Unidad 2: Análisis de vínculos

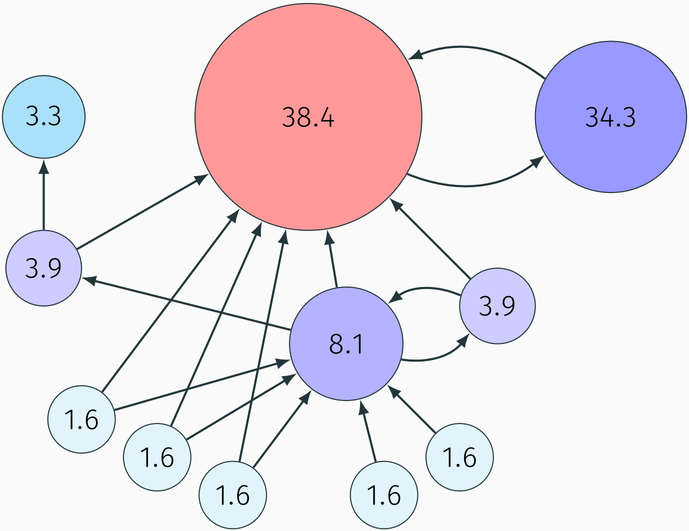
Asignación de relevancia (PageRank)

Blanca Vázquez y Gibran Fuentes-Pineda

6 de octubre de 2022

- 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

- La relevancia r_j para una página j está dada por

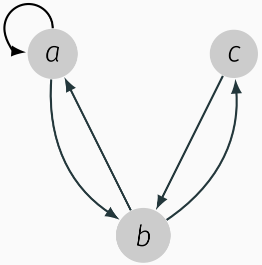
$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

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

$$r_a = {}^{r_a}\!/_2 + {}^{r_b}\!/_2$$

$$r_b = {}^{r_a}\!/_2 + r_c$$

$$r_c = {}^{r_b}\!/_2$$



Forma matricial

- Matriz de adyacencia estocástica M
  - La i-ésima página tiene d_i vínculos a otras páginas
  - Si $i \to j$ entonces $\mathsf{M}_{j,i} = \frac{1}{d_i}$, en caso contrario $\mathsf{M}_{j,i} = 0$
  - 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

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

PageRank y eigenvectores

- Forma matricial de ecuaciones de flujo

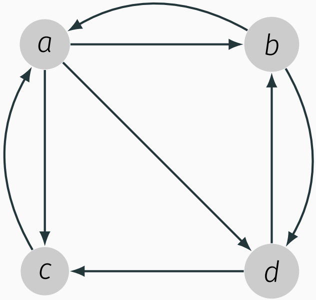
$$\mathbf{r} = \mathbf{M} \cdot \mathbf{r}$$

- 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. $\mathsf{r}^{(0)} = [1/n, 1/n, \ldots, 1/n]$
  2. $\mathsf{r}^{(t+1)} = \mathsf{M} \cdot \mathsf{r}^{(t)}$
  3. Repite 2 hasta que $\|\mathsf{r}^{(t+1)} - \mathsf{r}^{(t)}\|_1 < \epsilon$

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



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

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

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

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

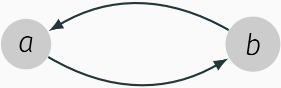
- $\mathsf{p}^{(t)}$ es la distribución estacionaria si

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

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

Problemas con formulación de PageRank: trampa de araña

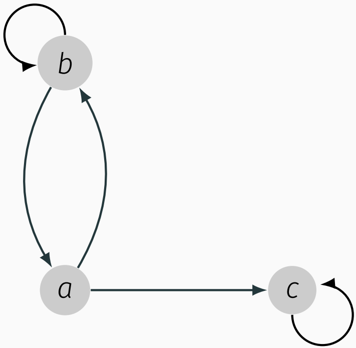
- Todos los vínculos salientes están dentro del grupo
  - Eventualmente absorben toda la relevancia



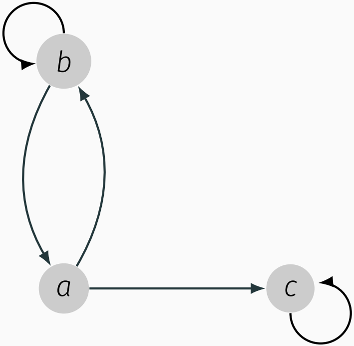
$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

$$\begin{array}{cc} & \begin{array}{ccccc} t=0 & t=1 & t=2 & t=3 & \cdots \end{array} \\ \begin{bmatrix} \mathrm{r}_a \\ \mathrm{r}_b \end{bmatrix} = & \begin{bmatrix} 1 & 0 & 1 & 0 & \cdots \\ 0 & 1 & 0 & 1 & \cdots \end{bmatrix} \end{array}$$

Ejemplo de PageRank con una trampa de araña



Ejemplo de PageRank con una trampa de araña



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

$$
\begin{bmatrix} \mathrm{r}_a \\ \mathrm{r}_b \\ \mathrm{r}_c \end{bmatrix} =
\begin{array}{c}
\begin{array}{cccccc} t=0 & t=1 & t=2 & t=3 & \cdots & \end{array} \\
\begin{bmatrix} 1/3 & 1/6 & 2/12 & 3/24 & \cdots & 0 \\ 1/3 & 2/6 & 3/12 & 5/24 & \cdots & 0 \\ 1/3 & 3/6 & 7/12 & 16/24 & \cdots & 1 \end{bmatrix}
\end{array}
$$

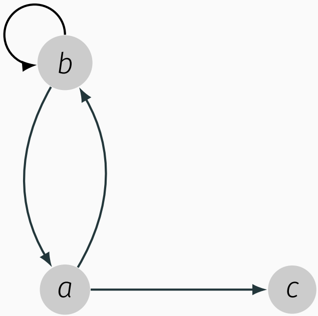
Problemas con formulación de PageRank: callejón sin salida

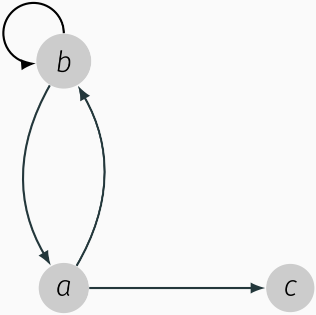
- Callejones sin salida: páginas sin vínculos salientes
  - Causa fuga en la relevancia

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

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

Ejemplo de PageRank con un callejón sin salida



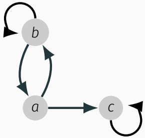


$$\mathsf{M} = \begin{array}{cc} & \begin{array}{ccc} a & b & c \end{array} \\ \begin{array}{c} a \\ b \\ c \end{array} & \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \end{bmatrix} \end{array}$$

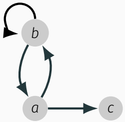
$$\begin{bmatrix} r_a \\ r_b \\ r_c \end{bmatrix} = \begin{array}{c} \begin{array}{ccccc} t=0 & t=1 & t=2 & t=3 & \cdots \end{array} \\ \begin{bmatrix} 1/3 & 1/6 & 2/12 & 3/24 & \cdots & 0 \\ 1/3 & 2/6 & 3/12 & 5/24 & \cdots & 0 \\ 1/3 & 3/6 & 1/12 & 2/24 & \cdots & 0 \end{bmatrix} \end{array}$$

Teletransportación aleatoria

- 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



$$\mathsf{M} = \begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0 & 1/2 & 1/3 \\ 1/2 & 1/2 & 1/3 \\ 1/2 & 0 & 1/3 \end{bmatrix} \end{array}$$

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

Método de potencias para teletransportaciones

- PageRank con teletransportaciones

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

- La matriz de Google

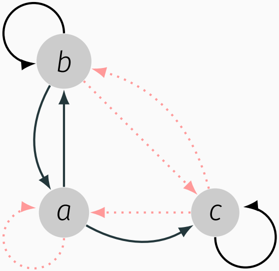
$$\mathsf{A} = \beta \cdot \mathsf{M} + (1 - \beta) \cdot \frac{1}{n} \mathsf{e} \cdot \mathsf{e}^\top$$

donde e es un vector de tamaño n cuyos elementos son 1

- A es estocástica, aperiodica e irreducible, por lo que

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

Ejemplo de teletransportación



$$\mathbf{A} = \overbrace{0.8}^{\beta} \cdot \overbrace{\begin{array}{cc} & \begin{matrix} a & \;\; b & \;\; c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{bmatrix} \end{array}}^{\mathbf{M}} + \overbrace{0.2}^{1-\beta} \cdot \overbrace{\begin{array}{cc} & \begin{matrix} a & \;\; b & \;\; c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \end{array}}^{\frac{1}{n}\mathbf{e}\cdot\mathbf{e}^{\top}}$$

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

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

Cómputo de PageRank para datos masivos (1)

- 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

- Reorganizando la expresión $\mathsf{r}^{(t+1)} = \mathsf{A} \cdot \mathsf{r}^{(t)}$

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

Cómputo de PageRank para datos masivos (2)

- 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 ${\left(1-\sum_j \hat{r}_j^{(t+1)}\right)}/{n}$ a los elementos de $\hat{\mathbf{r}}^{(t+1)}$

- PageRank se puede ver como una cadena de Markov
    - Las páginas son el conjunto de estados de la cadena
    - La matriz de transición $\mathbf{M}$ es $M_{i,j} = p(x^{[t+1]} = i | x^{[t]} = j)$
    - $\boldsymbol{\pi}$ es la distribución estacionaria: $\boldsymbol{\pi} = \mathbf{M}\boldsymbol{\pi}$

- Para cualquier $\boldsymbol{\pi}$ inicial, el método de las potencias convergerá a su distribución estacionaria si $\mathbf{M}$ 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

Multiplicación de una matriz y un vector (1)

- Dada una matriz de $n \times n$ y un vector de tamaño n, calcular $\mathbf{x} = \mathbf{M} \cdot \mathbf{v}$

$$x_i = \sum_j M_{i,j} \cdot v_j$$

- Función de mapeo
  - se lee $\mathbf{v}$ a la memoria principal (si no se ha hecho)
  - Se aplica a cada elemento de $\mathbf{v}$
  - Cada tarea se realiza en un fragmento de $\mathbf{M}$
  - Regresa pares $(i, M_{i,j} \cdot v_j)$

- Función de reducción
  - Suma todos los valores asociados a la llave i
  - Regresa pares (i, x_i)

Multiplicación de una matriz y un vector (2)

- Si **v** no cabe en memoria principal, se divide **M** en k segmentos verticales y **v** en k segmentos horizontales

- Mapeo se aplica a cada elemento de los segmentos correspondientes de **M** y **v**

- Las tareas de mapeo y de reducción se realizan de la misma manera que cuando **v** cabe en memoria principal