Análisis de Vínculos

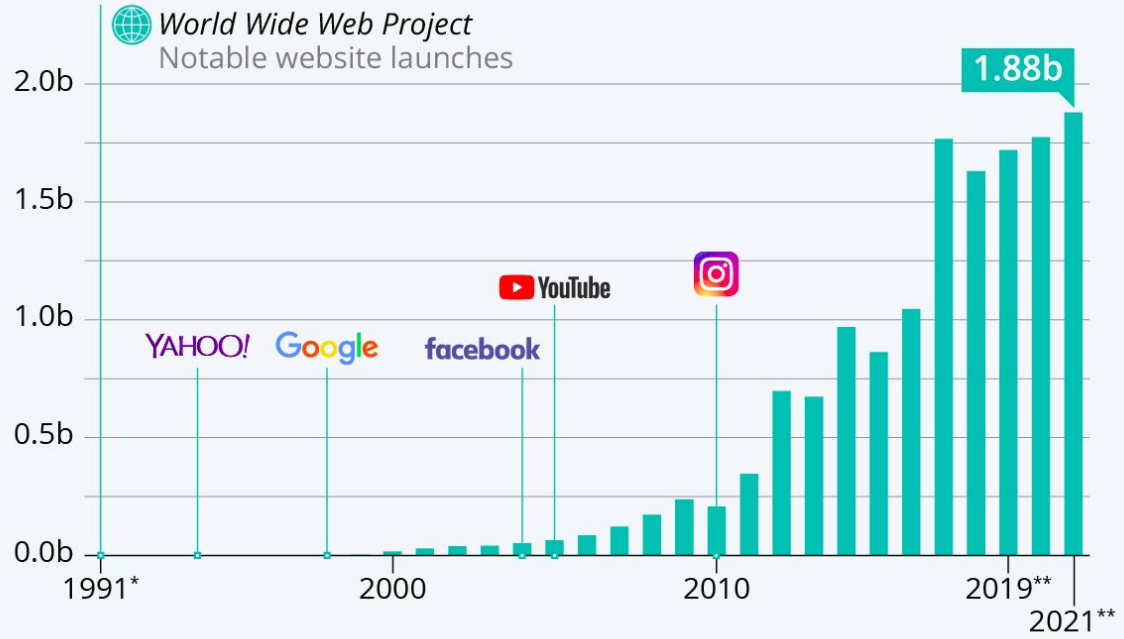
Alondra Berzunza

Page Rank Sensible al Tópico

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

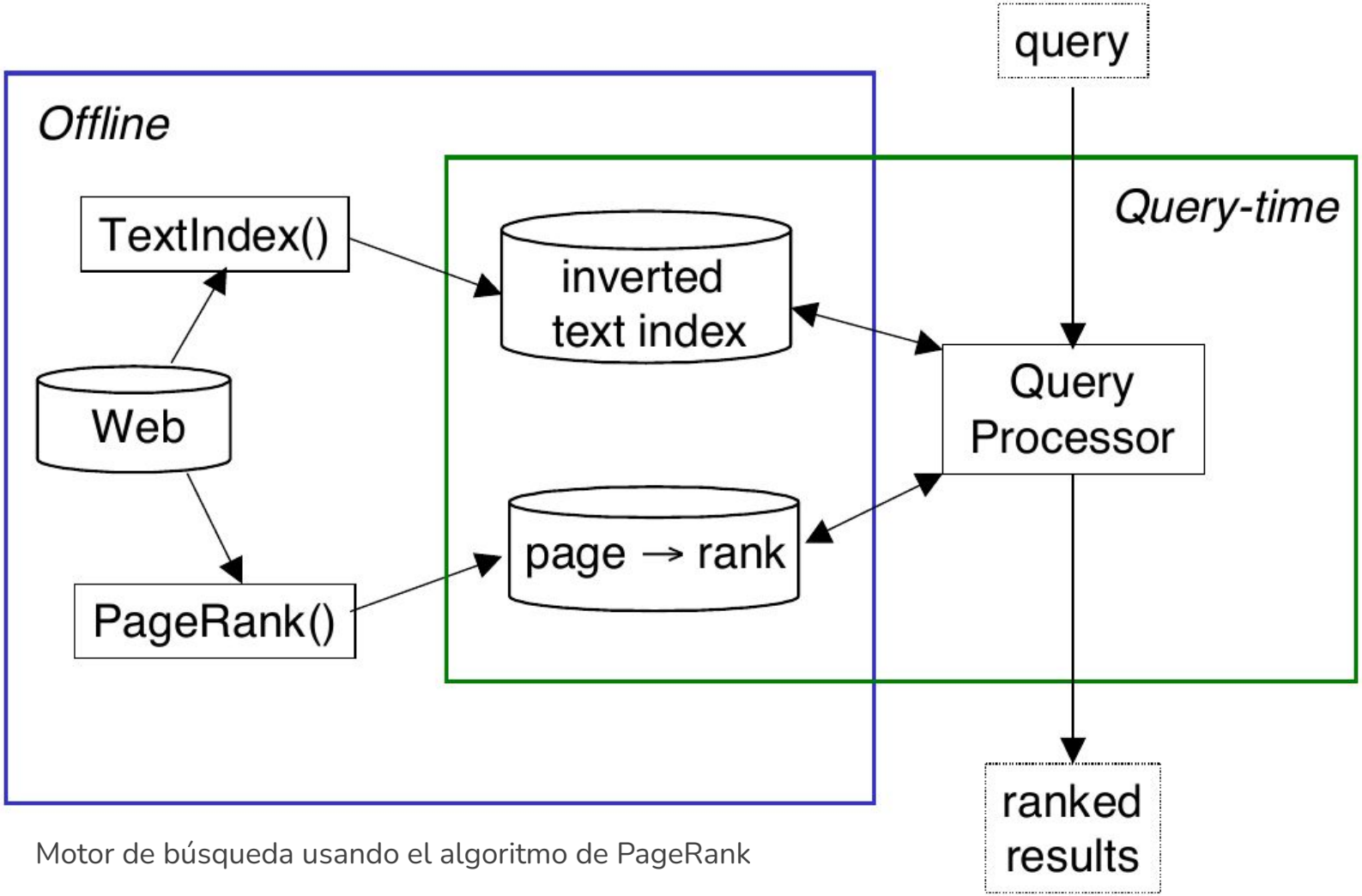
How Many Websites Are There?

Number of websites online from 1991 to 2021



* As of August 1, 1991.
** Latest available data for 2019: October 28, for 2020: June 2, for 2021: August 6.
Source: Internet Live Stats

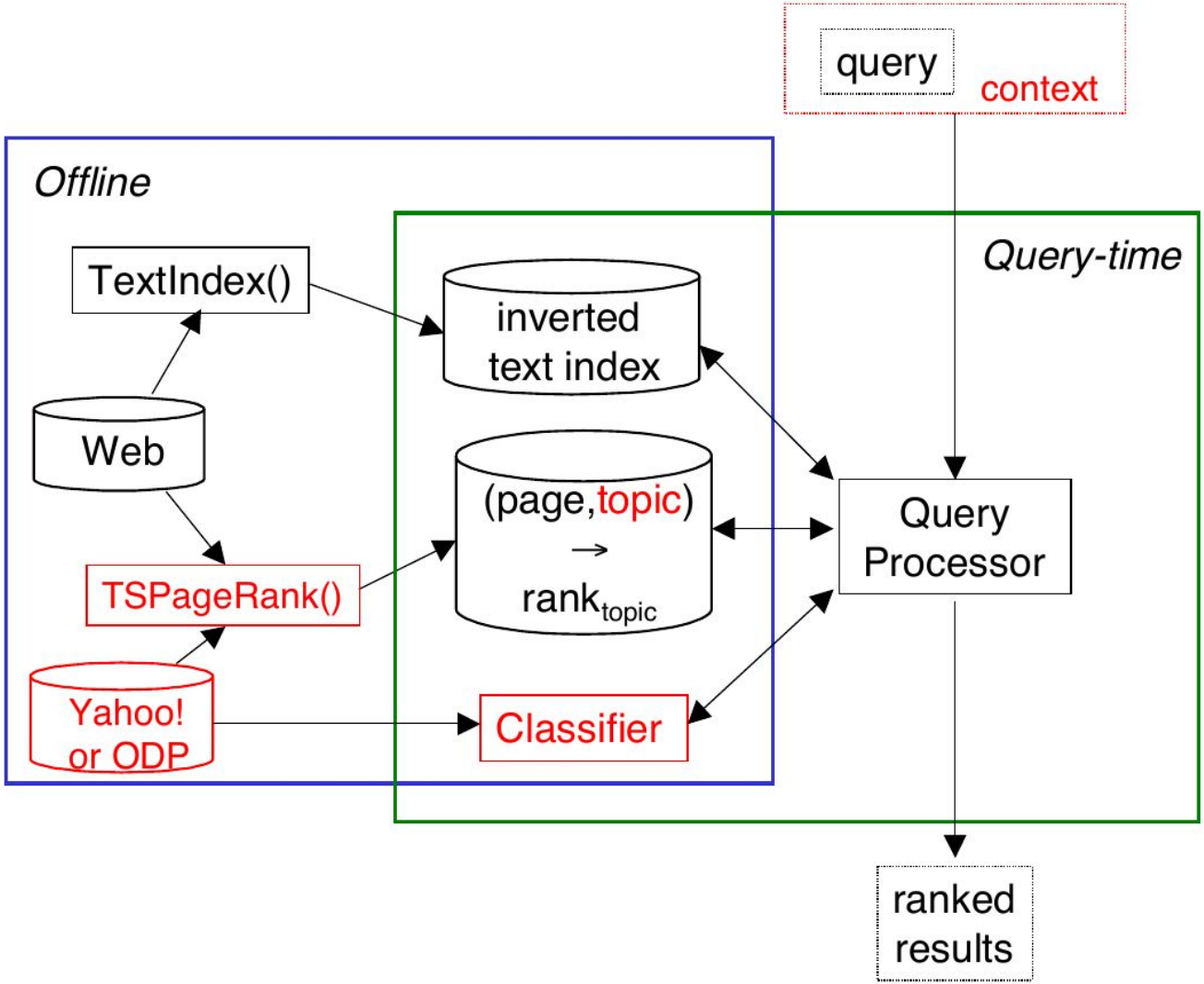
	HITS	Page Rank
Ventajas	★ Simple e iterativo. ★ Puntuación específica de la consulta.	★ Poco costoso (en tiempo de ejecución). ★ Las puntuaciones se calculan utilizando el grafo completo.
Desventajas	➔ Costoso (tiempo de ejecución). ➔ Las puntuaciones se calculan utilizando un subgrafo a partir de todo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



Motor de búsqueda usando el algoritmo de PageRank

Page Rank 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)?



Motor de búsqueda usando el algoritmo de PageRank sensible al tópico.

TSPR

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.

Una variante de PageRank donde la importancia está sesgada por los vectores de temas, mejorando las clasificaciones contextuales.

Caminatas Aleatorias Sesgadas

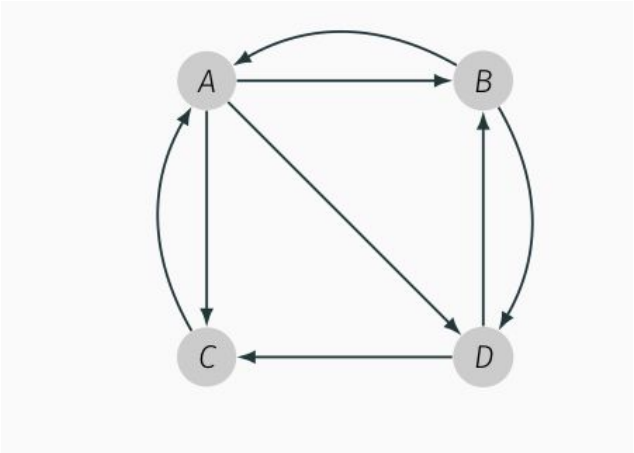
Su formulación es similar a la de Page Rank:

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 esto.
- |S| es el tamaño del conjunto S.

Ejemplo

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



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.

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

Identificamos las palabras más frecuentes de cada página.

Tomamos un conjunto de páginas especializadas de un cierto tópico, y extraemos las palabras más frecuentes.

Inferir Tópicos Basado en Palabras

- Sea S1 , S2 ...Sk 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 cada uno de Si .
- 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