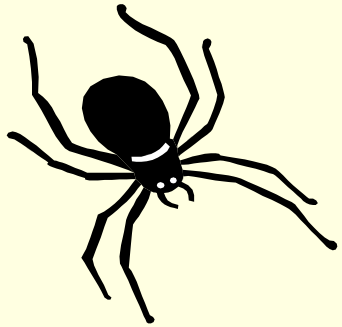
Gestión de Información en la Web
Máster en Ingeniería Informática
Parte II: Recuperación de información y Sistemas de recomendación

Tema 4: Recuperación de información
4.5. RI en la Web

Juan Manuel Fernández Luna
Dpto. Ciencias de la Computación e Inteligencia Artificial
jmfluna@decsai.ugr.es

RI en la Web

- Hasta ahora, las colecciones de documentos estaban disponibles...
- ...Pero en la Web hay que ir a buscar las páginas web que queremos indexar.



Robots

Crawling

- Robots, crawlers, arañas, agente web:
  - Recorren documentos web siguiendo los enlaces de manera recursiva y se los pasan al módulo de indexación.
  - Deciden qué documento visitar.
  - Deciden qué partes del documento mantener.
  - Comprueban que las páginas ya indexadas sigan existiendo.
  - Visitan de nuevo páginas para actualizarlas.

Proceso de captura de páginas

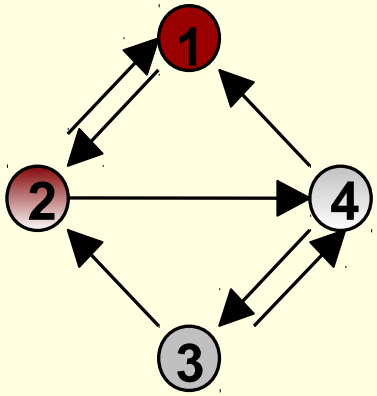
1. Inicializar la una cola con un conjunto de semillas.
2. Obtener el URL del tope de la pila y descargar la página.
3. Analizar la página para encontrar otros URLs.
4. Descartar aquellos URLs tales que no puedan ser analizados fácilmente (.exe, .jpg, ...) o hayan sido vistos antes.
5. Añadir los URLs a la cola: estrategias de búsqueda primero en anchura o primero en profundidad.
6. Continuar al paso 2 mientras no sea momento de parar.

Características de los doc. Web

- Distribuidos
- Volátiles
- Gran volumen.
- No estructurado.
- Redundantes
- Baja calidad.
- Diferentes idiomas.
- Heterogéneos: diferentes tamaños y tipos.
- Basura.

Características de los doc. Web

- Existencia de enlaces.
- Se puede ver la Web como un grafo dirigido:
  - Cada página es un nodo.
  - Cada enlace es un arco.



Enlaces

- Representan relaciones entre páginas web conectadas.
- Elementos:
  - Página web conteniendo el enlace:
    - Web de la UGR
  - Texto del enlace importante: “Web de la UGR”.
  - Página referenciada: www.ugr.es
  - Enlace entrante de una página p: enlace de otra página a p.
  - Enlace saliente de una página p: enlace de p a otra página.

Enlaces

Suposiciones en el uso de enlaces:

1) Enlazando una página, el autor la recomienda.
2) Páginas enlazadas tratan probablemente sobre la misma materia que aquellas que no lo están.
3) El texto del enlace describe la página objetivo.

Indexación en la Web

- Algo diferente que con documentos habituales:
  - Se hace un estudio de las etiquetas HTML.
  - Ponderación de términos variando según aparezcan en el título de la página, estén enfatizados, o pertenezcan al texto de un enlace.
    - Probablemente ofrezcan mejores descripciones que la propia página.
    - Probablemente contengan términos más significativos.
    - Representación de páginas no indexables.

Análisis de enlaces

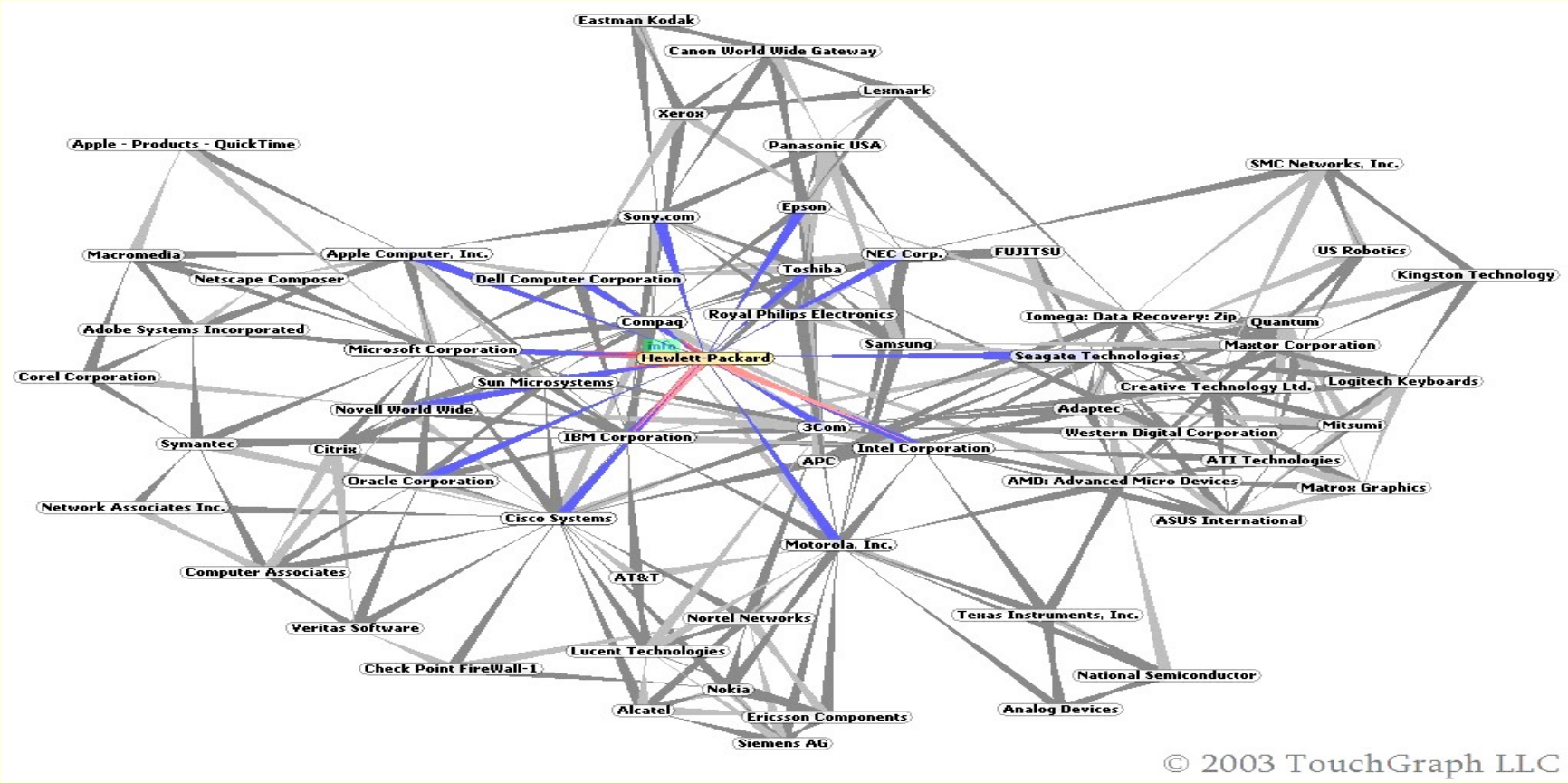
- Un estudio de las características de la Web pueden ayudar en la R.I. Web.
  - Al incluir un hiperenlace reflejamos nuestra opinión sobre la calidad de un sitio web.
  - Cuantos mas enlaces de entrada de un sitio, mayor será la calidad que se le presupone.

Objetivos:

- Repasar distintos algoritmos para el análisis de enlaces.

Análisis de enlaces

- Nodos representan las páginas Web y los arcos se asocian con hiperenlaces



Análisis de enlaces

- Los arcos pueden ser dirigidos o no
- Es un grafo muy dinámico
  - Los arcos y nodos se añaden/eliminan muy frecuentemente
  - El contenido de los nodos (páginas webs) también puede cambiar
- No es necesiamente un grafo conexo, podemos encontrar pequeñas componentes del mismo que no están conectadas con ninguna otra componente del grafo.

Análisis de enlaces: primeros modelos

Bray 1996

- La popularidad (prestigio) de un sitio se mide mediante el número de otros sitios que apuntan hacia él.
- El "visibilidad" de un sitio se define como el número a los que apunta.

✂ → Problemas: No tiene en cuenta la importancia relativa de los sitios adyacentes (padres/hijos).

Análisis de enlaces: primeros modelos

Marchiori (1997)

- El peso final de un documento es una combinación de la información sobre los hiperenlaces y la inf. textual

$$S(v) = s(v) + h(v)$$

- S(v): Peso global
- s(v): información textual
- h(v): información de hiperenlaces

- $h(v) = \sum_{w \in |ch[v]|} F^{r(v, w)} S(w)$

- F: constante , $F \in (0, 1)$
- r(v, w): la posición de w tras ordenar los hijos de v por S(w)

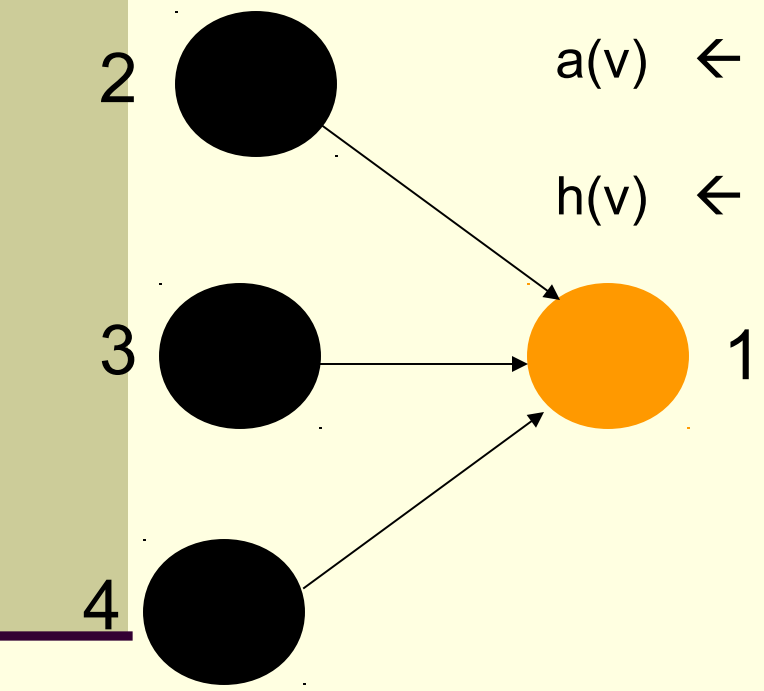
HITS – Kleinberg (1998)

- HITS – Hypertext Induced Topic Selection
- Para cada vértice v Є V en el subgrafo

  a(v) – la autoridad de v
  h(v) – (hub) la centralidad de v

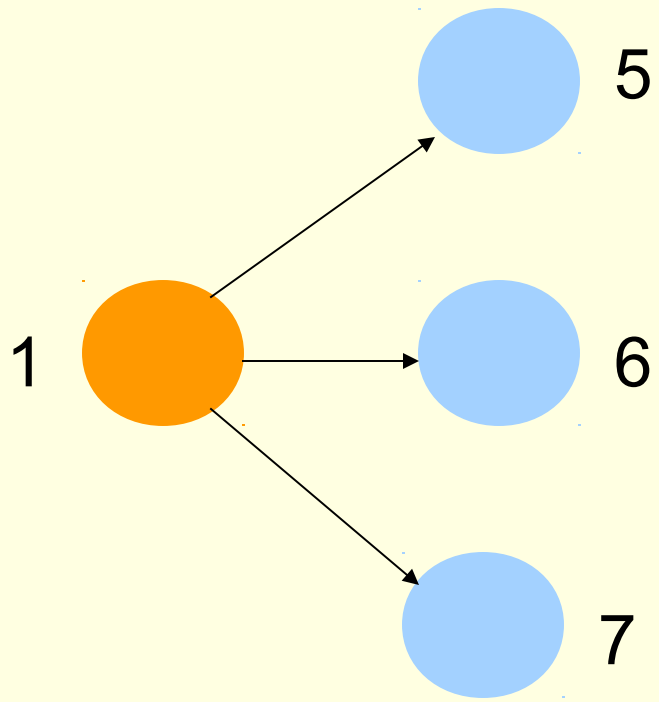
- Un sitio tiene mucha autoridad si recibe muchas citas (el peso de los enlaces depende de la importancia del sitio del que proceden)
- h(v) nos indica cuan buena es la información que se consigue siguiendo los enlaces que tiene a otras páginas. La centralidad de un sitio depende de la autoridad de los sitios a los que apunta

Autoridad (A) y Centralidad (h)



$a(v) \leftarrow \Sigma_{w \in pa(v)} h(w)$

$h(v) \leftarrow \Sigma_{w \in pa(v)} a(w)$



$a(1) = h(2) + h(3) + h(4)$

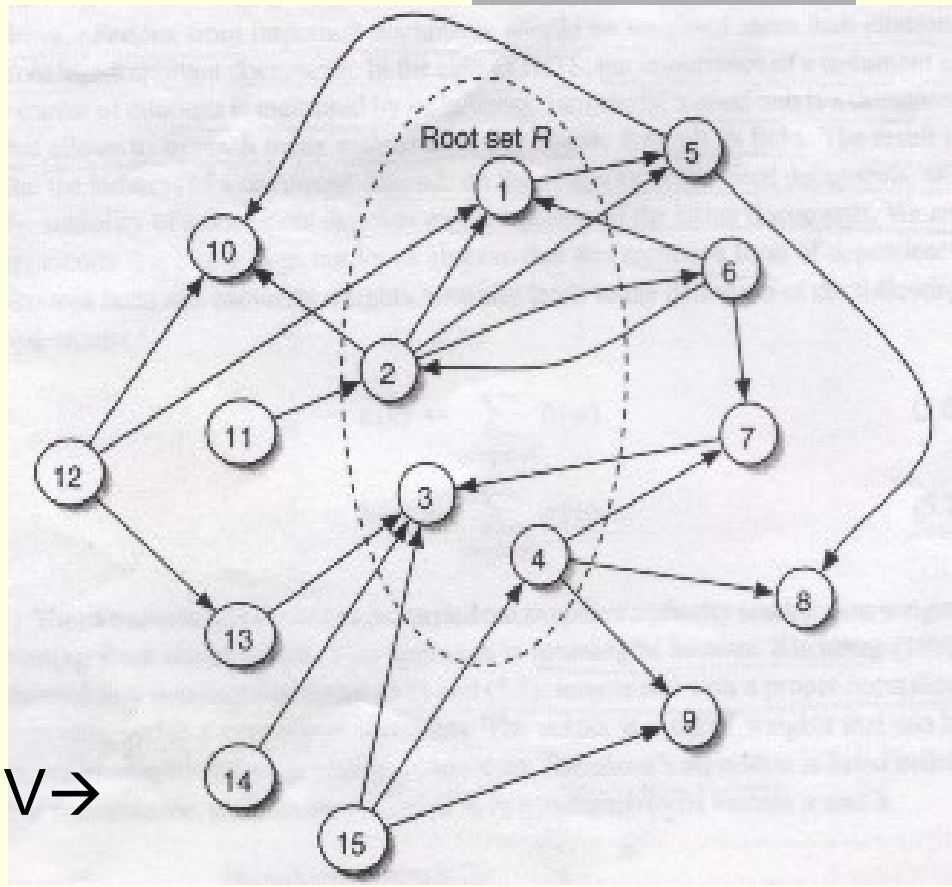
$h(1) = a(5) + a(6) + a(7)$

Autoridad y Centralidad convergen

Ejemplo HITS

Encontrar el subgrafo base

- Comenzar por el conjunto raíz
  R {1, 2, 3, 4}
- {1, 2, 3, 4} –nodos relevantes a la consulta
- Expandir R para incluir todos los hijos y un numero fijo de padres en R



→ Obtenemos el subgrafo base V→

Algoritmo HITS

HubsAuthorities(G)

1 $\quad$ **1** $\leftarrow$ [1,...,1] ϵ $R^{|V|}$

2 $\quad$ $a_0 \leftarrow h_0 \leftarrow$ **1**

3 $\quad$ $t \leftarrow 1$

4 $\quad$ repeat

5 $\quad\quad$ for each v in V

6 $\quad\quad$ do $a_t(v) \leftarrow \sum_{w \,\epsilon\, pa[v]} h_{t-1}(w)$

7 $\quad\quad\quad$ $h_t(v) \leftarrow \sum_{w \,\epsilon\, ch[v]} a_{t-1}(w)$

8 $\quad\quad$ $a_t \leftarrow a_t / \| a_t \|$

9 $\quad\quad$ $h_t \leftarrow h_t / \| h_t \|$

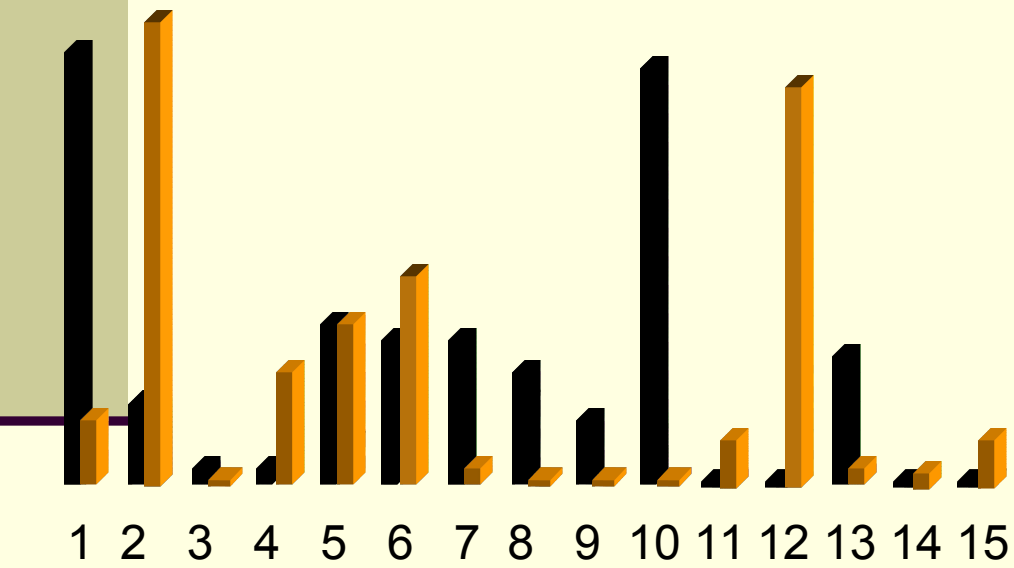
10 $\quad\quad$ $t \leftarrow t + 1$

11 $\quad$ until $\| a_t - a_{t-1} \| + \| h_t - h_{t-1} \| < \varepsilon$

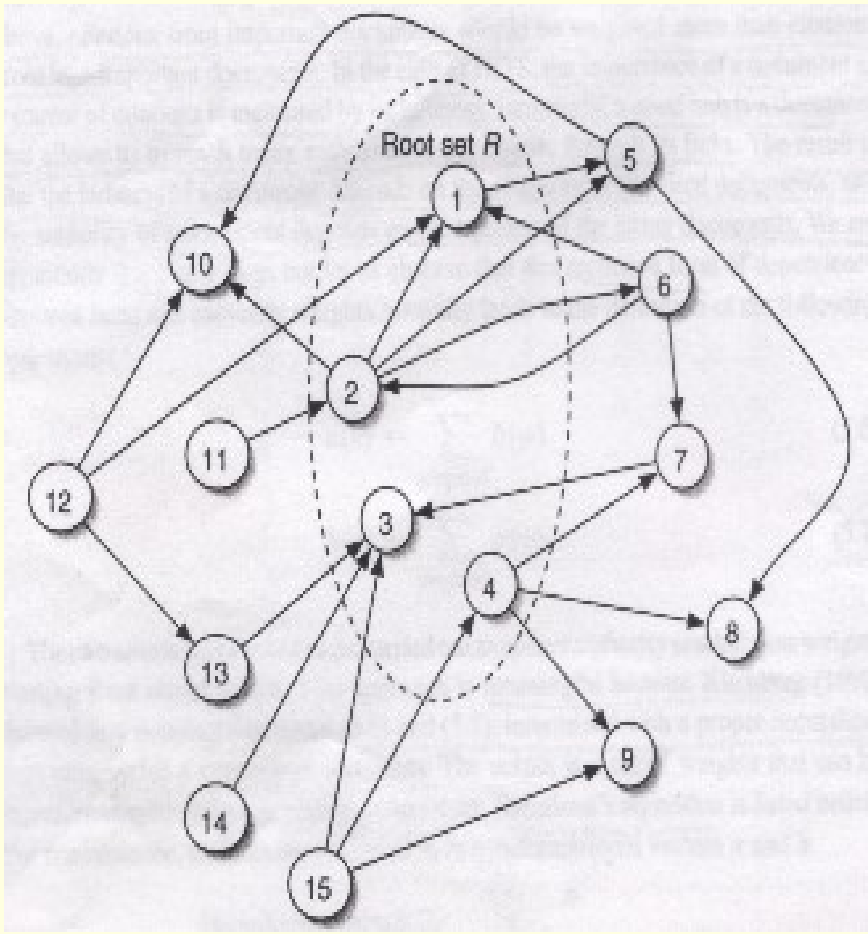
12 $\quad$ return (a_t , h_t)

Ejemplo HITS

A (autoridad)
H (centralidad)



Pesos



PageRank (Page et al.,1998)

- PageRank (PR) es uno de los métodos que utiliza Google para determinar la relavance de una página web.
- PR se puede considerar como un “voto”, de todas las páginas en la Web, sobre la importancia de una página. Un enlace a dicha página representa un voto “positivo”.

$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

donde d factor de “relajación” en (0,1). Usulx 0.85

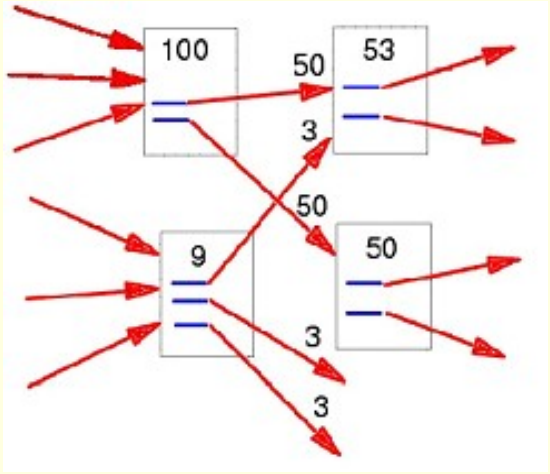
T1, …, Tn son páginas que citan a A

C(X) es el número de enlaces que salen de X

PageRank (Page et al.,1998)

- Es peso es determinado únicamente por el peso de los padres

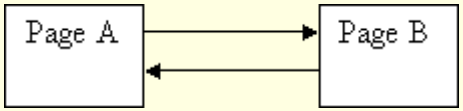
$$r(v) = \alpha \sum_{w \in pa[v]} \frac{r(w)}{|ch[w]|},$$



- Diferencia con HITS
  - HITS considera pesos de Hubness & Authority
  - El ranking de la página es proporcional es proporcional al ranking de los padres e inversamente proporcional al grado de salida de los padres

Ejemplo PageRank

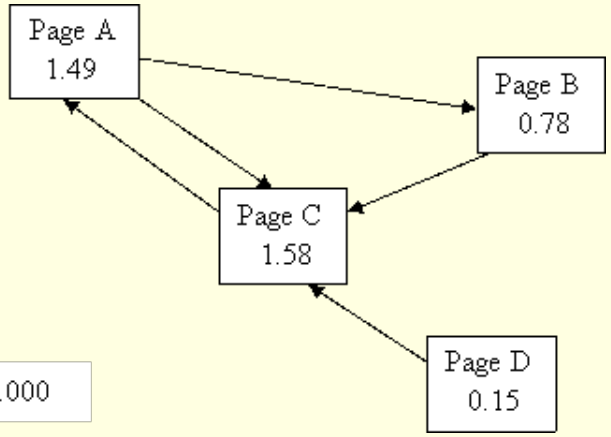
- Se puede calcular utilizando un algoritmo iterativo que converge a los valores reales.
- Inicialmente, se le asignan valores de PR a cada página
- Ejemplo:



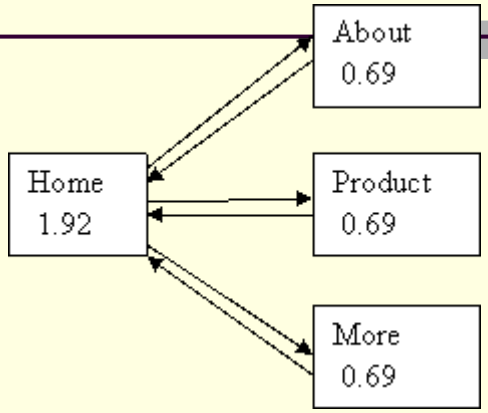
Supongamos PR(X) = 0
PR(A)= 0.15 + 0.85 * 0 = 0.15
PR(B)= 0.15 + 0.85 * 0.15 = 0.277
PR(A)= 0.15 + 0.85 * 0.277 = 0.385
PR(B)= 0.15 + 0.85 * 0.385 = 0.478
PR(A)= 0.15 + 0.85 * 0.478 = 0.556
PR(B)= 0.15 + 0.85 * 0.556 = 0.623
……..
PR(A) = 1 y PR(B) = 1

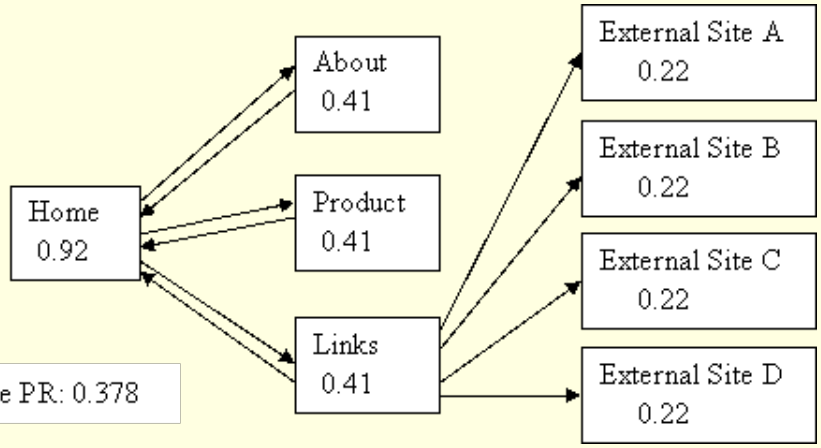
Supongamos PR(X) = 40
PR(A)= 0.15 + 0.85 * 40 = 34.25
PR(B)= 0.15 + 0.85 * 34.25= 29.17
PR(A)= 0.15 + 0.85 * 29.17= 24.95
PR(B)= 0.15 + 0.85 * 24.95= 21.35
……..
PR(A) = 1 y PR(B) = 1

Ejemplo PageRank (Ejemplos)

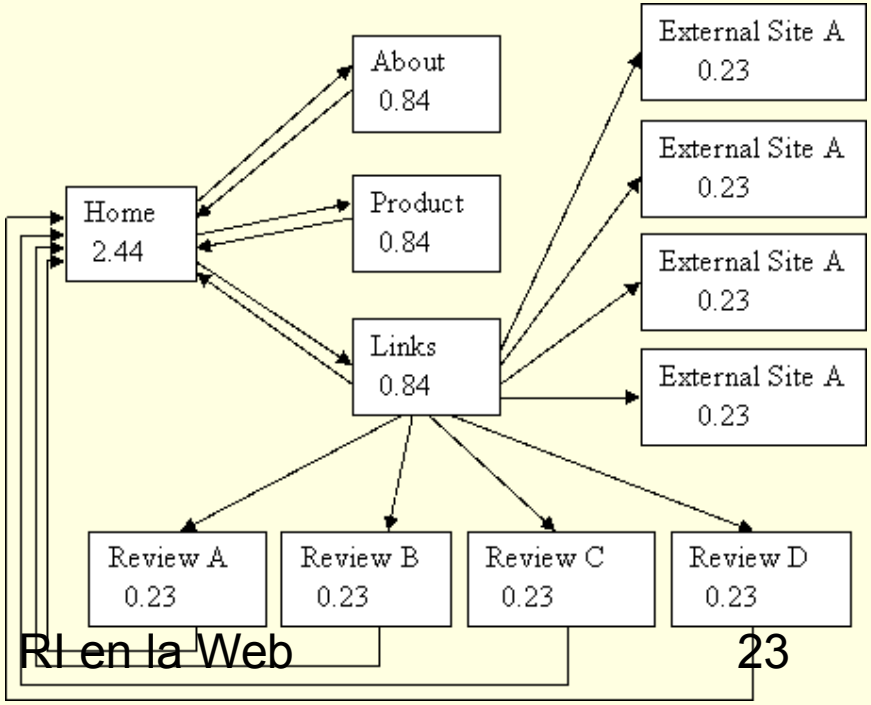


Average PR: 1.000





Average PR: 0.378



Algoritmo PageRank

PAGERANK(M, n, ϵ)

1 $\mathbf{1} \leftarrow [1, \ldots, 1] \in \mathbb{R}^n$
2 $z \leftarrow \frac{1}{n}\mathbf{1}$
3 $x_0 \leftarrow z$
4 $t \leftarrow 0$
5 repeat
6 $\quad t \leftarrow t + 1$
7 $\quad x_t \leftarrow M^{\mathrm{T}} x_{t-1}$
8 $\quad d_t \leftarrow \|x_{t-1}\|_1 - \|x_t\|_1$
9 $\quad x_t \leftarrow x_1 + d_t z$
10 $\quad \delta \leftarrow \|x_{t-1} - x_t\|_1$
11 until $\delta < \epsilon$
12 return x_t

* Page et al, 1998