

UNIDAD 2: ANÁLISIS DE VÍNCULOS

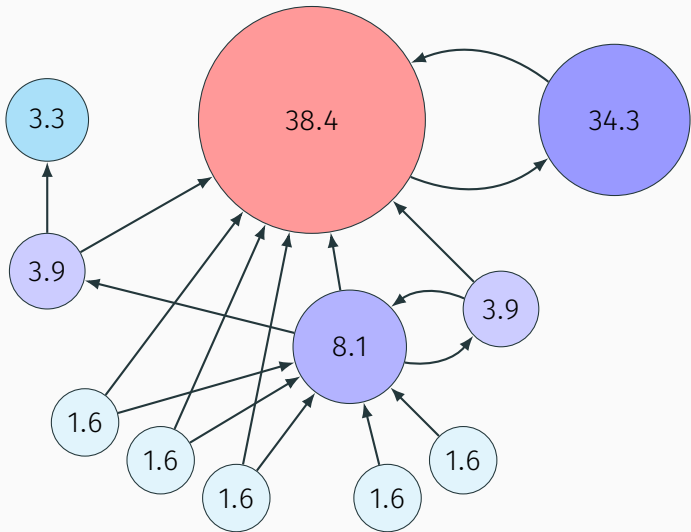
ASIGNACIÓN DE RELEVANCIA (PAGERANK)

Blanca Vázquez y Gibran Fuentes-Pineda

22 de septiembre de 2021

- Desarrollado por Larry Page, cofundador de Google, para asignar relevancia a páginas Web en motores de búsqueda
- Cuenta el número y calidad de los vínculos a una página para estimar su relevancia
- Páginas más relevantes tienen mayor probabilidad de recibir vínculos de más páginas
- Toma en cuenta Hubs y autoridades

EJEMPLO



VÍNCULOS COMO VOTOS

- 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

FORMULACIÓN DE FLUJO

- La relevancia r_j para una página j está dada por

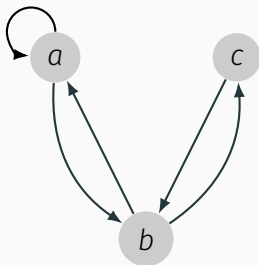
$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$$

- Genera un sistema de n ecuaciones con n variables
 - No hay solución única
 - Restricción adicional: $\sum_i r_i = 1$

$$r_a = r_a/2 + r_b/2$$

$$r_b = r_a/2 + r_c$$

$$r_c = r_b/2$$



- Matriz de adyacencia estocástica \mathbf{M}
 - La i -ésima página tiene d_i vínculos a otras páginas
 - Si $i \rightarrow j$ entonces $\mathbf{M}_{j,i} = \frac{1}{d_i}$, en caso contrario $\mathbf{M}_{j,i} = 0$
 - \mathbf{M} es una matriz columna estocástica: cada columna suma a 1
- Vector de relevancia \mathbf{r}
 - r_i es la relevancia de la i -ésima página
 - $\sum_{i=1}^n r_i = 1$
- Ecuaciones de flujo

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$$

- Forma matricial de ecuaciones de flujo

$$\mathbf{r} = \mathbf{M} \cdot \mathbf{r}$$

- El vector de relevancia \mathbf{r} es un eigenvector de la matriz de adyacencia \mathbf{M}
 - Debido a que \mathbf{M} es una matriz estocástica, su primer eigenvector tiene un eigenvalor asociado de $\lambda = 1$
 - \mathbf{r} es un vector estocástico y las columnas de \mathbf{M} suman, por lo que $\mathbf{M} \cdot \mathbf{r} \leq 1$
- Podemos calcular las relevancias de las páginas si encontramos el primer eigenvector de la matriz \mathbf{M}

- Esquema iterativo

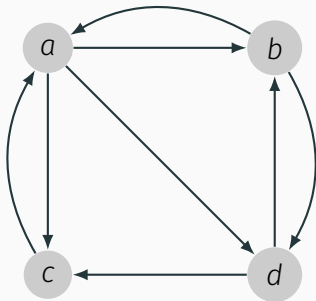
1. $\mathbf{r}^{(0)} = [1/n, 1/n, \dots, 1/n]$

2. $\mathbf{r}^{(t+1)} = \mathbf{M} \cdot \mathbf{r}^{(t)}$

3. Repite 2 hasta que $\|\mathbf{r}^{(t+1)} - \mathbf{r}^{(t)}\|_1 < \epsilon$

EJEMPLO DEL MÉTODO DE LAS POTENCIAS

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



	$t = 0$	$t = 1$	$t = 2$	$t = 3$	\dots
$\begin{bmatrix} r_a \\ r_b \\ r_c \\ r_d \end{bmatrix}$	$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$	$\begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}$	$\begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}$	$\begin{bmatrix} 11/32 \\ 7/32 \\ 7/32 \\ 7/32 \end{bmatrix}$	\dots

$$= \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

- 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
- $\mathbf{p}^{(t)}$ 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

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

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

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

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

- El vector \mathbf{r} corresponde a la distribución estacionaria \mathbf{p} de la caminata aleatoria
 - Esta distribución es única sin importar qué probabilidad inicial $\mathbf{p}^{(0)}$ se elija

PROBLEMAS CON FORMULACIÓN DE PAGERANK: TRAMPA DE ARAÑA

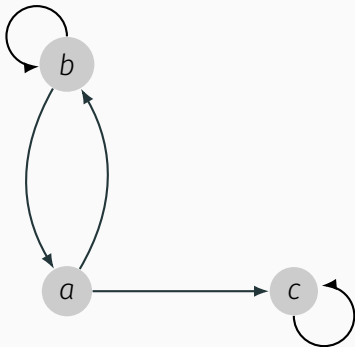
- Todos los vínculos salientes están dentro del grupo
 - Eventualmente absorben toda la relevancia



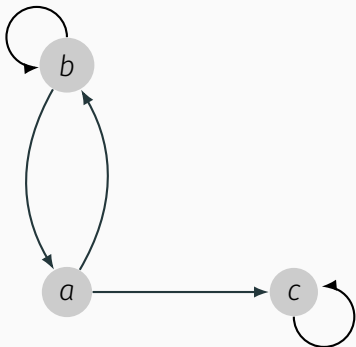
$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i}$$

$$\begin{matrix} & t=0 & t=1 & t=2 & t=3 & \dots \\ \begin{bmatrix} r_a \\ r_b \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & \dots \\ 0 & 1 & 0 & 1 & \dots \end{bmatrix} \end{matrix}$$

EJEMPLO DE PAGERANK CON UNA TRAMPA DE ARAÑA



EJEMPLO DE PAGERANK CON UNA TRAMPA DE ARAÑA



$$\mathbf{M} = \begin{matrix} & \begin{matrix} a & b & c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{bmatrix} \end{matrix}$$

$$\begin{bmatrix} r_a \\ r_b \\ r_c \end{bmatrix} = \begin{matrix} & \begin{matrix} t=0 & t=1 & t=2 & t=3 & \dots \end{matrix} \\ \begin{matrix} r_a \\ r_b \\ r_c \end{matrix} & \begin{bmatrix} 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 2/6 & 3/12 & 5/24 & \dots & 0 \\ 1/3 & 3/6 & 7/12 & 16/24 & \dots & 1 \end{bmatrix} \end{matrix}$$

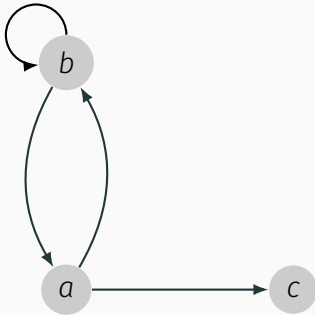
- Callejones sin salida: páginas sin vínculos salientes
 - Causa fuga en la relevancia



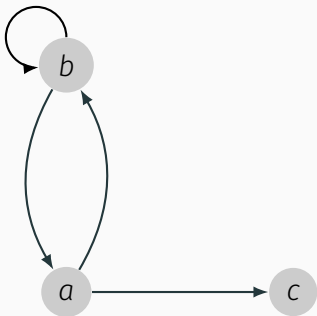
$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i}$$

$$\begin{bmatrix} r_a \\ r_b \end{bmatrix} = \begin{array}{c} \begin{matrix} t=0 & t=1 & t=2 & t=3 & \dots \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \end{bmatrix} \end{array}$$

EJEMPLO DE PAGERANK CON UN CALLEJÓN SIN SALIDA



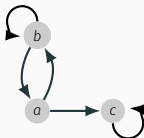
EJEMPLO DE PAGERANK CON UN CALLEJÓN SIN SALIDA



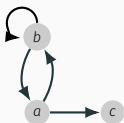
$$M = \begin{matrix} & \begin{matrix} a & b & c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \end{bmatrix} \end{matrix}$$

$$\begin{bmatrix} r_a \\ r_b \\ r_c \end{bmatrix} = \begin{matrix} & \begin{matrix} t=0 & t=1 & t=2 & t=3 & \dots \end{matrix} \\ \begin{matrix} r_a \\ r_b \\ r_c \end{matrix} & \begin{bmatrix} 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 2/6 & 3/12 & 5/24 & \dots & 0 \\ 1/3 & 3/6 & 1/12 & 2/24 & \dots & 0 \end{bmatrix} \end{matrix}$$

- Elige un vínculo de forma aleatoria con probabilidad β o salta a una página aleatoria con probabilidad $1 - \beta^1$



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



$$\mathbf{M} = \begin{matrix} & \begin{matrix} a & b & c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{bmatrix} 0 & 1/2 & 1/3 \\ 1/2 & 1/2 & 1/3 \\ 1/2 & 0 & 1/3 \end{bmatrix} \end{matrix}$$

¹En la práctica es común que β sea un valor entre 0.8 y 0.9

- PageRank con teletransportaciones

$$r_j^{(t+1)} = \sum_{i \rightarrow j} \beta \cdot \frac{r_i^{(t)}}{d_i} + (1 - \beta) \cdot \frac{1}{n}$$

- La matriz de Google

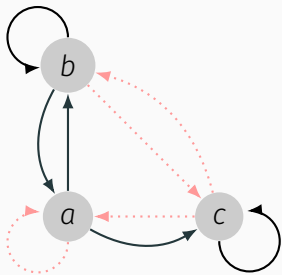
$$\mathbf{A} = \beta \cdot \mathbf{M} + (1 - \beta) \cdot \frac{1}{n} \mathbf{e} \cdot \mathbf{e}^\top$$

donde \mathbf{e} es un vector de tamaño n cuyos elementos son 1

- \mathbf{A} es estocástica, aperiodica e irreducible, por lo que

$$\mathbf{r}^{(t+1)} = \mathbf{A} \cdot \mathbf{r}^{(t)}$$

EJEMPLO DE TELETRANSPORTACIÓN



$$A = \underbrace{0.8}_{\beta} \cdot \underbrace{\begin{matrix} & \begin{matrix} a & b & c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{bmatrix} \end{matrix}}_M + \underbrace{0.2}_{1-\beta} \cdot \underbrace{\begin{matrix} & \begin{matrix} a & b & c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \end{matrix}}_{\frac{1}{n} \mathbf{e} \cdot \mathbf{e}^T}$$

$$A = \begin{matrix} & \begin{matrix} a & b & c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 13/15 \end{bmatrix} \end{matrix}$$

$$\begin{bmatrix} r_a \\ r_b \\ r_c \end{bmatrix} = \begin{matrix} & t=0 & t=1 & t=2 & t=3 & \dots \\ \begin{bmatrix} r_a \\ r_b \\ r_c \end{bmatrix} & \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} & \begin{bmatrix} 0.20 \\ 0.33 \\ 0.46 \end{bmatrix} & \begin{bmatrix} 0.20 \\ 0.24 \\ 0.52 \end{bmatrix} & \begin{bmatrix} 0.18 \\ 0.26 \\ 0.56 \end{bmatrix} & \dots & \begin{bmatrix} 5/33 \\ 7/33 \\ 21/33 \end{bmatrix} \end{matrix}$$

CÓMPUTO DE PAGERANK PARA DATOS MASIVOS (1)

- \mathbf{M} es una matriz usualmente dispersa: solo se requiere almacenar en memoria una fracción de elementos
- \mathbf{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
- Reorganizando la expresión $\mathbf{r}^{(t+1)} = \mathbf{A} \cdot \mathbf{r}^{(t)}$

$$\begin{aligned}\mathbf{r}^{(t+1)} &= \left[\beta \cdot \mathbf{M} + \frac{1 - \beta}{n} \cdot \mathbf{e} \cdot \mathbf{e}^\top \right] \cdot \mathbf{r}^{(t)} \\ &= \beta \cdot \mathbf{M} \cdot \mathbf{r}^{(t)} + \frac{1 - \beta}{n} \cdot \mathbf{e} \cdot \mathbf{e}^\top \cdot \mathbf{r}^{(t)} \\ &= \beta \cdot \mathbf{M} \cdot \mathbf{r}^{(t)} + \frac{1 - \beta}{n}\end{aligned}$$

- Sin callejones sin salida
 1. Calcula $\hat{\mathbf{r}}^{(t+1)} = \beta \cdot \mathbf{M} \cdot \mathbf{r}^{(t)}$
 2. Agrega $(1-\beta)/n$ a los elementos de $\hat{\mathbf{r}}^{(t+1)}$
- Con callejones sin salida
 1. Calcula $\hat{\mathbf{r}}^{(t+1)} = \beta \cdot \mathbf{M} \cdot \mathbf{r}^{(t)}$
 2. Agrega $(1-\sum_j \hat{r}_j^{(t+1)})/n$ a los elementos de $\hat{\mathbf{r}}^{(t+1)}$

PAGERANK COMO CADENA DE MARKOV

- PageRank se puede ver como una cadena de Markov
 - Las páginas son el conjunto de estados de la cadena
 - La matriz de transición \mathbf{P} es $P_{i,j} = p(x^{[t+1]} = i | x^{[t]} = j)$
 - π es la distribución estacionaria: $\pi = \mathbf{P}\pi$
- Para cualquier π inicial, el método de las potencias convergerá a su distribución estacionaria si \mathbf{P} 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