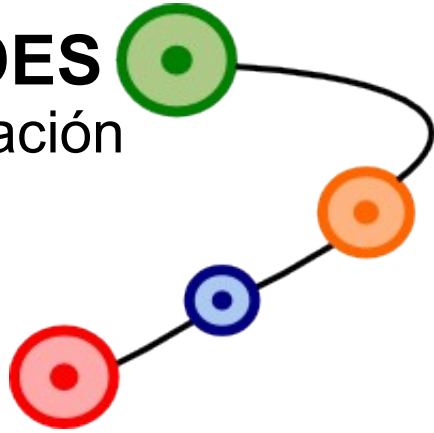
Laboratorio de REDES
Recuperación de Información
y Estudios de la Web



Recuperación de Información en la Web y Motores de Búsqueda

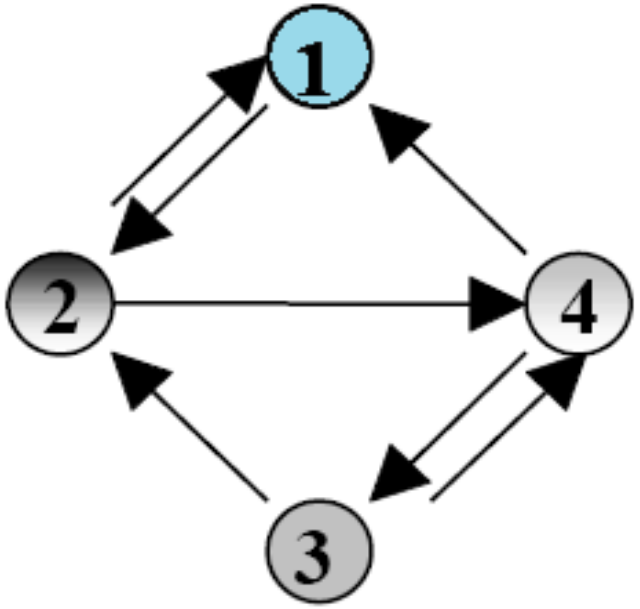
Dr. Gabriel H. Tolosa
tolosoft@unlu.edu.ar

Análisis de Enlaces

El grafo de la web

Se modela la web como un grafo dirigido

- Cada página es un nodo
- Cada hyperlink es un arco dirigido
  - Grado entrante
  - Grado saliente
  -
- Se puede representar mediante la matriz de adyacencia:



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

Estructura

Los enlaces (hyperlinks)

- Representan una relación entre páginas conectadas
- Documento origen -> link

<a href="http://www.unlu.edu.ar">Universidad Nacional de Luján****

- In-links → indegree
- Out-links → outdegree

- Los enlaces son fuente de evidencia, pero también pueden aportar ruido

Estructura

Suposiciones sobre la creación de enlaces

- Recomendación
  El autor recomienda la página destino

- Localidad temática
  Las páginas conectadas tienen mayor probabilidad de ser del mismo tema que las que no lo están.

- "anchor text" descriptor
  El texto del "ancla" describe el destino

Para la indexación:

- (Probablemente) Provea una descripción consisa de la página misma
- (Probablemente) Contenga más términos significativos que la página misma
- Representa el contenido de páginas aún no recolectadas
- Representa objetos no textuales (imágenes, programas, etc.)

Ranking

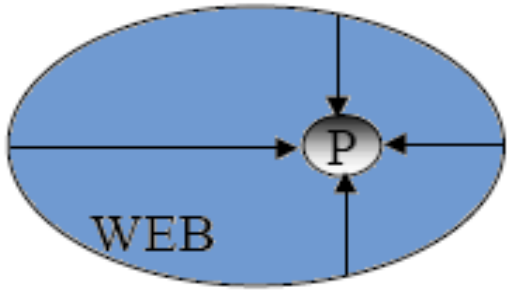
Algoritmos de Ranking

- Ranking basado en contenido
  - Modelos booleano, vectorial, etc.
- Ranking basado en enlaces
  - Mediante el análisis de los enlaces se determina la calidad de la página
  - Clásicos: HITS [Kleinberg] y PageRank [Brin & Pag]
- Combinación de los anteriores

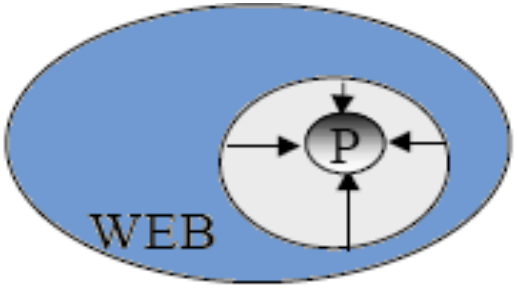
Análisis de enlaces

Algoritmos

- Dos enfoques:
  - Análisis global: la calidad de la página es **independiente** de la consulta



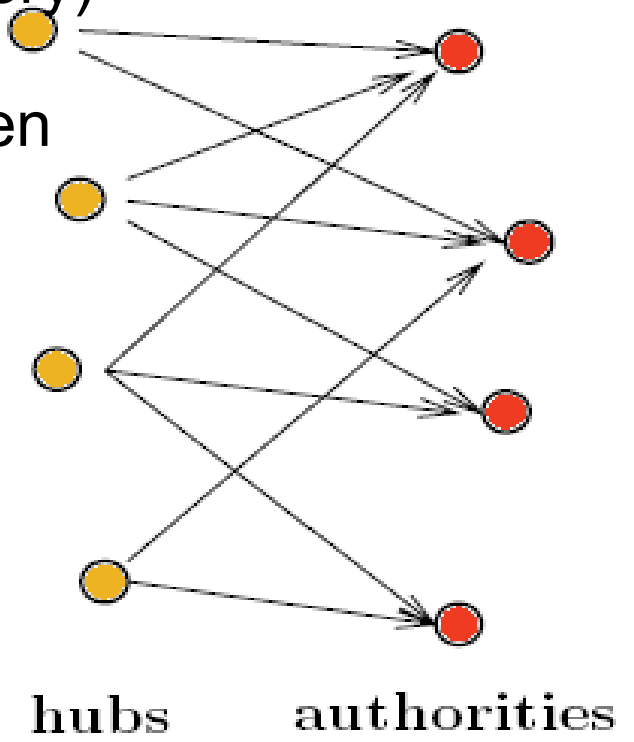
  - Análisis local: la calidad de la página es **dependiente** de la consulta



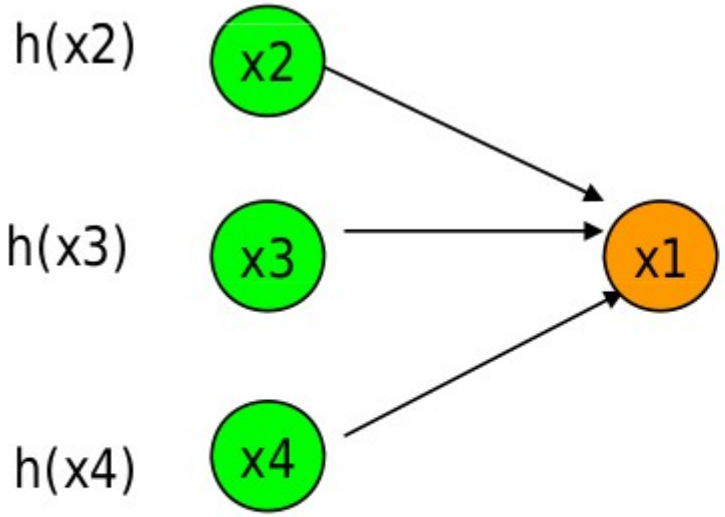
Análisis de enlaces

HITS – Hypertext Induced Topic Search

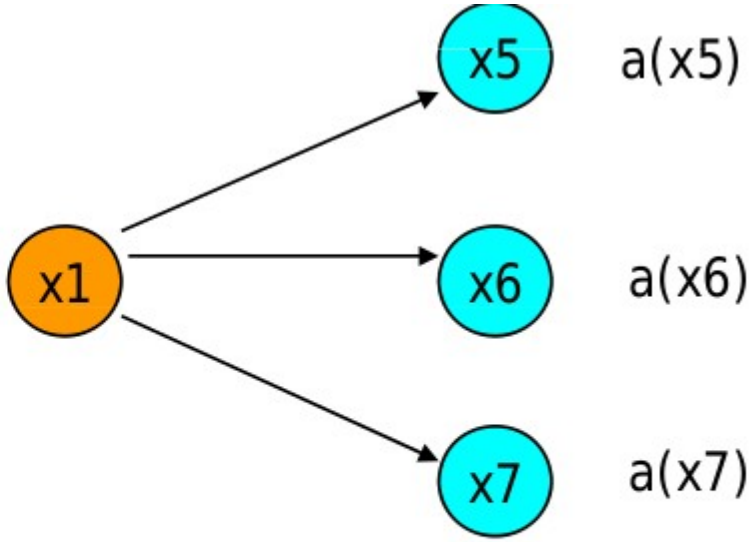
- Kleinberg, 1997.
- Identifica para un tema determinado (Query)
  - Autoridades: Páginas que contienen información relevante respecto de Q.
  - Hubs: Páginas que poseen links salientes (¨apuntan¨) a páginas útiles.
- El valor de autoridad viene de los inlinks
- El valor de hub viene de los outlinks
- Refuerzo mutuo



HITS



$$a(x1) = h(x2) + h(x3) + h(x4)$$

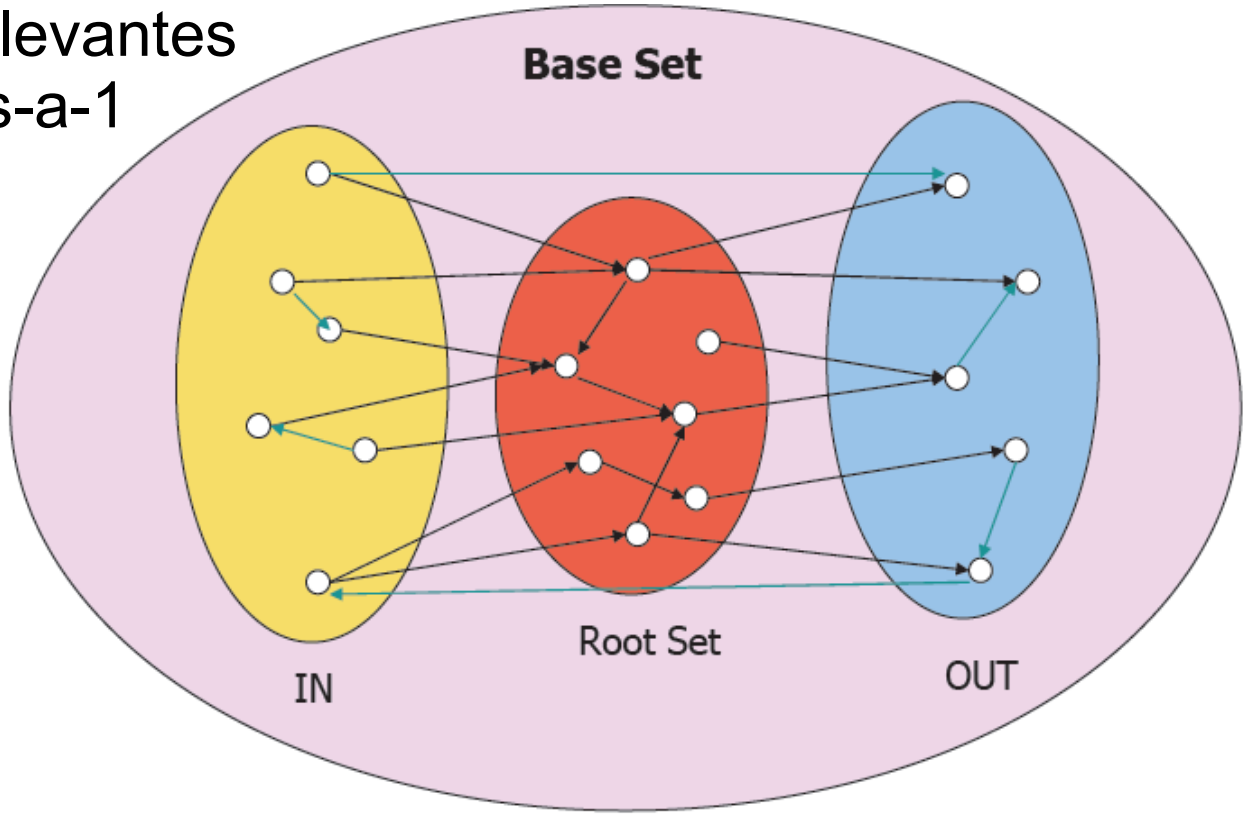


$$h(x1) = a(x5) + a(x6) + a(x7)$$

HITS

Proceso

- Dado un query, identifica:
  - Root set – top k relevantes
  - Base set – vecinos-a-1
- Construye el grafo correspondiente
- Construye la matriz de adyacencia
- Calcula
  - Hub-score
  - Auth-score



HITS

Cómputo de los Scores

- Inicializar todos los pesos a 1

- Repetir hasta converger

#Operación O – los hubs suman los pesos de las autoridades

$$h_i = \sum_{j:i \to j} a_j$$

#Operación I – las autoridades suman los pesos de los hubs

$$a_i = \sum_{j:j \to i} h_j$$

Normalizar los pesos

Análisis de Enlaces

PageRank

- Brin & Page, 1998.

- **Idea**: Una página es importante si otras páginas importantes apuntan a ésta (*Buenas autoridades apuntan a buenas autoridades)*
- Cada link entrante es un voto
- Entonces:

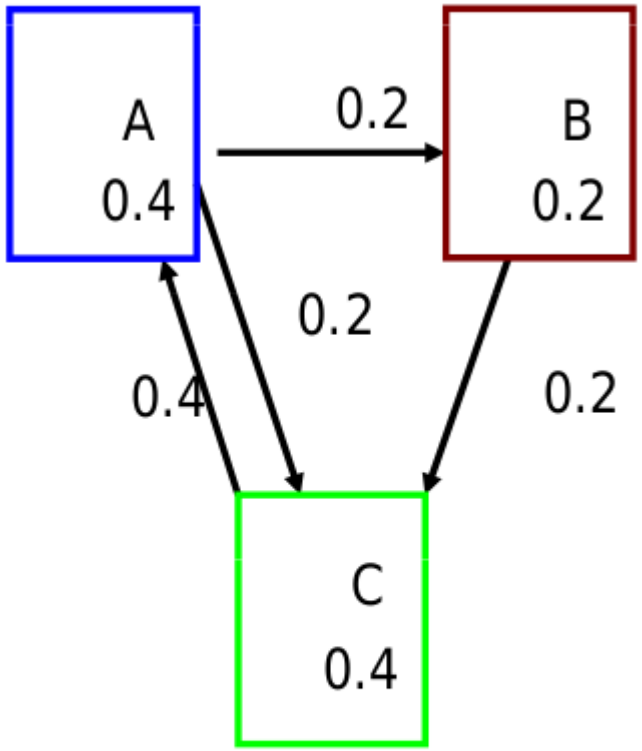
$$r_i = \sum_{j \in L_i} r_j / N_j, \qquad i = 1, 2, \ldots, n.$$

Donde:

$N_j \rightarrow$ #de oulinks de P_j

$L_i \rightarrow$ Páginas que apuntan a P_i

Pagerank



$\mathbf{PR}^k$=[0.4; 0.2; 0.4]

Dado:

El grafo de la web G = (V, E)

$$PR(i) = \sum_{(j,i)\in E} \frac{PR(j)}{O_j}$$

$$H = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

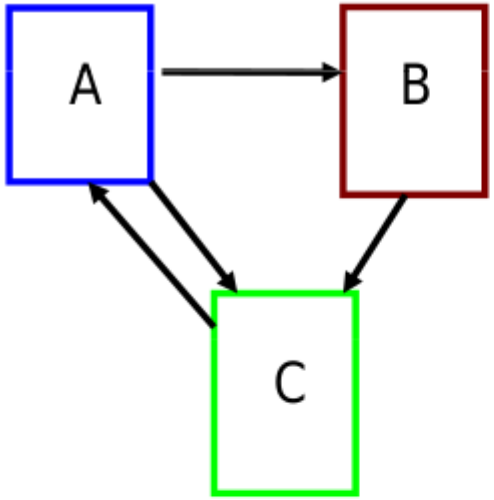
Matriz estocástica!

Pagerank

Hay que resolver

- i.e. $PR^{k+1} = H^T PR^k$

Donde la solución corresponde al autovector del autovalor 1.



$$H = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

PR^0	PR^1	PR^2	PR^3	PR^k
0.33	0.33	0.33	0.42	0.40
0.33	0.17	0.25	0.17	0.20
0.33	0.50	0.42	0.42	0.40

Pagerank: Random Surfing

Cadena de Markov
* Cada página web es un estado
* Cada enlace es una transición de un estado a otro
* Se modela la navegación como un proceso estocástico

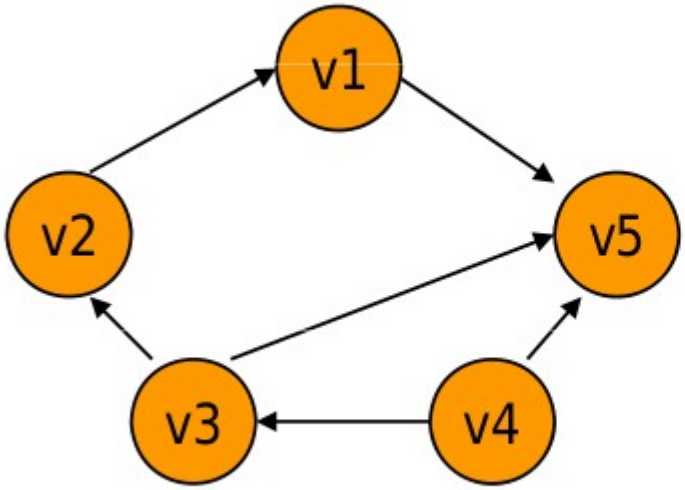
Es el modelo del **"random surfer"**, que elije siempre un link saliente

Cuando se alcanza el estado estacionario (*steady*) resulta:

$$\mathbf{PR^k = PR^{k+1} = PR}$$

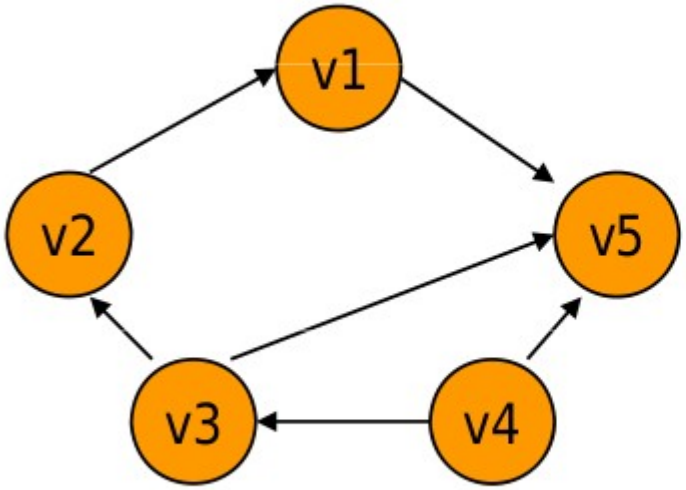
PR es el autovector principal de H^T con autovalor 1

H es estocástica?



$$\boldsymbol{H} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

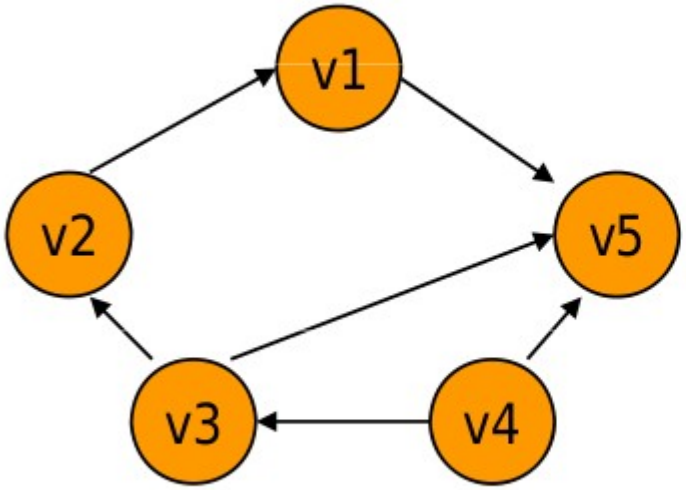
H es estocástica?



$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

dangling pages

H es estocástica?



$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

dangling pages

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}.$$

Solucionando H

Agregar un link de todas las páginas a todas las páginas, asignándole una probabilidad de transición pequeña, controlada por un parámetro d (damping factor).

Luego, la nueva matriz es **irreducible**, ya que cada par de nodos se puede comunicar con cierta P.

$$PR(i) = (1-d) + d\sum_{j=1}^{n} H_{ji} PR(j)$$

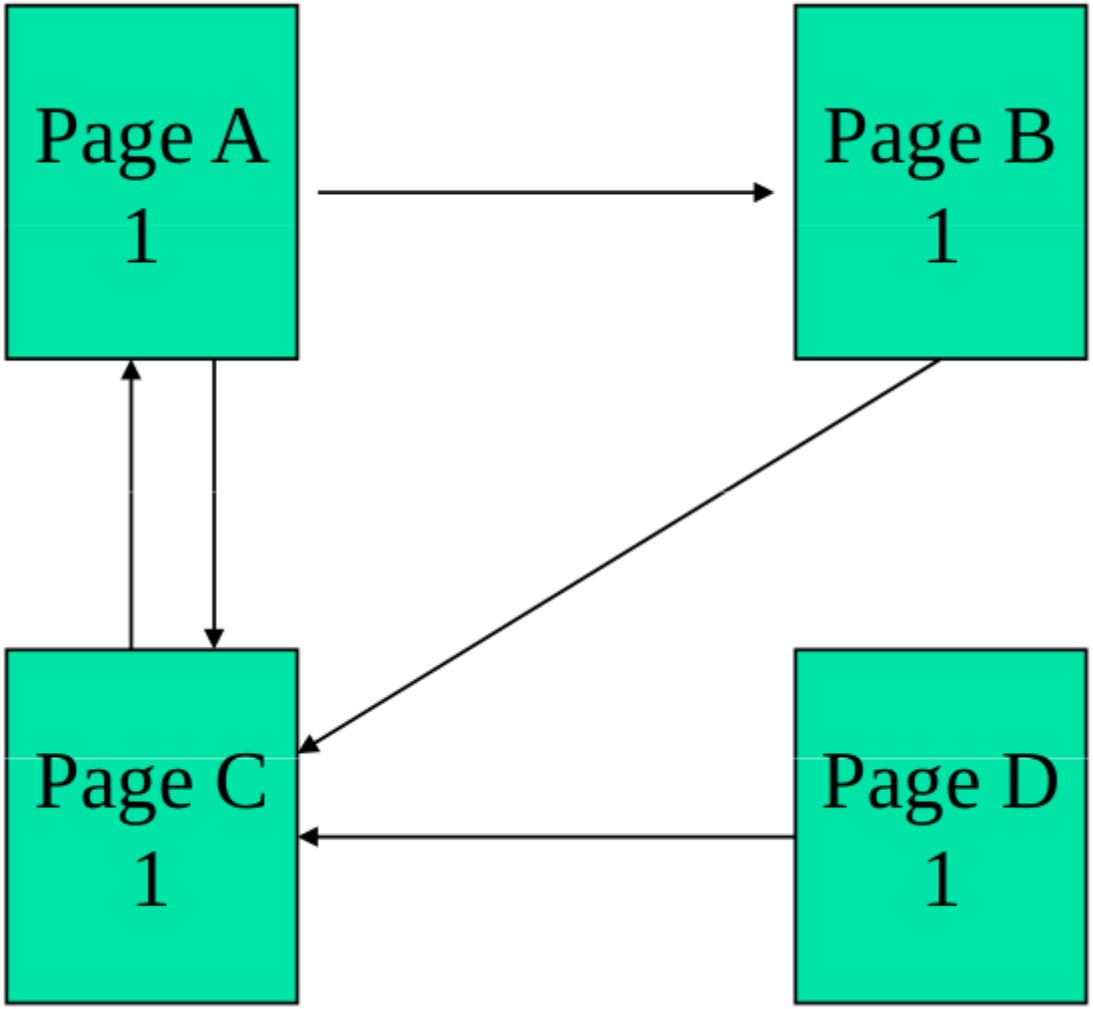
$$= (1-d) + d \sum_{(j,i)\in E} \frac{PR(j)}{O_j}$$

El random surfer tiene dos opciones:

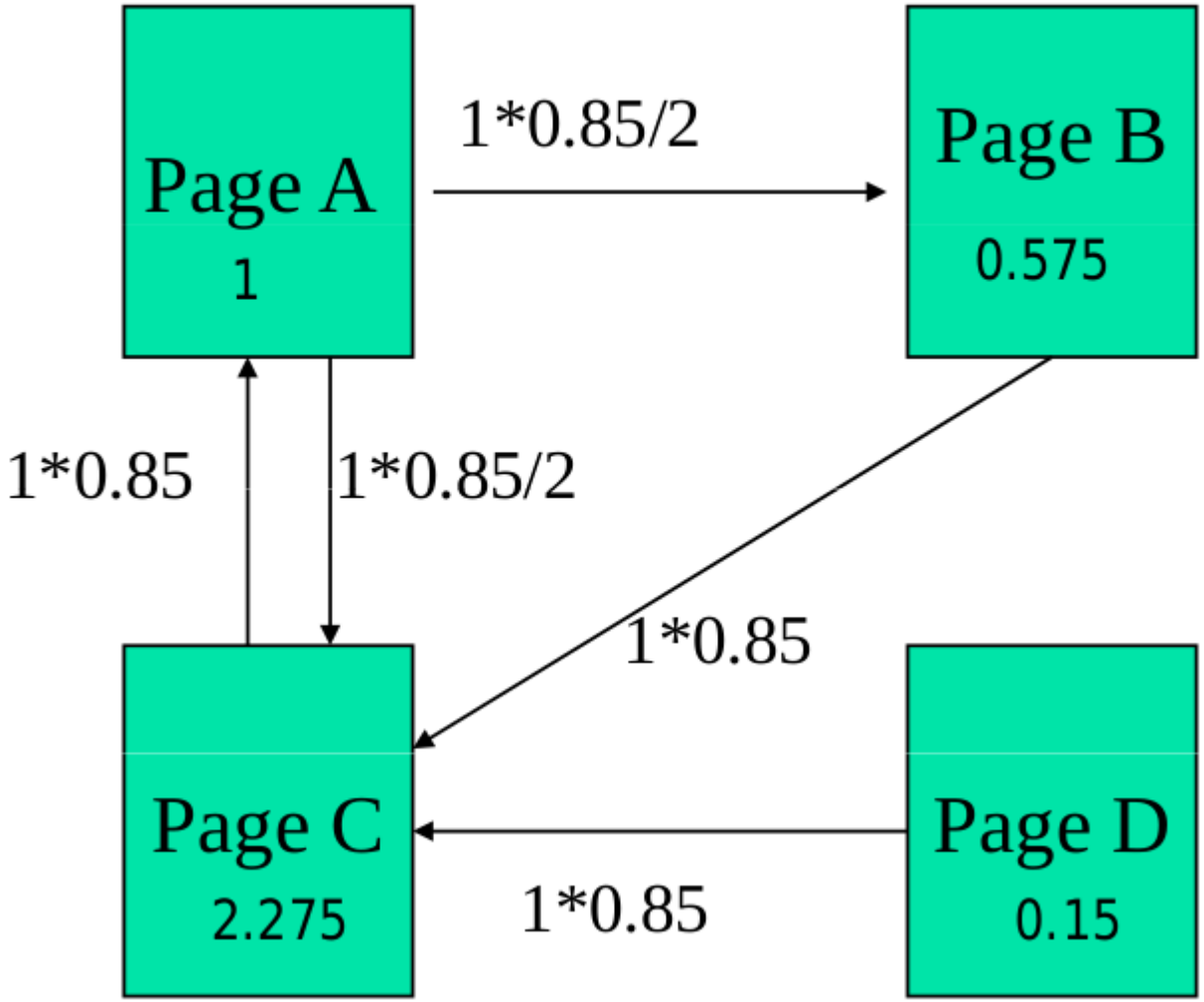
•Con probabilidad d, elige al azar uno de los links salientes de la página.

•Con probabilidad (1 - d), “salta” a una página al azar.

Ejemplo



Ejemplo



Page A:
0.85 (de Page C) +
0.15 (random jump) = **1**

Page B:
0.425 (de Page A) +
0.15 (random jump) = **0.575**

Page C:
0.85 (de Page D) +
0.85 (de Page B) +
0.425 (de Page A) +
+ 0.15 (random jump) = **2.275**

Page D:
Recibe nada + 0.15 = **0.15**

Pagerank

- **Algunas consideraciones:**
  - Actualmente, se considera el problema de cómputo con matrices más grande en el mundo.
  - Las operaciones se realizan sobre matrices de más 20 mil millones de filas/columnas.
  - La matriz es esparcida, la cantidad media de enlaces es 8.
  - Con el factor de damping seteado en 0.15 se requieren aproximadamente 100 iteraciones.
  - La detección de web spam es – en la actualidad – una tarea importante para no sesgar artificialmente los valores.

Funciones de Ranking

- Combinación lineal de todas las “señales” consideradas en el ranking

- Ejemplo con PR:

$$R(p, Q) = \alpha \; BM25(p, Q) \; + \; (1 - \alpha) PR(p)$$

Comparación

PageRank	HITS
Google	CLEVER (IBM)
Independiente del query. Calculado offline para todas las páginas web en el índice.	Dependiente del query. Calculado online para un subconjunto de páginas (Root-set + Base-set)
Calcula sólo un score de autoridad	Calcula dos scores: Hub y Autoridad
Cálculo sobre un grafo muy grande	Cálculo sobre un grafo reducido
Trivial y rápido de calcular (la dificultad es de escala)	Fácil de calcular, pero de ejecución más compleja en tiempo real.
Menos suceptible a ataques de Spam	Más suceptible a ataques de Spam
Más estable	Menos estable, la calidad depende del seed