

UNIDAD 3: ANÁLISIS DE VÍNCULOS

ALGORITMO HITS

Blanca Vázquez y Gibran Fuentes-Pineda

26 de septiembre de 2022



🕒 UNAM







- 1960: surgen los inicios de la recuperación automática de información (antes de la creación de WWW)
- Se diseñó para buscar en los repositorios artículos, documentos legales basados en **palabras claves**.


Retos


- Son limitadas
- Son cortas
- No expresivas
- Problemas de sinonimia
- Problemas de polisemia


RETOS EN LAS PALABRAS CLAVES





 Todo

 Maps

 **Imágenes**


 Noticias


 Videos


 Más


Ajustes


Herramientas




 madera


 dinero


 animado


 ahorro


 trabajo





 Todo

 **Imágenes**

 Videos


 Maps

 Noticias

 Más


Ajustes


Herramientas





Detrás de la naturaleza del jaguar - Funda...
fundacioncarlosslim.org

Búsquedas relacionadas


 jaguar f type 2020

 jaguar negro

 jaguar 2020 deportivo



Home Page | Jaguar Mexico
jaguarmexico.com.mx



Jaguar | WWF
wwf.org.mx

- 1980: la recuperación automática de información se convirtió en pieza importante para los bibliotecarios, abogados de patentes... realizaban consultas *efectivas / complejas* para la búsqueda de documentos.
 - Vocabularios específicos
 - Estilos

World Wide Web

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give univers

Everything there is online about W3 is linked directly or indirectly to this document, including an [executiv](#)

[What's out there?](#)

Pointers to the world's online information, [subjects](#) , [W3 servers](#), etc.

[Help](#)

on the browser you are using

[Software Products](#)

A list of W3 project components and their current state. (e.g. [Line Mode](#) ,X11 [Viola](#) , [NeXTStep](#) , [Se](#)

[Technical](#)

Details of protocols, formats, program internals etc

[Bibliography](#)

Paper documentation on W3 and references.

[People](#)

A list of some people involved in the project.

[History](#)

A summary of the history of the project.

[How can I help ?](#)

If you would like to support the web..

[Getting code](#)

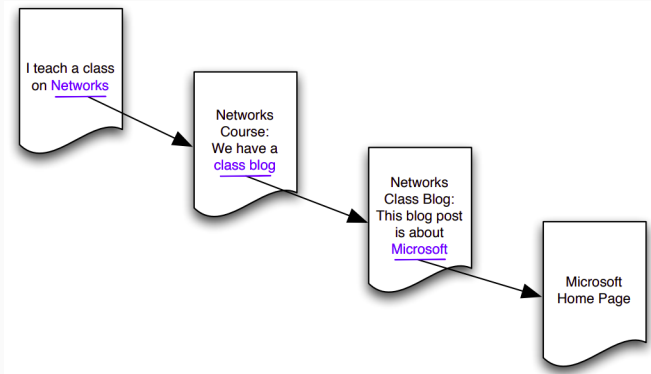
Getting the code by [anonymous FTP](#) , etc.

World Wide Web: creada por Tim Berners-Lee , 1989

De manera simplificada, la concepción de la WWW tenía dos objetivos:

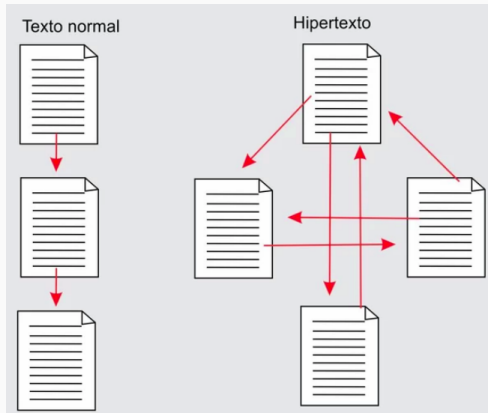
- Compartir información que estuviera disponible para cualquiera
 - A través de la creación de páginas web
- Proporcionar una manera para que todos pudieran acceder a esa información
 - A través de un buscador

SECUENCIAS DE PÁGINAS DENTRO DE UN BUSCADOR



La organización de la información en la web puede verse como un grafo dirigido.
Imagen tomada de Easley, 2010.

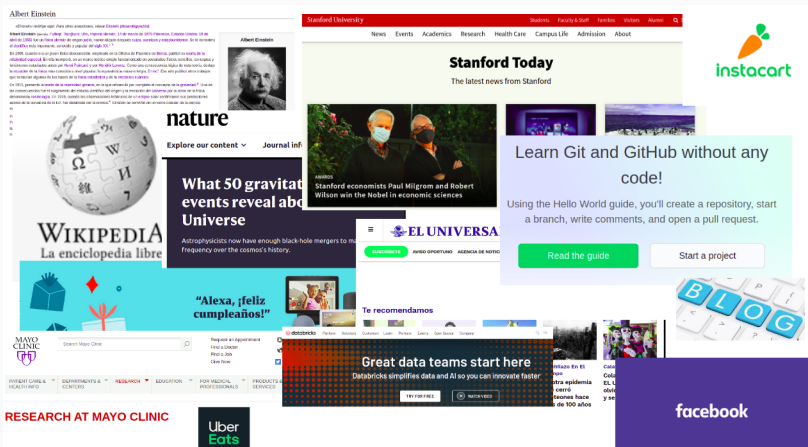
HIPERTEXTO



La idea del hipertexto es reemplazar una estructura lineal de texto hacia una estructura de red.

Imagen tomada de Pimentel, 2011

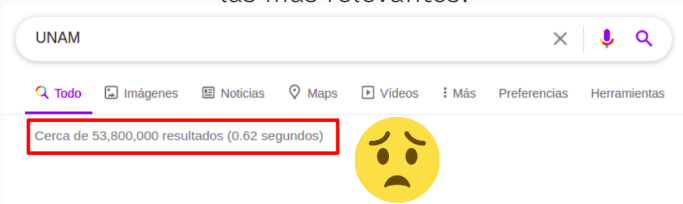
EXPANSIÓN Y CRECIMIENTO DE LA WEB



Retos

- No es posible usar las técnicas tradicionales de recuperación de información
- Crecimiento constante (no. páginas)
- Cantidad y tipo de contenido (audio, vídeo, imágenes, texto)
- Discrepancia: entre hechos que sucedían al momento vs historia
- Pasamos de la escasez a la abundancia.

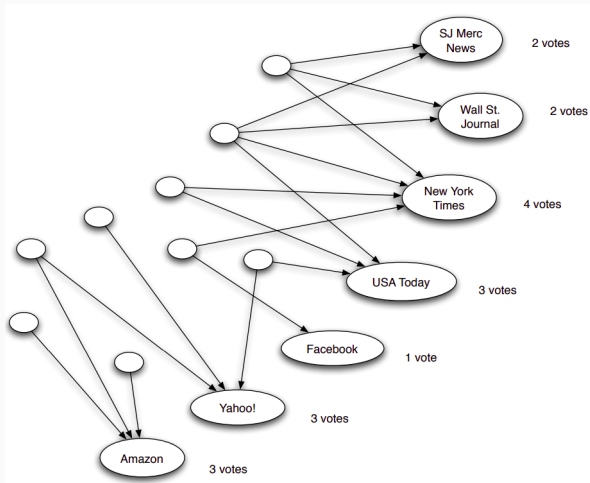
¿Cómo filtrar de un conjunto de millones de páginas,
las más relevantes?



- La importancia de una página no se decide únicamente por las características internas de la página
- Su 'calidad' puede ser juzgada a partir de los enlaces que apuntan a la página.
- Los enlaces son un respaldo colectivo
- Se llama respaldo colectivo cuando una página recibe enlaces de otras páginas relevantes.
- Los enlaces serán claves para la relevancia de una página (críticas, anuncios pagados)

- **Contar los votos / enlaces:** es un tipo de medida simple para descubrir la relevancia de una página web.

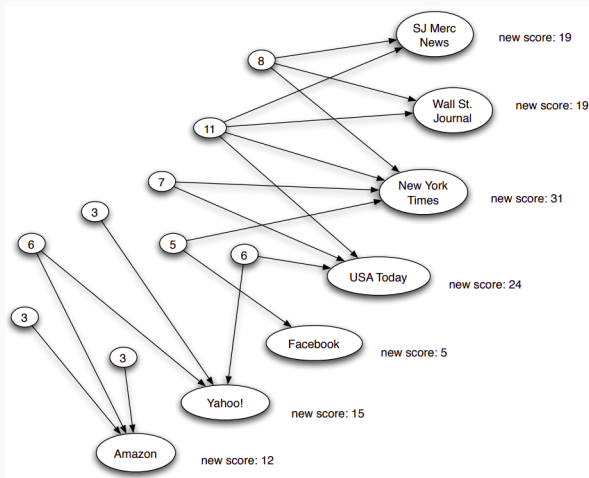
CONTEO DE VOTOS



Contando enlaces de páginas para la consulta *newspapers*.

Imagen tomada de Easley, 2010.

BÚSQUEDA EN LISTAS

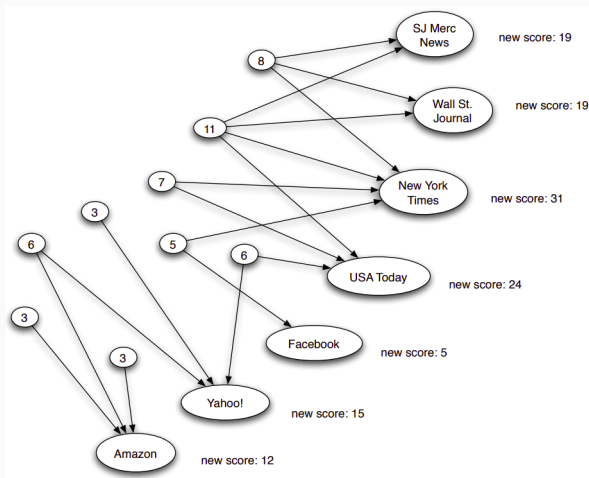


El valor de una página como lista es igual a la suma de los votos recibidos por todas las páginas.
Imagen tomada de Easley, 2010.



Uso de listas en nuestra vida diaria

PRINCIPIO DE MEJORA REPETIDA

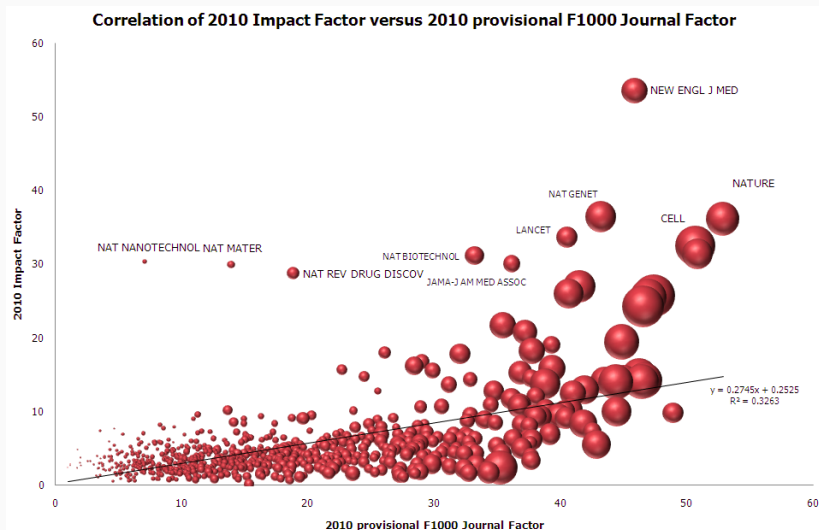


Cada refinamiento de un lado de la figura permite un refinamiento adicional al otro lado.

Imagen tomada de Easley, 2010.

- HITS es el acrónimo de *Hypertext Induced Topic Selection* conocido como el algoritmo de *Hubs* y *autoridades*
- Desarrollado por Jon Kleinberg.
- Es un algoritmo de análisis de enlaces web para descubrir y clasificar las páginas relevantes a partir de una búsqueda.

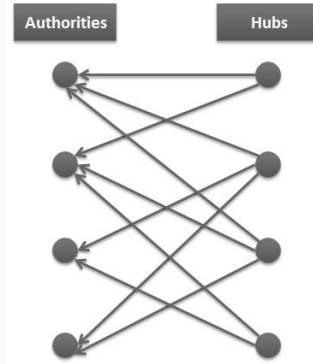
MOTIVACIÓN



HITS está inspirado en el método de clasificación de las revistas académicas.
Imagen tomada de 2011 Journal Citation Reports.

LA IDEA BÁSICA DEL ALGORITMO HITS

La importancia de una página web se mide por 2 indicadores: el valor de autoridad (*Authority*) y el valor de hub (*Hub*)



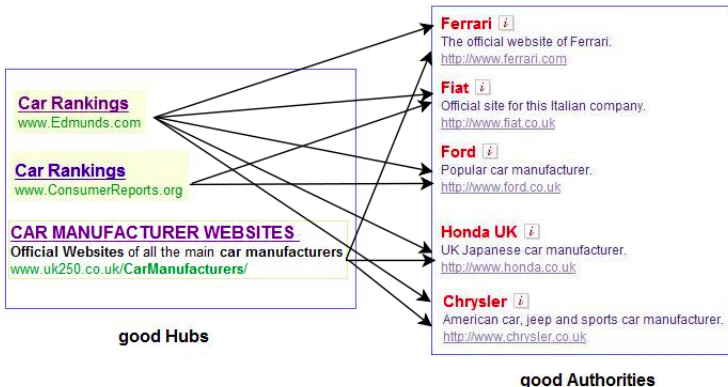
Indicadores de importancia: *hub* y *authority*
Imagen tomada de Hussain,2019.

Dada una consulta en un buscador:

- Las páginas *Hubs* son aquellas que no aportan mucha información sobre un tema, pero enlazan a otras que si lo hacen.
- Las páginas *Authority* son aquellas que aportan mucha información sobre un tema y por ello muchas páginas *Hubs* la enlazan.

- Una buena página *Hub* es aquella que apunta a muchas páginas de autoridad.
- Una buena página *Authority* es aquella que es apuntada por muchas páginas hub.
- Toda página tiene dos indicadores: uno de Hub y uno de autoridad.
- Ambos indicadores son interdependientes y se influyen mutuamente.

EJEMPLO DE HUBS Y AUTORIDADES



Ejemplos de hubs: blogs, foros, sitios de renta

Ejemplo de autoridad: sitios oficiales de fabricantes de coches

Imagen tomada de Cornell, 2009.

$$auth(p) = \sum_{i=1}^n hub(i)$$

dónde n es el número total de páginas enlazadas a p e i es una página conectada a p .

- Por lo tanto, $auth(p)$ es la suma de todas las puntuaciones de hub de las páginas que apuntan a ella.

$$hub(p) = \sum_{i=1}^n auth(i)$$

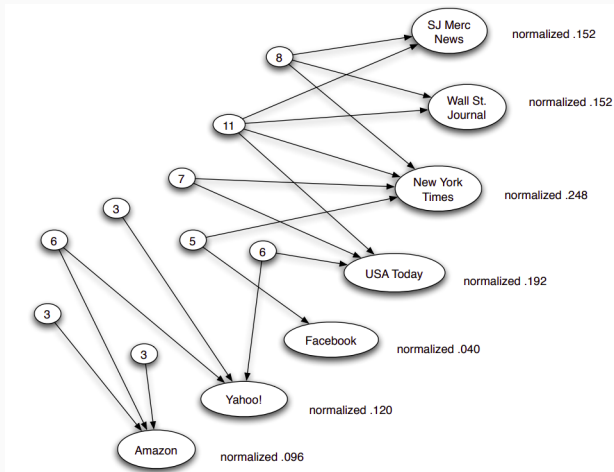
dónde n es el número total de páginas enlazadas desde p e i es una página conectada desde p .

- Por lo tanto, $hub(p)$ es la suma de todas las puntuaciones de $auth$ de todas sus páginas de enlace

El cálculo de los indicadores de *auth* y *hub* se realiza a través de un algoritmo de k iteraciones.

- Debido a que los valores finales de *auth* y *hub* pueden ser divergentes, al final se aplica un proceso de normalización, consiste en:
 - Dividir cada valor final de *auth* de cada página entre la suma total de todos los valores *auth*
 - Dividir cada valor final de *hub* de cada página entre la suma total de todos los valores *hub*

NORMALIZANDO LOS RESULTADOS



1. Sea k el número de iteraciones
2. Cada nodo se asigna a un valor $hub = 1$ y un valor de $auth = 1$
3. Repetimos k veces:

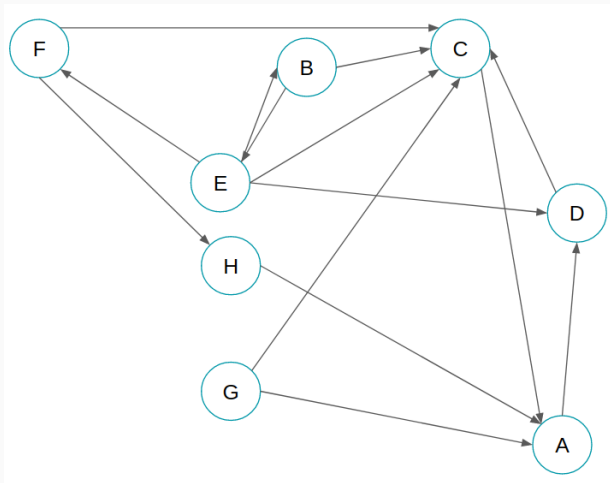
3.1 Actualizamos $auth(p) = \sum_{i=1}^n hub(i)$

3.2 Actualizamos $hub(p) = \sum_{i=1}^n auth(i)$

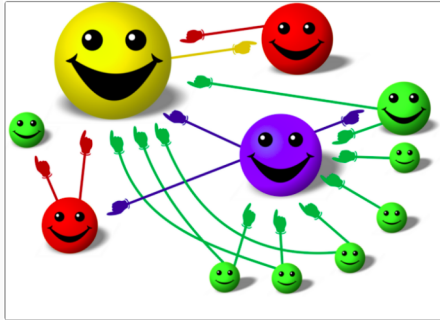
4. Normalizamos* $auth(p)$ y $hub(p)$

EJEMPLO DEL ALGORITMO DE HITS

Calcular el indicador de *auth* y *hub* del siguiente grafo ($k=3$):



RESUMEN DEL ALGORITMO HITS



- Calcula la relevancia de una página a través de *auth* y *hub*.
- Se realiza sobre conjuntos pequeños de páginas.