

Análisis de Vínculos

Alondra Berzunza





HITS



HITS

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.



Recuperación de Información por palabras clave

Retos:

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

Google

banco



Todo

Imágenes

Noticias

Maps

Shopping

Videos

Videos cortos

Más ▾

Herramientas ▾

Guardados



Dibujo



Animado



Icono



Logo



Madera



Dinero



Edificio



Fondo



Recibidor



Jardin



Parque



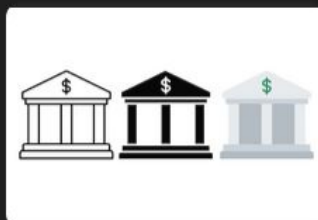
Gob MX
Banco del Bienestar, Sociedad Nacional de Crédito, Institu...



www.banxico.org.mx
Banxico, banco central, Ba...



Lumen · Disponible
Banco de madera Esco ...



Freepik
Imágenes de Bancos - Descarga gr...



Vexels
Diseño PNG Y SVG De...



Vecteezy
arte de línea de constru...



Pixabay



Grup Fábregas



X X



Amazon · Disponible



Alto Nivel



Freepik



Banco de México



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

La World Wide Web (WWW o Web) es un sistema global de información accesible a través de Internet, que permite a los usuarios acceder a documentos y recursos interconectados mediante hipervínculos.

Es un servicio que opera sobre Internet, no la red en sí.

Fue inventada por Tim Berners-Lee en el CERN y publicada en 1991, sentando las bases para el acceso generalizado a la información digital.



World Wide Web

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.



Hipertexto

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

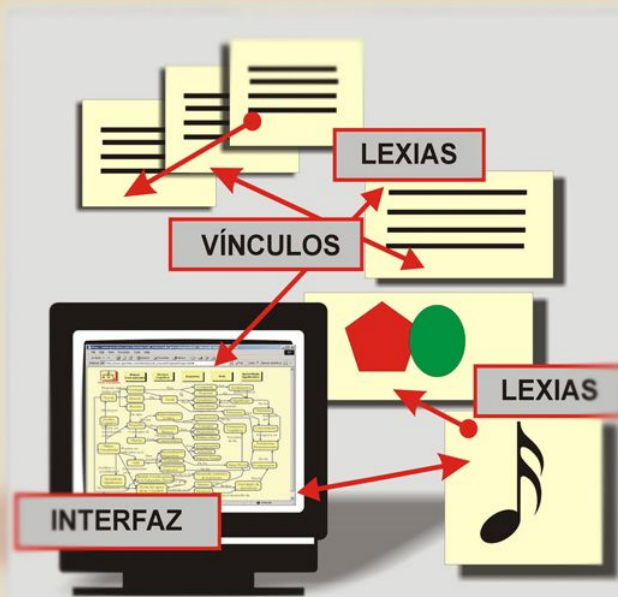
Texto normal



Hipertexto



Hipertexto

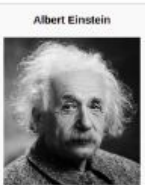


«Einstein recibió así. Para otros acceptance, vision Einstein (disambiguation)

Albert Einstein (sevilla: *Albért*, *Algrinst*; Ulm, Imperio alemán; 14 de marzo de 1879-Princeton, Estados Unidos; 18 de abril de 1955) fue un físico alemán de origen judío, nacionalizado después suizo, suabílo y estadounidense. Se le considera el científico más importante, conocido y popular del siglo XX.¹ 2

En 1905, cuando era un joven físico desconocido, empleado en la Oficina de Patentes de Berna, publicó su teoría de la relatividad especial. En ella incorporó, en un marco teórico simple fundamentado en postulados físicos sencillos, conceptos y fenómenos establecidos antes por Henri Poincaré y por Hendrik Lorentz. Como una consecuencia lógica de esta teoría, dedujo las ecuaciones de la física más conocida a nivel popular: la ecuación más famosa, *E=mc²*. Ese año publicó otros trabajos que sentaron algunas de las bases de la física moderna y de la mecánica cuántica.

En 1915, presentó la teoría de la relatividad general, en la que reformuló por completo el concepto de la gravedad.³ Una de las consecuencias fue el surgimiento del estudio científico del origen y la evolución del Universo por la teoría de la física denominada cosmología. En 1919, cuando las observaciones británicas de un eclipse solar confirmaron sus predicciones acerca de la curvatura de la luz, fue reconocido por la ciencia.⁴ Trasladó su domicilio a un apartamento de la ciudad de Princeton.





WIKIPEDIA

La enciclopedia libre

nature

Explore our content ▾ Journal info

What 50 gravitational events reveal about the Universe

Astrophysicists now have enough black-hole mergers to map frequency over the cosmos's history.



"Alexa, ¡feliz cumpleaños!"





MAYO CLINIC

Request an Appointment

Find a Doctor

Find a Job

Give Now

PATIENT CARE & HEALTH INFO

DEPARTMENTS & CENTERS

RESEARCH

EDUCATION

FOR MEDICAL PROFESSIONALS


PRODUCTS & SERVICES

RESEARCH AT MAYO CLINIC





Stanford Today

The latest news from Stanford



AWARDS
Stanford economists Paul Milgrom and Robert Wilson win the Nobel in economic sciences






Learn Git and GitHub without any code!

Using the Hello World guide, you'll create a repository, start a branch, write comments, and open a pull request.

Read the guide

Start a project




databricks

Platform Solutions Customers Learn Partners Events OpenSource Company


Great data teams start here

Databricks simplifies data and AI so you can innovate faster

TRY FOR FREE WATCH VIDEO







facebook

Calatino
Colo
EL U
olivo
y se

Chilazo En El
mpo
otra epidemia
cerró
teones hace
de 100 años



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.



Expansión y Crecimiento De La Web

UNAM



Todo



Imágenes



Noticias



Maps



Vídeos



Más

Preferencias

Herramientas

Cerca de 53,800,000 resultados (0.62 segundos)



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



Votación por enlaces

- 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).



Votación por enlaces

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



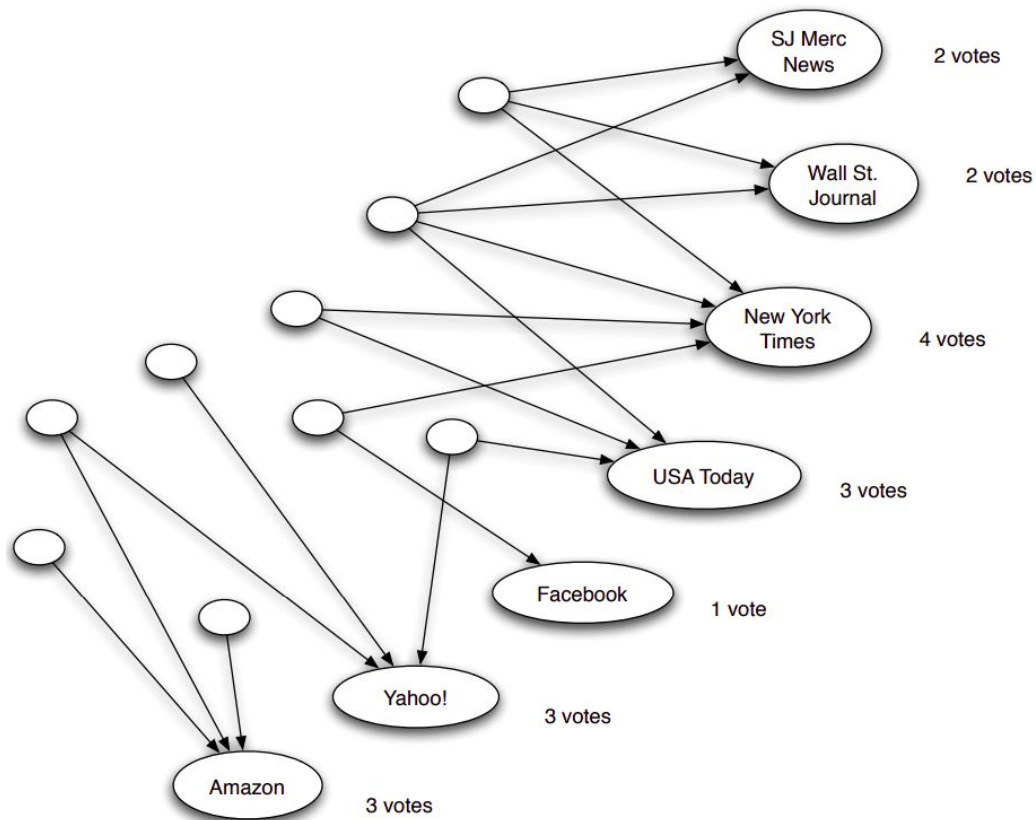
Conteo de Votos

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



Conteo de Votos

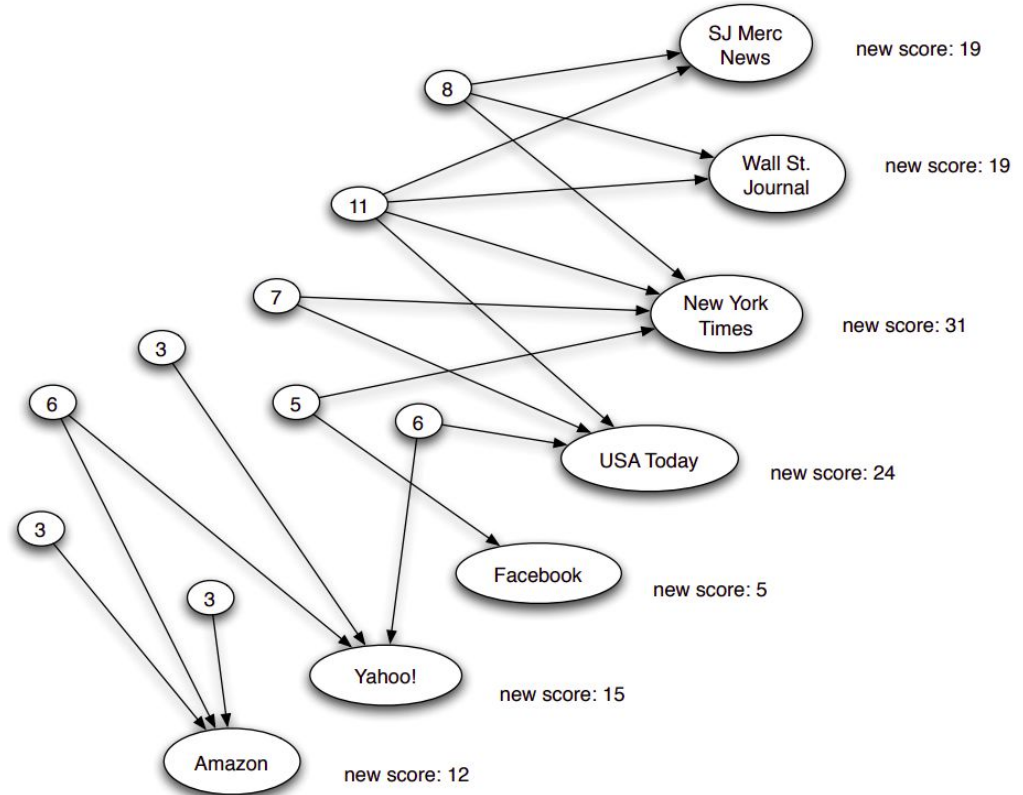
Contando enlaces de páginas para la consulta newspapers.





Conteo de Votos

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





Algoritmo de HITS

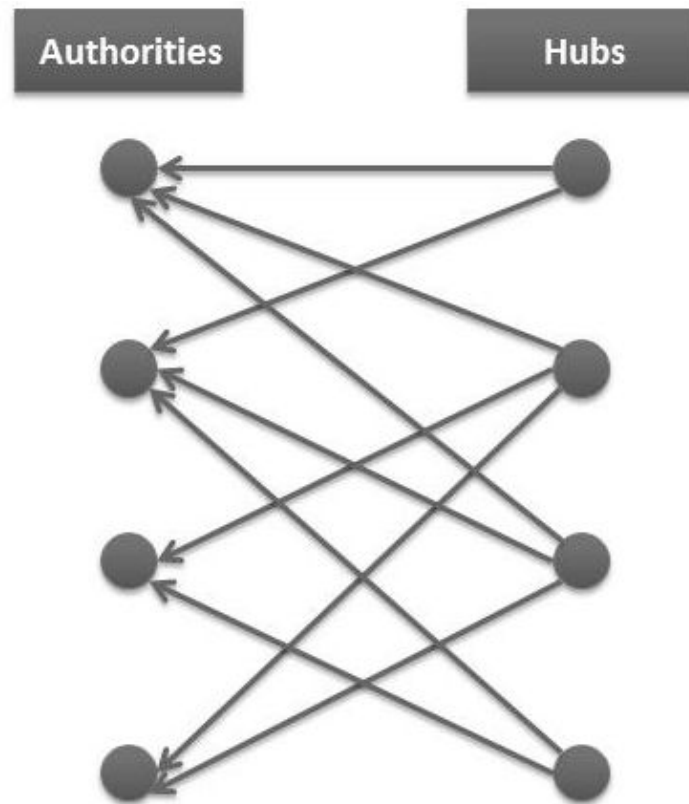
- HITS es el acrónimo de Hypertext Induced Topic Selection conocido como el algoritmo de Hubs y autoridades.
- Desarrollado por Jon Kleinberg.
- Es un algoritmo de análisis de enlaces web para descubrir y clasificar las páginas relevantes a partir de una búsqueda.

Surgió del hecho de que un sitio web ideal debería enlazar a otros sitios relevantes y también ser enlazado por otros sitios importantes.



Algoritmo de HITS

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

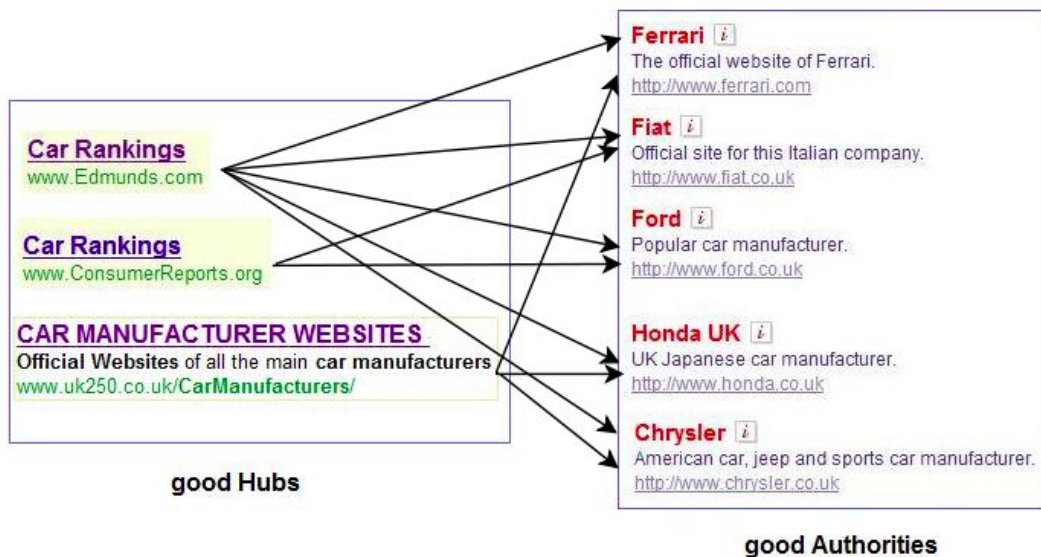




Algoritmo de HITS

- 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.

Algoritmo de HITS



Ejemplos de hubs: blogs, foros, sitios de renta.

Ejemplo de autoridad: sitios oficiales de fabricantes de coches.



Algoritmo de HITS

Indicador de autoridad.

- n es el número total de páginas enlazadas a p
- 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.



Algoritmo de HITS

Indicador de hub.

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

Por lo tanto, $\text{hub}(p)$ es la suma de todas las puntuaciones de auth de todas sus páginas de enlace.

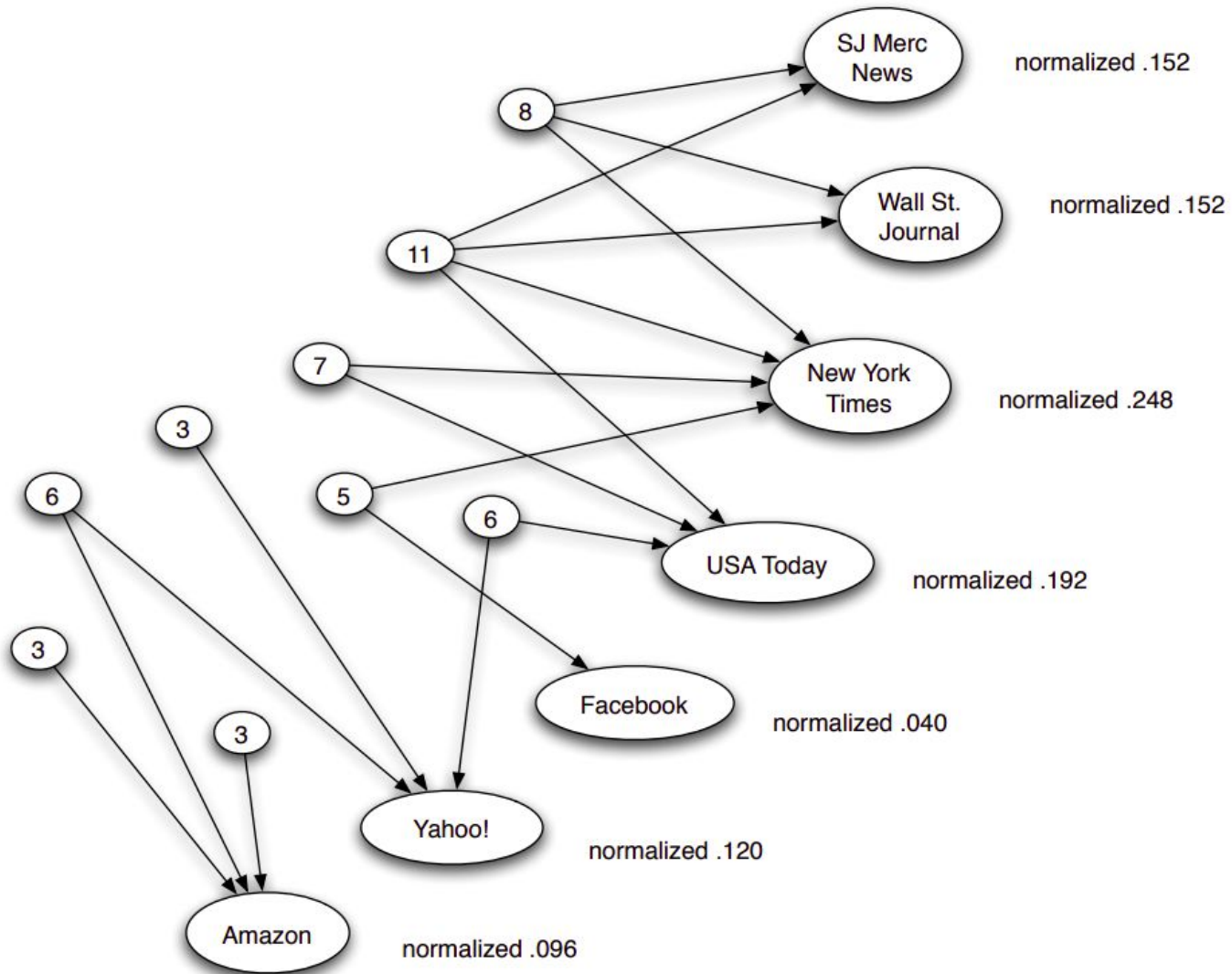


Algoritmo de HITS - Normalización

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.



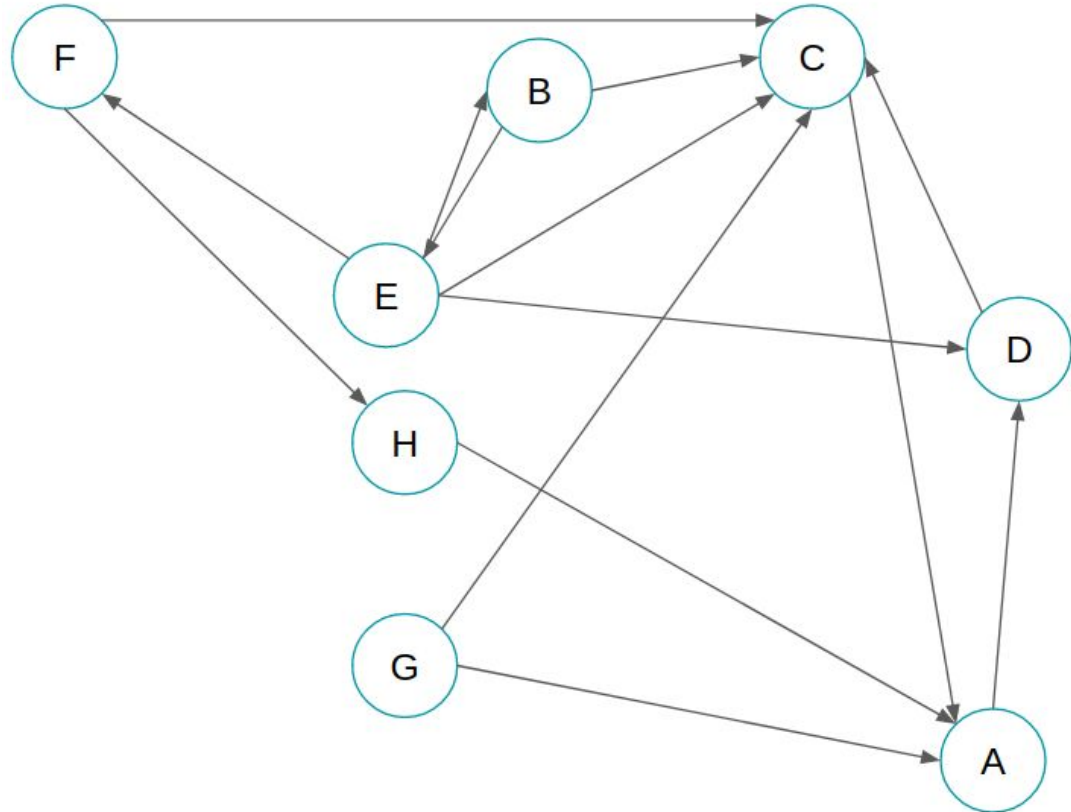


Algoritmo de HITS - Pasos

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:
 - a. Actualizamos $auth(p)$
 - b. Actualizamos $hub(p)$
4. Normalizamos* $auth(p)$ y $hub(p)$.

Algoritmo de HITS - Ejemplo

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



Algoritmo de HITS - Resumen

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

