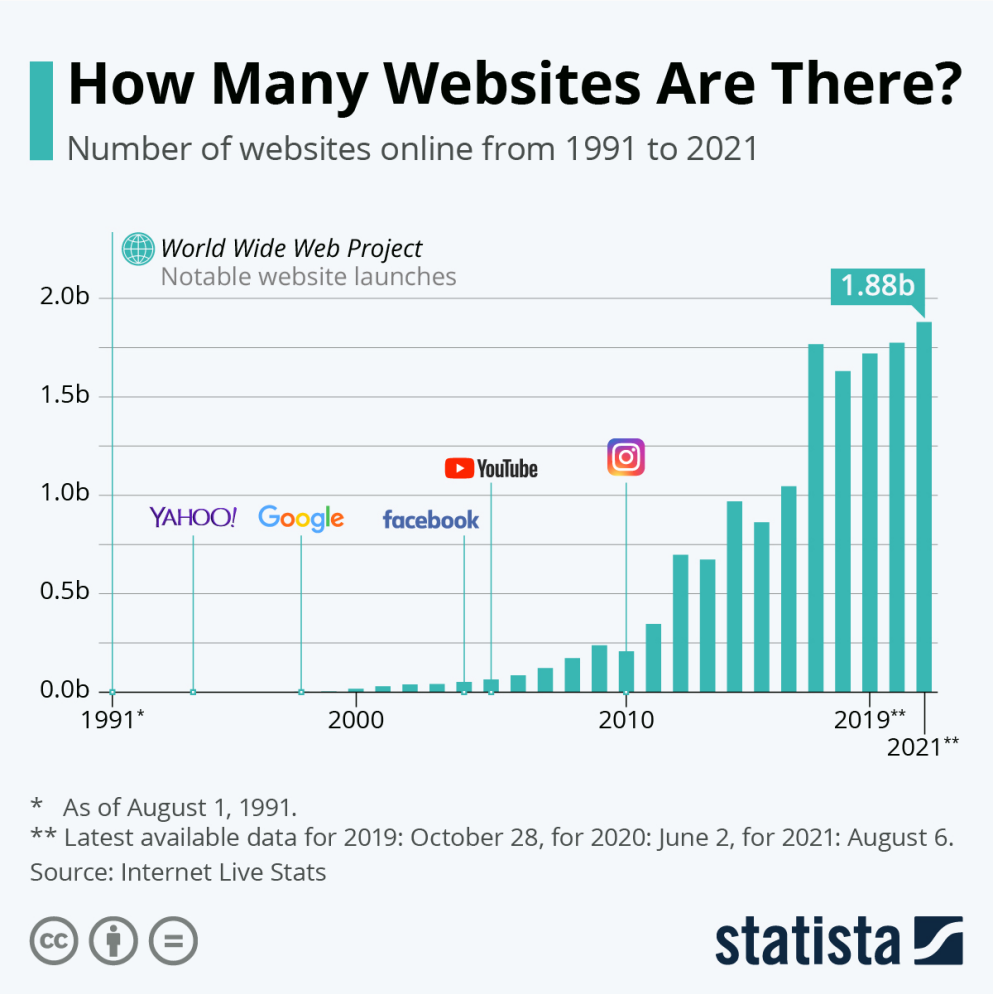
Unidad 3: Análisis de vínculos

Page rank sensible al tópico

Blanca Vázquez y Gibran Fuentes-Pineda

11 de octubre de 2022

Introducción

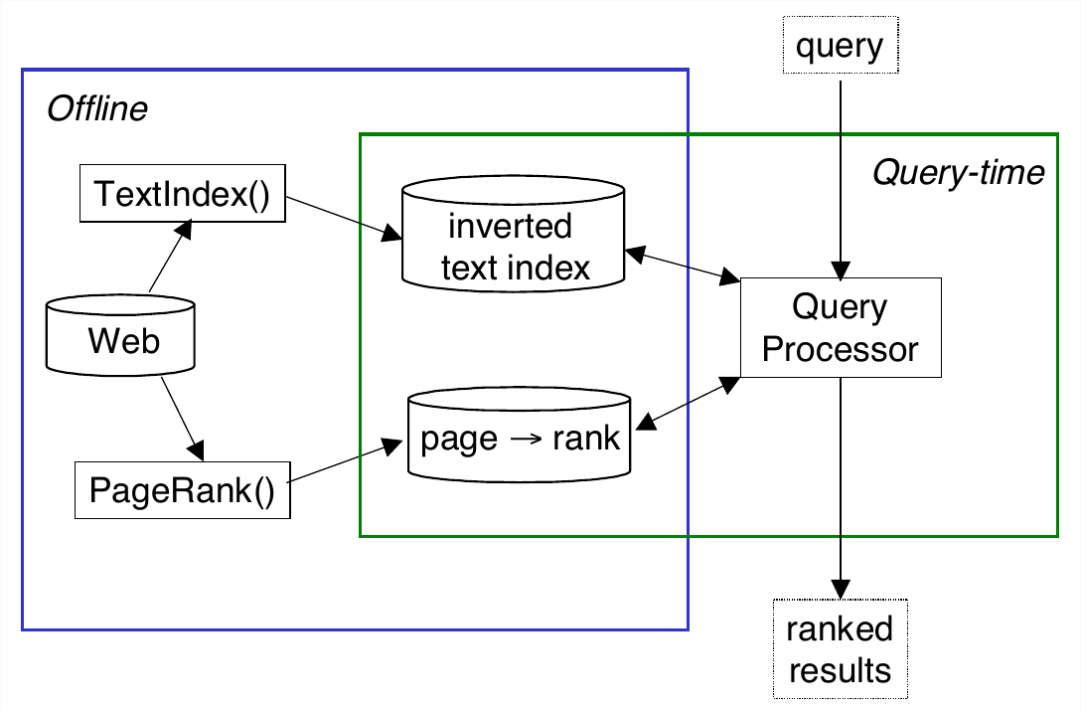


Casi el 90 % del tráfico se encuentra en los buscadores.

Críticas a las soluciones existentes

	HITS	PageRank
Ventajas	- Simple e iterativo - Puntuación específica de la consulta	- Poco costoso (en tiempo de ejecución) - Las puntuaciones se calculan utilizando el grafo completo
Desventajas	- Costoso (tiempo de ejecución) - Las puntuaciones se calculan utilizando un subgrafo a partir de todo el grafo.	- La puntuación es independiente de la consulta - El algoritmo es propenso a manipulaciones (granjas de enlaces)

Algoritmo PageRank

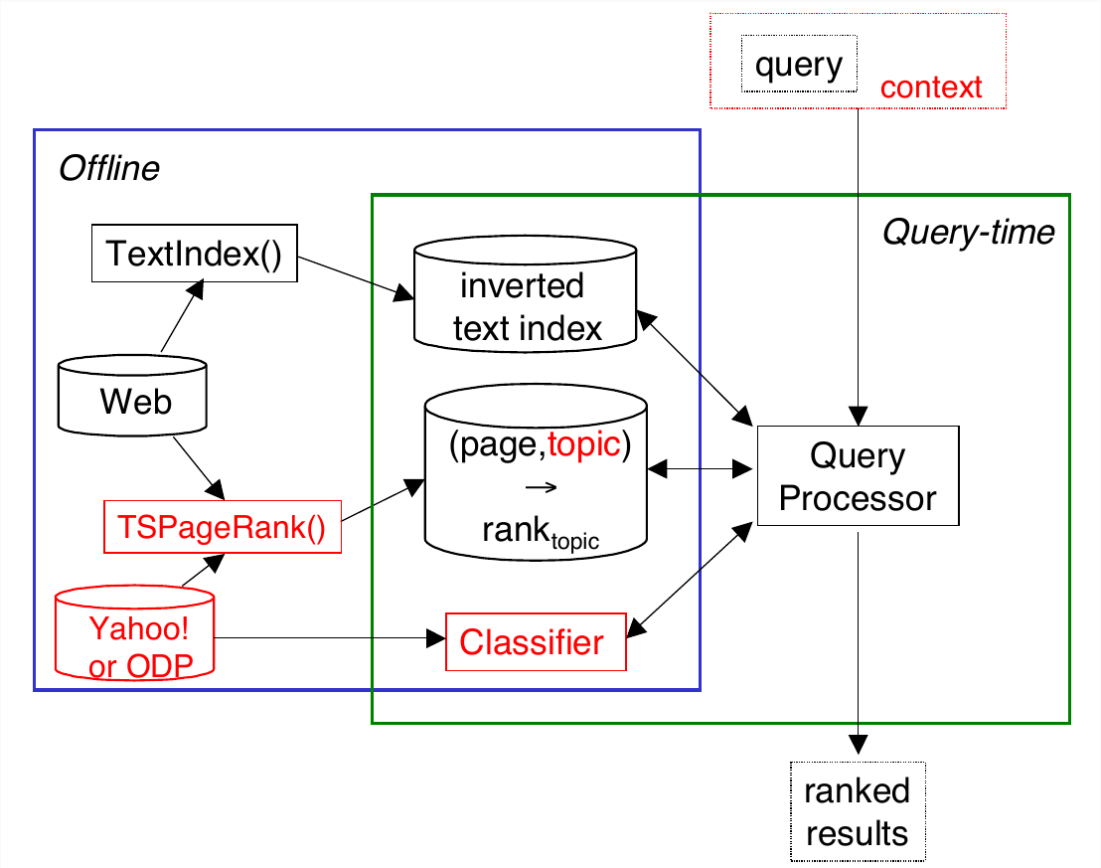


Motor de búsqueda usando el algoritmo de PageRank
Imagen tomada de Haveliwala, 2003.

PageRank sensible al tópico

- TSPR son las siglas de Topic-Sensitive PageRank
- Propuesto por Taher H. Haveliwala de la Universidad de Stanford en el 2003.
- Es la versión personalizada de *Page Rank*.
- En lugar de calcular un solo vector de rango, ¿por qué no calcular un conjunto de vectores de rango (uno por cada tópico)?

Algoritmo Page Rank Sensible al tópico



Motor de búsqueda usando el algoritmo de PageRank sensible al topico
Imagen tomada de Haveliwala, 2003.

Intuición

- Supongamos que creamos un vector único para cada tópico usando PageRank.
- Si se pudiera determinar ¿cuál de estos tópicos son de interés para el usuario?, entonces :
  - Se podría usar el vector de Page Rank de ese tópico cuando se clasifiquen las páginas por relevancia.

Tópicos en TSPR

- Usa el proyecto *Open Directory Project* como fuente de selección de tópicos ***https://dmoz-odp.org/***
  - También conocido como *DMoz* por *directory.mozilla.org*
- Es una colección de páginas web clasificadas por humanos.
- Consta de 16 tópicos (deportes, medicina,...)

Caminatas aleatorias sesgadas

Su formulación es similar a la de Page Rank:

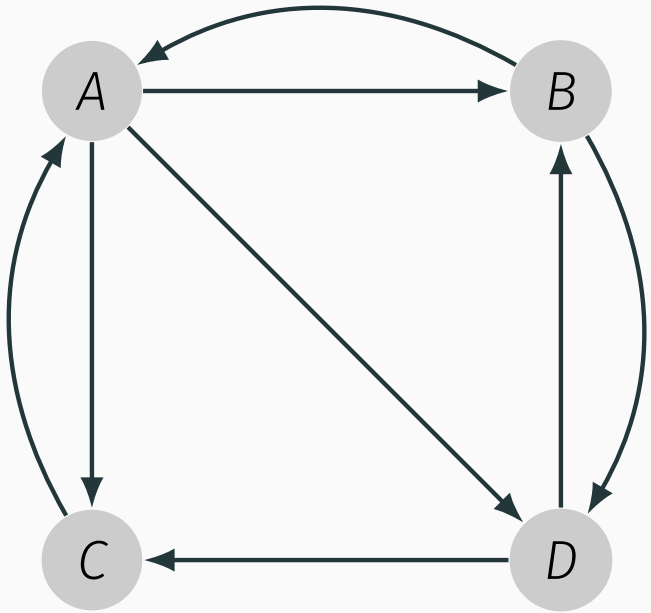
$$v' = \beta M v + (1 - \beta) e_S / |S|$$

dónde:

- β es la probabilidad de elegir un vínculo de forma aleatoria
- M es la matriz de adyacencia
- v es el vector de Page Rank
- S indica la páginas que pertenecen a cierto tópico
- e_S es un vector que tiene 1s en los componentes S y 0s en el resto.
- $|S|$ es el tamaño del conjunto S.

Ejemplo

Calcular el Page Rank sensible al tópico, dónde β = 0.8 y $S = \{B, D\}$

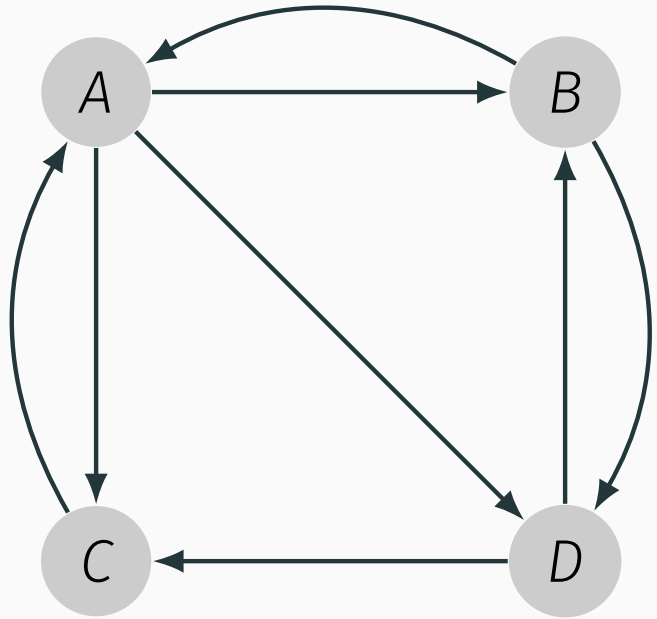


Ejemplo

$$v' = \beta M v + (1 - \beta) e_S / |S|$$

Paso 1: matriz de adyacencia

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



Ejemplo

$$v' = \beta M v + (1 - \beta) e_S / |S|$$

Paso 2: matriz de adyacencia * β

$$\begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} * 0.8 = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix}$$

Ejemplo

$$v' = \beta M v + (1 - \beta) e_S / |S|$$

Paso 3: resolver $(1 - \beta) e_S / |S|$

$$(1 - 0.8) \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} /2 = \frac{1}{5} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} /2 = \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

Ejemplo

$$v' = \beta M v + (1 - \beta) e_S / |S|$$

Unimos los resultados previos

$$v' = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} v + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

Ejemplo

Cálculo de las primeras iteraciones: t_0

$$t_0 = \begin{bmatrix} 0 \\ 1/2 \\ 0 \\ 1/2 \end{bmatrix}$$

Recordemos, solo aplica en los nodos del conjunto S

Ejemplo

Cálculo de las primeras iteraciones: t_1

$$t_1 = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} v + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$t_1 = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1/2 \\ 0 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix} = \begin{bmatrix} 1/5 \\ 3/10 \\ 1/5 \\ 3/10 \end{bmatrix}$$

Ejemplo

Iteraciones

$$t_0 = \begin{bmatrix} 0/2 \\ 1/2 \\ 0/2 \\ 1/2 \end{bmatrix}, t_1 = \begin{bmatrix} 1/5 \\ 3/10 \\ 1/5 \\ 3/10 \end{bmatrix}, t_2 = \begin{bmatrix} 42/150 \\ 41/150 \\ 25/150 \\ 41/150 \end{bmatrix}, t_3 = \begin{bmatrix} 62/250 \\ 71/250 \\ 46/250 \\ 71/250 \end{bmatrix} \cdots \begin{bmatrix} 54/210 \\ 59/210 \\ 38/210 \\ 59/210 \end{bmatrix}$$

¿Cómo integrar TSPR al buscador?

1. Decidir sobre los tópicos para crear vectores de Page Rank especializados
2. Encontrar una manera de determinar el tópico o los tópicos que sean más relevantes
3. Usar los vectores de Page Rank de esos tópicos para responder la consulta del usuario.

¿Cómo identificar los tópicos?

- Permitir que el usuario los seleccione (usando un menú)
- Inferir los tópicos usando:
  - Las búsquedas previas del usuario.
  - La información del usuario (marcadores, Facebook).

Inferir tópicos basado en palabras

- Ejemplo: las palabras sarampión y gol aparecen frecuentemente en las páginas web:
  - Sarampión $---> T_{medicina}$
  - Gol $---> T_{deportes}$
- Supongamos que identificamos las palabras más frecuentes de cada página.
- Supongamos que tomamos un conjunto de páginas especializadas de un cierto tópico, y extraemos las palabras más frecuentes.

Inferir tópicos basado en palabras

- Sea $S_1, S_2 ... S_k$ son el conjunto de palabras que definen cada tópico.
- Sea P el conjunto de palabras que aparecen en una página p.
- Calcular la medida de similitud de Jaccard entre P y en cada uno de S_i.
- Clasificar la página al tópico con mayor similitud.

Inferir tópicos basado en palabras

computer vision	
Computers	0.24
Business	0.14
Reference	0.09

cruises	
Recreation	0.65
Regional	0.18
Sports	0.04

death valley	
Regional	0.28
Society	0.14
News	0.10

gardening	
Home	0.63
Shopping	0.14
Regional	0.04

graphic design	
Computers	0.36
Business	0.23
Shopping	0.09

gulf war	
Society	0.21
Kids & Teens	0.18
Regional	0.17

java	
Computers	0.53
Games	0.10
Kids & Teens	0.06

lipari	
Home	0.19
Kids & Teens	0.17
News	0.13

lyme disease	
Health	0.96
Regional	0.01
Recreation	0.01

national parks	
Regional	0.42
Recreation	0.16
Kids & Teens	0.09

parallel architecture	
Computers	0.70
Science	0.10
Reference	0.07

recycling cans	
Home	0.42
Business	0.38
Kids & Teens	0.06

Imagen tomada de Haveliwala, 2003.