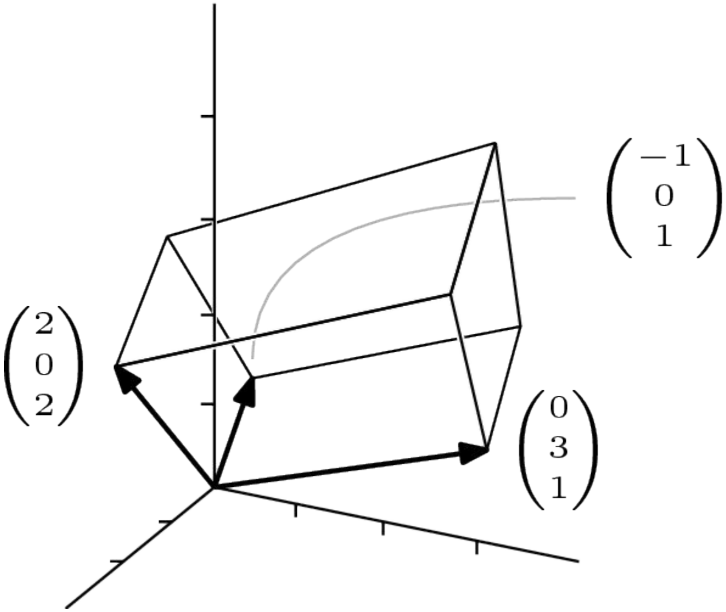
$$PAP^T = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$



Es fácil ver que el efecto de la permutación simétrica en la matriz de adyacencia de un grafo es simplemente cambiar el nombre de los nodos del grafo. En el gráfico se ha producido la permutación $(3\,1\,4\,2)$ de los nodos, en el sentido de que: el nuevo nodo 1 es el antiguo nodo 3 , el nuevo nodo 2 es el antiguo 1 , el nuevo nodo 3 es el antiguo 4 y el el nuevo nodo 4 es el antiguo 2 . Vale la pena señalar que la única multiplicación que queda de la matriz de adyacencia por una permutación P no es interesante y no está relacionada con una permutación de los nodos en el gráfico.

$$PA = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$



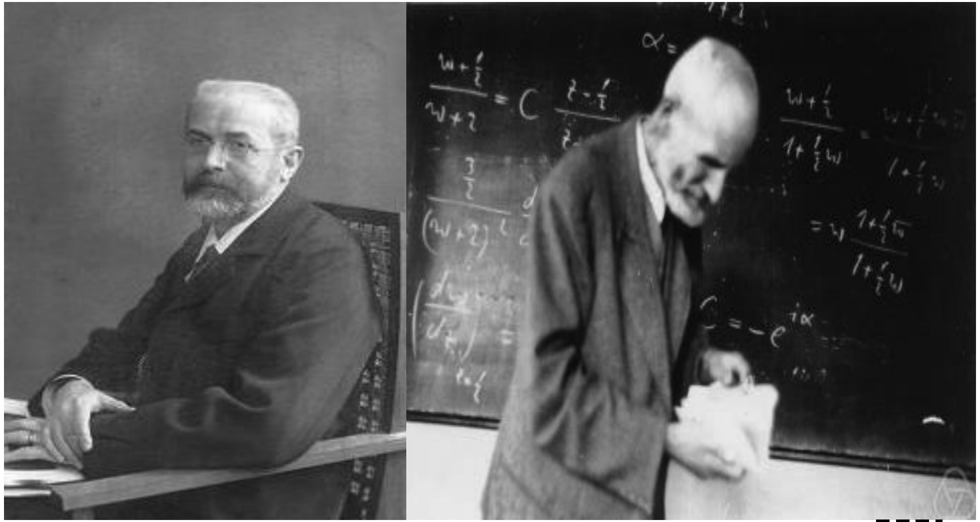


03
Álgebra Lineal

$$\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$$

Teorema de Perron-Frobenius

Nos da información acerca de los valores y vectores propios de una matriz positiva, una primitiva y una no negativa.



Positiva

$$A = [\, a_{i,j} \,]$$

$$a_{i,j} > 0$$

No negativa

$$A = [\, a_{i,j} \,]$$

$$a_{i,j} \geq 0$$

Primitiva

Si es no negativa y tiene alguna potencia positiva

Es equivalente a ser no negativa, irreducible y aperiódica

Irreducible

Una matriz A se dice irreducible si

(i) A no puede ser conjugada en forma de bloque triangular superior por una matriz de permutación P

$$PAP^{-1} \neq \begin{pmatrix} E & F \\ O & G \end{pmatrix}$$

(ii) Si asociamos la matriz A a la matriz de adyacencia de un grafo dirigido G_A, entonces A es irredicible si y solo si su grafo asociado G_A es fuertemente conexo

Teorema de Perron-Frobenius

Puede ser extendido a no negativas y primitivas

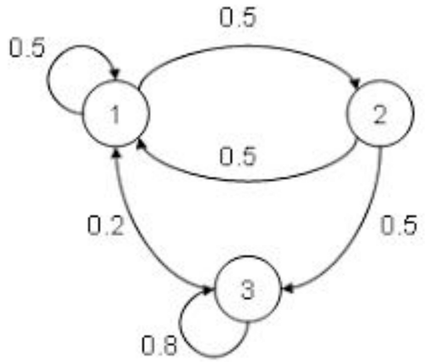
Teorema 1 (Teorema de Perron-Frobenius). *Sea A una matriz positiva $n \times n$ con $\rho(A) = r$, donde $\rho(A)$ es el radio espectral de A. Entonces los siguientes enunciados son verdaderos:*

(i) $r > 0$

(ii) r *es un valor propio con vector propio* $v > 0$. v *es el único vector propio no negativo (salvo múltiplos) y es llamado el vector de Perron.*

(iii) r *es el único valor propio en el circulo espectral de* A

(iv) r *es una raíz simple del polinomio característico de* A.



04
Cadenas de Markov

Para cada $t = 1, 2, \ldots$, para cada par de estados i, j y para cada secuencia de estados $v_0, v_1, \ldots, v_{t-1} \in E$

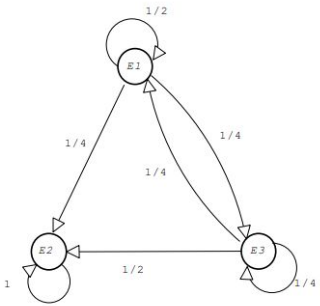
$$Prob(X_{t+1} = j \mid X_0 = v_0, X_1 = v_1, \ldots, X_{t-1} = v_{t-1}, X_t = i) = Prob(X_{t+1} = j \mid X_t = i).$$

Estos significa que la probabilidad de que el sistema pase del estado i al estado j en el tiempo t, es independiente de los estados anteriores al tiempo t.

Matriz de transición

1. $p_{ij} \geq 0$ para cada i, j
2. $\sum_{j=1}^{n} p_{ij} = 1$ para cada i, j

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & 1 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$



Definición 8. *Una cadena de Markov es irreducible si su matriz de transición P es irreducible.*

Definición 10. *Un vector de probabilidad p es un vector de probabilidad estacionario para una cadena de Markov con matriz de transición P si:*

$$p^T P = P^T$$

$$\boxed{p^T(t) = (p_1(t), \ldots, p_n(t))}$$

$$\begin{aligned}
p_j(t+1) &= \text{Prob}(X_{t+1} = j) = \text{Prob}((X_{t+1} = j) \cap (X_t = 1 \cup \ldots \cup X_t = n)) \\
&= \text{Prob}((X_{t+1} = j \cap X_t = 1) \cup \ldots \cup (X_{t+1} \cap X_t = n)) \\
&= \sum_{i=1}^{n} \text{Prob}(X_{t+1} = j \cap X_t = i) \\
&= \sum_{i=1}^{n} \text{Prob}(X_{t+1} = j \mid X_t = i) \cdot \text{Prob}(X_t = i) \\
&= \sum_{i=1}^{n} p_{ij} \cdot p_i(t)
\end{aligned}$$

$$\boxed{p^T(t+1) = p^T(t) P.}$$

Iteration 1

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.213 \\ 0.426 \\ 0.426 \\ 0.639 \\ 0.426 \end{bmatrix}$$

05

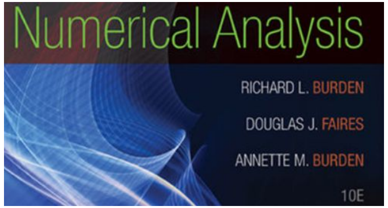
Método de potencias

Iteration 2

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.213 \\ 0.426 \\ 0.426 \\ 0.639 \\ 0.426 \end{bmatrix} = \begin{bmatrix} 0.426 \\ 0.852 \\ 1.065 \\ 1.278 \\ 1.065 \end{bmatrix} = \begin{bmatrix} 0.195 \\ 0.389 \\ 0.486 \\ 0.584 \\ 0.486 \end{bmatrix}$$

Iteration 3

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.195 \\ 0.389 \\ 0.486 \\ 0.584 \\ 0.486 \end{bmatrix} = \begin{bmatrix} 0.389 \\ 0.779 \\ 1.07 \\ 1.361 \\ 1.07 \end{bmatrix} = \begin{bmatrix} 0.176 \\ 0.352 \\ 0.484 \\ 0.616 \\ 0.484 \end{bmatrix}$$



Iteration 4

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.176 \\ 0.352 \\ 0.484 \\ 0.616 \\ 0.484 \end{bmatrix} = \begin{bmatrix} 0.352 \\ 0.792 \\ 1.100 \\ 1.320 \\ 1.100 \end{bmatrix}$$

Matriz A de tamaño $n \times n$

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$$

Definición iterativa

$$x^{(i)} = Ax^{(i-1)}, \text{ con } i = 1, 2, \ldots.$$

Convergencia

Para la sucesión $\{x^{(k)}\}_{k\in\mathbb{N}}$ *se cumplen las siguientes dos igualdades*

$$\lim_{k\to\infty} \frac{x_j^{(k+1)}}{x_j^{(k)}} = \lambda_1 \quad y \quad \lim_{k\to\infty} \frac{x^{(k)}}{x_j^{(k)}} = cv^1,$$

$$r(P) = \sum_{Q \text{ tiene enlace a } P} \frac{r(Q)}{|Q|}$$

$$a_{ij} = Prob(P_j \text{ tiene enlace a } P_i) = \begin{cases} \frac{1}{|P_j|} & \text{si } P_j \text{ tiene enlace a } P_i \\ 0 & \text{en otro caso,} \end{cases}$$

06

Algoritmo PageRank

$$r^{(0)} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)^T$$

$$r^{(k)} = Ar^{(k-1)}$$

Una matriz de Markov P converge únicamente a un vector estrictamente positivo si se cumple que P es estocástica, irreducible y aperiódica.

Mejoras al algoritmo

La matriz estocástica S

$$S = A + a\left(\frac{1}{n}e^T\right)$$

La matriz de Google G

$$G = \alpha S + \frac{(1-\alpha)}{n}ee^T$$

Parámetro de *random surfer* es $0 \leq \alpha \leq 1$

$$\begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array}
\begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\
\begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}
\longrightarrow
\begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

07

Implementaciones

neo4j

Referencias

- Gleich, David F.. "PageRank beyond the Web." (2014).
- Alberto Peretti, Alberto Roveda et al. On the mathematical background of Google PageRank algorithm. Inf. téc. 2014
- Carl D Meyer. Matrix analysis and applied linear algebra. Vol. 71. Siam, 2000
- Charles R MacCluer. “The many proofs and applications of Perron’s theorem”. En: Siam Review 42.3 (2000), págs. 487-498
- Amy N Langville y Carl D Meyer. “Google’s PageRank and beyond”. En: Google’s PageRank and Beyond. Princeton University Press, 2011
- Keshi Dai. PageRank Lecture Note. 2009.
- Hannah Cairns. “Perron’s Theorem in an Hour”. En: The American Mathematical Monthly 128.8 (2021), págs. 748-752

¡Gracias por su atención!