¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web?

En primer lugar, ¿cómo funcionaban primeros buscadores?

ALIWEB: Los administradores de los sitios web debían registrarse en el buscador proporcionando la URL a una ficha descriptiva del sitio (unas pocas palabras clave) que era incluida en una base de datos. No hay información sobre la relevancia de los resultados pero se presume escasa (¿búsqueda booleana?)

WWW Worm: Para cada documento se almacenaba el título, URL y texto de los enlaces recibidos. Exploraba la Web en busca de nuevos recursos. Usaba **`egrep`** para las búsquedas (coincidencia con una expresión regular).

Web Crawler: También exploraba la Web para localizar nuevos documentos (pero el índice solo podía almacenar 50.000). Empleaba un modelo vectorial y eliminaba palabras vacías. Por primera vez se ofrecen datos sobre exhaustividad (adecuada) y precisión (escasa).

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web?

En primer lugar, ¿cómo funcionaban primeros buscadores?

Lycos: También explora la Web en busca de nuevos documentos (no parece tener un límite arbitrario). No indexa el texto completo del documento (título, cabeceras, 100 palabras más relevantes *tf*idf* y primeras 20 líneas). Como *WWW Worm*, también utiliza el texto de los enlaces entrantes. No emplea exactamente un modelo vectorial pues el cálculo de la relevancia se hace en base a: número de términos de la consulta que aparecen en el documento, frecuencia de los mismos o proximidad.

Naturalmente, hubo más buscadores (*Altavista*, *inktomi*, etc.) pero no hay muchos detalles sobre su funcionamiento.

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web?

En resumen, el mejor buscador Web antes de 1998 sería así…

- Empleaba robots para explorar la Web en busca de documentos
- Almacenaba el texto completo de las páginas web además del texto de los enlaces entrantes
- No tenía en cuenta las palabras vacías en documentos ni en consultas
- Los términos podían ponderarse mediante *tf*idf*
- Retornaba resultados ordenados por relevancia decreciente
- La relevancia se calculaba *ad hoc* teniendo en cuenta no sólo el peso de los términos según el modelo vectorial sino relativos a la proximidad entre los términos o aspectos de "formateo" (título, cabeceras, etc.)

Y no funcionaba "bien"...

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

Un momento, no tan rápido...

- ¿Búsqueda booleana?
- ¿Modelo vectorial?
- ¿*tf*idf*?
- ¿Palabras vacías?
- ¿*Stemming*?
- ¿Precisión y exhaustividad?

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

Búsqueda booleana

Modelo *bag-of-words*, los términos están presentes o no

Las consultas son expresiones lógicas que combinan términos y operadores lógicos

Problemas

- Las consultas retornan o demasiados documentos o muy pocos
- No hay ninguna forma de ordenar los resultados por relevancia.

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

Búsqueda booleana con medidas de asociación

Tanto documentos y consultas se representan mediante *bags-of-words*

Se dispone de coeficientes que determinan cuán relevante es un documento para una consulta

Coeficiente de Dice	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
Coeficiente de Jaccard	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$
Coseno	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert \cdot \lvert Y \rvert}$
Coeficiente de solapamiento	$\frac{\lvert X \cap Y \rvert}{\min(\lvert X \rvert, \lvert Y \rvert)}$

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

Modelo vectorial

- Los documentos son puntos en un entorno *T*-dimensional, donde *T* es el número de términos diferentes en la colección
- Los términos son generalmente palabras o raíces (*stems*) o lemas de palabras
- Cada coordenada de un vector documental tendrá un peso que será nulo si el término no aparece en el documento y no nulo en caso contrario
- Pueden usarse distintos métodos de ponderación, habitualmente ***tf*idf***
- Es posible definir distancias (y similitudes) entre los documentos de manera algebraica
- La función del coseno es la medida más común

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

tf

Método para ponderar los términos de un documento en base a la frecuencia de aparición de los mismos en el propio documento; se supone que un término muy repetido será muy importante

idf (inverse document frequency)

Método para ponderar los términos de un documento en base al número de documentos de la colección que los contienen. Un término es tanto más informativo (i.e. importante) cuanto menor es el número de documentos que lo emplean

*tf*idf*

Método para ponderar los términos de un documento que combina los dos anteriores

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

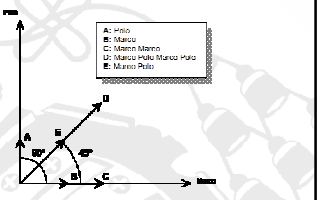
Función del coseno *(cosine similarity)*

Medida de similitud empleada en el modelo vectorial

En la siguiente ecuación n es el número de términos (dimensiones del espacio vectorial) y q_i y d_i son, respectivamente, el i-ésimo término de los documentos q y d.

La función del coseno admite una interpretación geométrica sencilla puesto que mide el ángulo formado por los vectores de los documentos a comparar.

$$\frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$



¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

Palabras vacías *(stop words)*

Se denominan stop words o palabras vacías aquellas palabras que, a pesar de un uso frecuente, aportan por sí solas poco significado a un texto

Eliminarlas no siempre es una buena idea. Riloff, E. 1995, "Little words can make a big difference for text classification", en *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 130-136.

Además, ¿qué es una palabra vacía? Por ejemplo, **`ser`**

- Verbo (palabra vacía)
- **`Cadena SER`** (no es palabra vacía)
- **`SER Society for Ecological Restoration`** (no es castellano)

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

Stemming (reducción a la raíz)

Algoritmos que colapsan múltiples formas de una palabra en un único término.

Por ejemplo, **`investigación`**, **`investigaciones`**, **`investigador`**, **`investigadora`** e **`investigadores`** colapsan en **`investig`**. En cambio **`universidad`** colapsa a **`univers`** mientras que **`universitario`** lo hace a **`universitari`**.

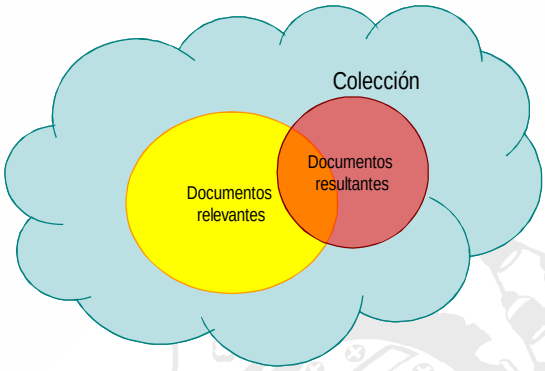
Aplicando *stemming* se reduce el número de términos y, al mismo tiempo, se consigue que una misma consulta abarque más palabras (algo que puede ser un problema, p.ej. **`universo`**)

`http://snowball.tartarus.org/`

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

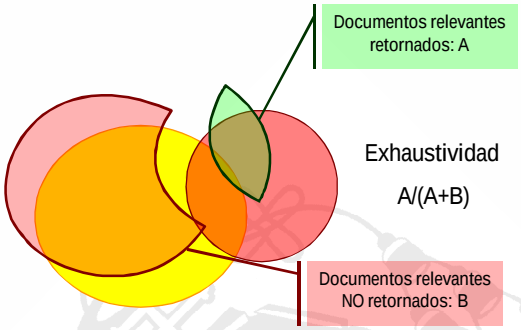
¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

Precisión *(precision)* y exhaustividad *(recall)*



Precisión *(precision)* y exhaustividad *(recall)*

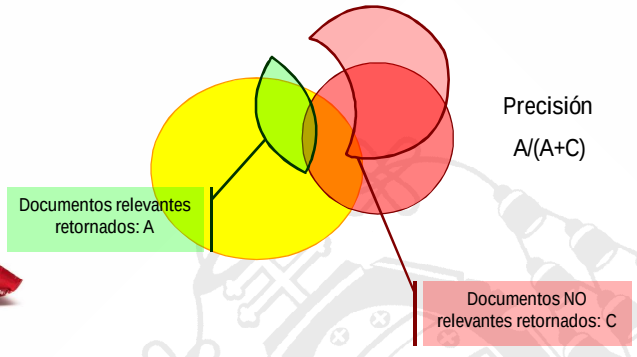
¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)



Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

Precisión *(precision)* y exhaustividad *(recall)*

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)



¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web? (Intermedio)

Precisión *(precision)* y exhaustividad *(recall)*

En resumen,

- Precisión es el porcentaje de los documentos resultantes que son verdaderamente relevantes
- Exhaustividad es el porcentaje de los documentos relevantes que son retornados al usuario

Un sistema *IR* perfecto tendría precisión y exhaustividad 1.00 siempre; sin embargo, eso es imposible.

Y ahí es donde volvemos a los buscadores Web pre-*Google*...

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web?

En 1998 la cota inferior para la Web era de 320 x 10^6 documentos.
Lawrence, S. y Giles, C.E. 1998, "Searching the World Wide Web", *Science*, vol. 280, no. 3, pp. 98-100.

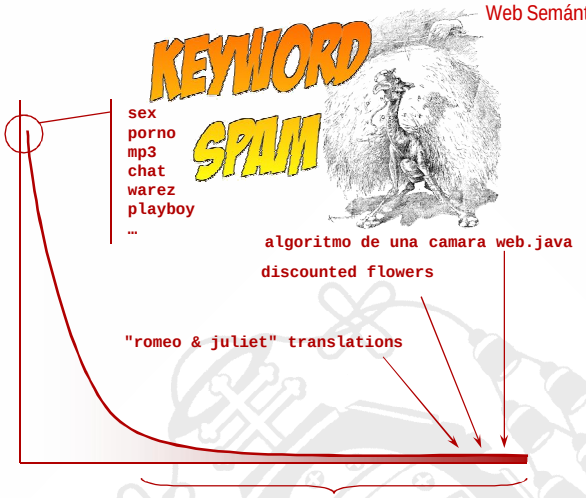
La mayor colección de evaluación de la época tenía "sólo" 7.5 x 10^6 documentos.

Características de las consultas

- Son muy cortas (casi el 90% constan de 3 términos o menos)
- Más del 60% son únicas

Dado el número de documentos y la longitud de las consultas la mayor parte de los resultados eran irrelevantes ≡ $\lim_{R\to\infty} P = 0$

¿Por qué las técnicas *IR* clásicas no funcionan bien en la Web?



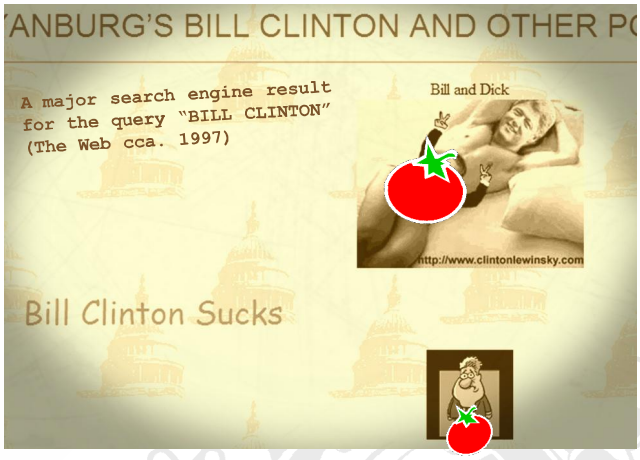
La Web es un grafo

Hasta aquí hemos llegado...

★ Brin, S. y Page, L. 1998, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117.

> *as of November 1997, only one of the top four commercial search engines finds itself.*
>
> *...*
>
> *[...] we have seen a major search engine return a page containing only "Bill Clinton Sucks" and picture from a "Bill Clinton" query. [...] If a user issues a query like "Bill Clinton" they should get reasonable results since there is a enormous amount of high quality information available on this topic. Given examples like these, we believe that the standard information retrieval work needs to be extended to deal effectively with the web.*

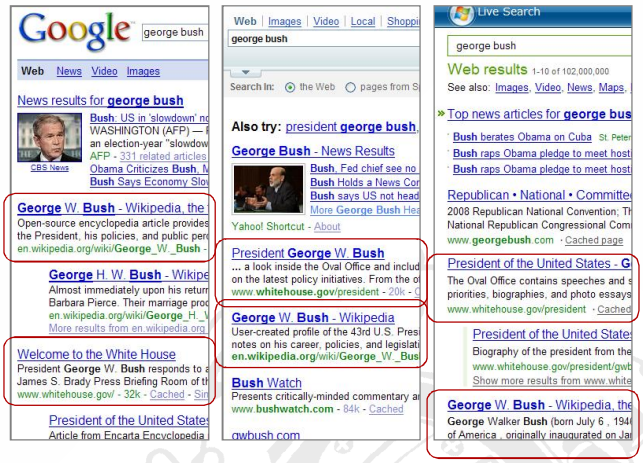
La Web es un grafo



Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

La Web es un grafo

Resultados ofrecidos por Google, Yahoo! y Live Search para la query "GEORGE BUSH". La Web, hoy (aproximadamente)



La Web es un grafo

Demos un paseo…

***Plagiar**, v. Adoptar el pensamiento o el estilo de otro escritor, a quien uno jamás ha leído.*
***Plagio**, s. Coincidencia literaria entre un antecedente carente de mérito y un consecuente honorable.*

…Siempre hay que acreditar las fuentes que hemos usado

Bierce, A. 1906, *The Devil's Dictionary*

En los trabajos científicos se citan trabajos de terceros por dos razones: para interpretarlos o en apoyo de la interpretación personal. Umberto, E. 1977, *Cómo se hace una tesis*.

Las citas deben aportar algo nuevo o confirmar lo sabido con autoridad.

Citando dotamos de autoridad a un tercero…

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

La Web es un grafo

Domingo, 17 de febrero de 2008 — LA NUEVA ESPAÑA

La orla dorada de la ciencia

Otín, Barluenga y Sanz Medel son los investigadores asturianos de mayor impacto internacional

Oviedo, P. Á.

¿Quién es quién en la Universidad de Oviedo? ¿Quiénes son los investigadores más productivos? ¿Es posible hacer un ranking de los científicos asturianos? Tal ranking –uno de los posibles rankings– existe, está disponible en internet y se halla en permanente evolución. Desde tiempo atrás, lo encabeza Carlos López Otín, catedrático de Biología Molecular y uno de los investigadores más destacados del panorama nacional.

La clasificación a la que se refiere esta información es la basada en el «índice h», un método que toma como referencia los

Los científicos asturianos de mayor impacto

Nombre	Índice H	Año de inicio	Especialidad
Carlos López Otín	57	1982	Bioquímica y Biología Molecular
José Barluenga Mur	39	1971	Química Orgánica
Alfredo Sanz Medel	35	1976	Química Analítica
José Gimeno Heredia	32	1974	Química Inorgánica

La Web es un grafo

Donde dije "cita" digo "hiperenlace"…

Marchiori, M. 1997 "The Quest for Correct Information on the Web: Hyper Search Engines". *The Sixth International WWW Conference (WWW 97).*

> *A great problem with search engines' scoring mechanisms is that they tend to score text more than hypertext.*
>
> *[...] focusing separately on the "textual" and "hyper" components.*
>
> *The presence of links in a Web object clearly augments the informative content with the information contained in the pointed Web objects.*
>
> *Recursively, links present in the pointed Web objects further contribute, and so on. Thus, in principle, the analysis of the informative content of a Web object A should involve all the Web objects that are reachable from it […]*
>
> *This is clearly unfeasible in practice, so, for practical reasons, we have to stop the analysis at a certain depth […]*

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

La Web es un grafo

Donde dije "cita" digo "hiperenlace"…

Marchiori, M. 1997 "The Quest for Correct Information on the Web: Hyper Search Engines". *The Sixth International WWW Conference (WWW 97).*

A great problem with search engines' scoring mechanisms is that they tend to score text more than hypertext.

[...] focusing separately on the "textual" and "hyper" components.

The presence of links in a Web object clearly augments the informative content with the information contained in the pointed Web objects.

Recursively, links present in the pointed Web objects further contribute, and so on. Thus, in principle, the analysis of the informative content of a Web object A should involve all the Web objects that are reachable from it […]

La Web es un grafo

Donde dije "cita" digo "hiperenlace"...

Jon Kleinberg define los conceptos de autoridad y *hub*

Kleinberg, J.M. 1998, "Authoritative sources in a hyperlinked environment", en *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pp. 668-677. ★

Una autoridad es un documento fuertemente enlazado

Un *hub* es un documento que enlaza a muchas autoridades

Esta técnica logró que el 50% de los resultados para las consultas fueran relevantes, frente al 40% de *Yahoo!* (un directorio) o *Altavista*

Chakrabarti, S., Dom, B.E., Gibson, D., Kleinberg, J., Raghavan, P. y Rajagopalan, S. 1998, "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text", en *Proceedings of the 7th World-Wide Web conference*, pp. 65-74. ★

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

La Web es un grafo

La Web es un grafo

Google comienza a operar en 1998

- Brin, S. y Page, L. 1998, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117.

El núcleo de su sistema de ponderación es el algoritmo *PageRank*, similar al método de Kleinberg

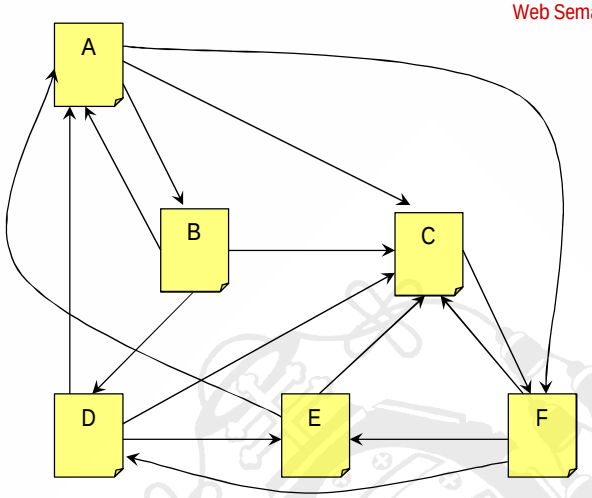
- Page, L., Brin, S., Motwani, R. y Winograd, T. 1998, *The PageRank Citation Ranking: Bringing Order to the Web*

El algoritmo asocia a cada documento un valor (tb. *PageRank*) de este modo:

- Un documento transmite a todos los documentos que enlaza su valor *PageRank* dividido por el número de enlaces salientes
- Un documento muy enlazado tendrá un *PageRank* elevado
- Un documento enlazado desde documentos prestigiosos tendrá un *PageRank* elevado

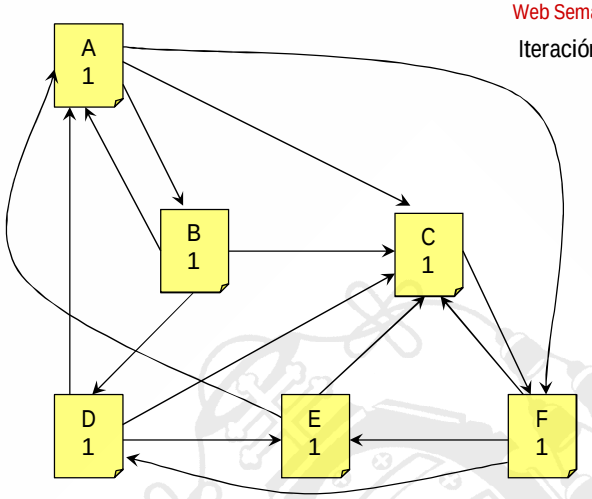
Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

PageRank

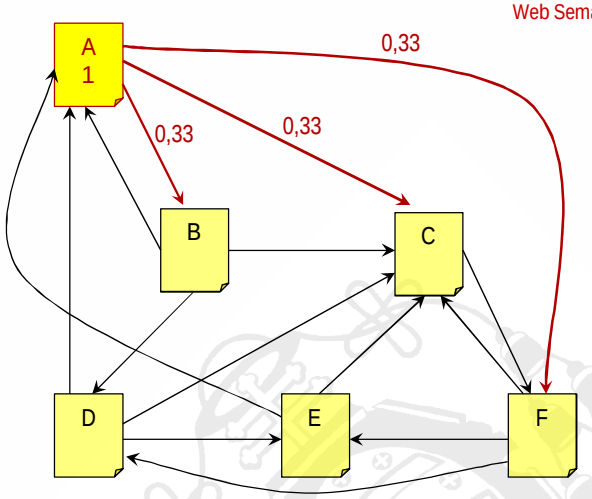


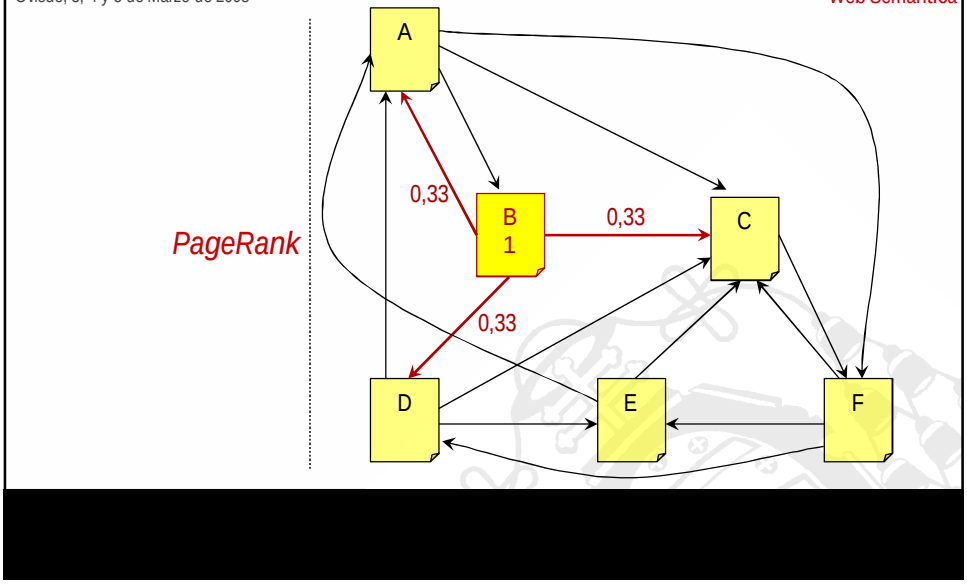
Iteración 0

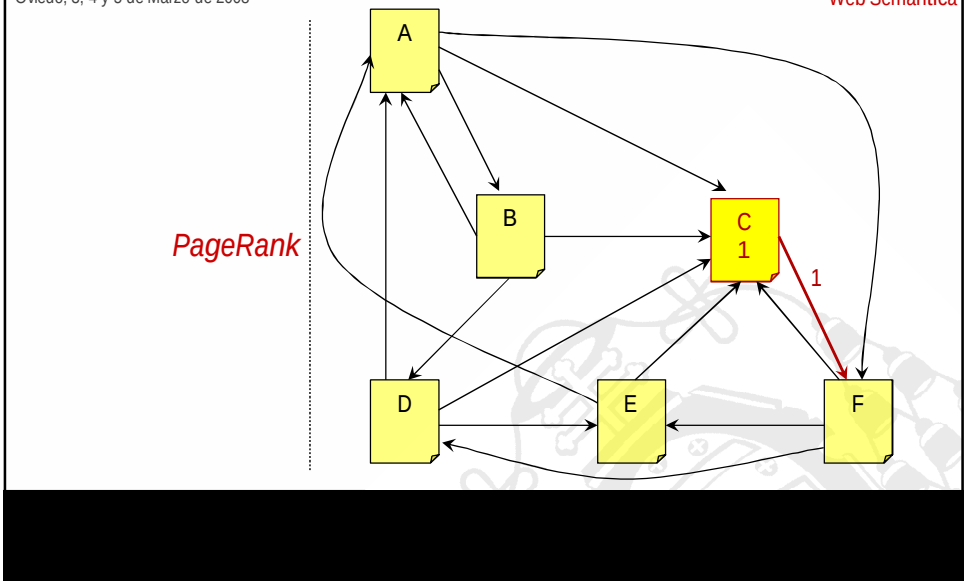
PageRank

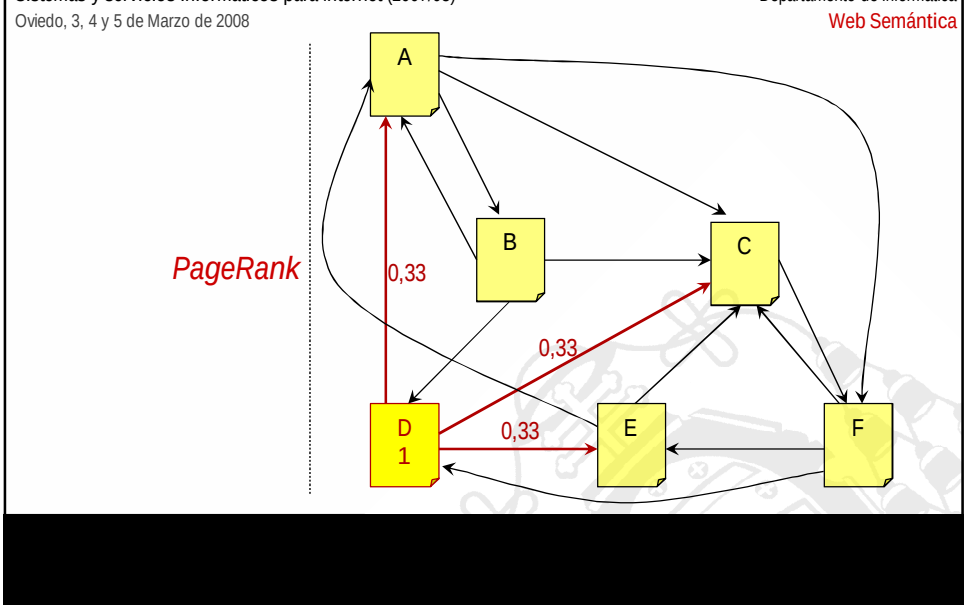


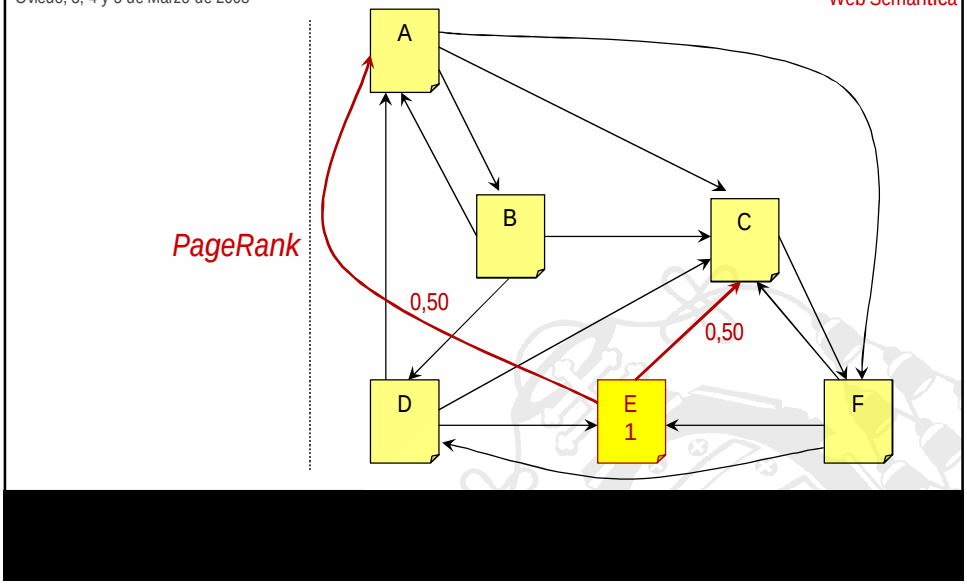
PageRank



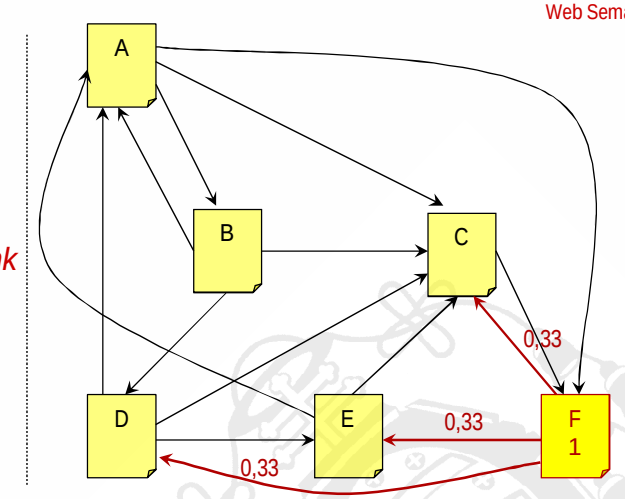

PageRank
A
B
1
C
D
E
F
0,33
0,33
0,33


PageRank
A
B
C
1
D
E
F
1


PageRank
A
B
C
D
1
E
F
0,33
0,33
0,33

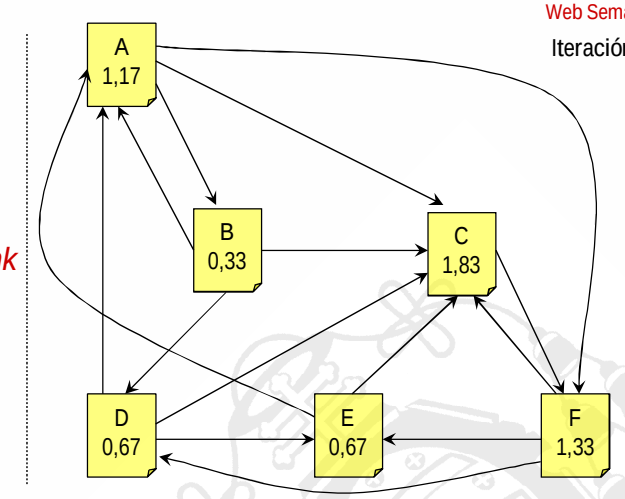

PageRank
A
B
C
D
E
1
F
0,50
0,50

PageRank



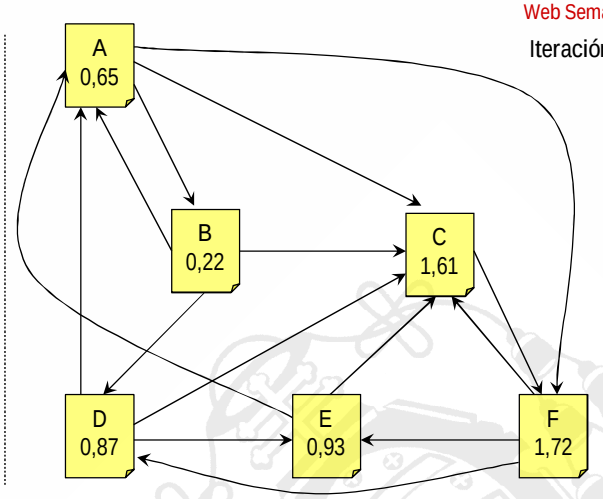
Iteración 1

PageRank



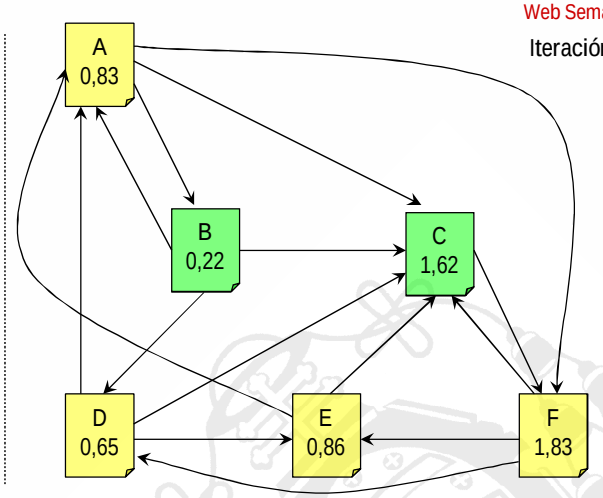
Iteración 2

PageRank



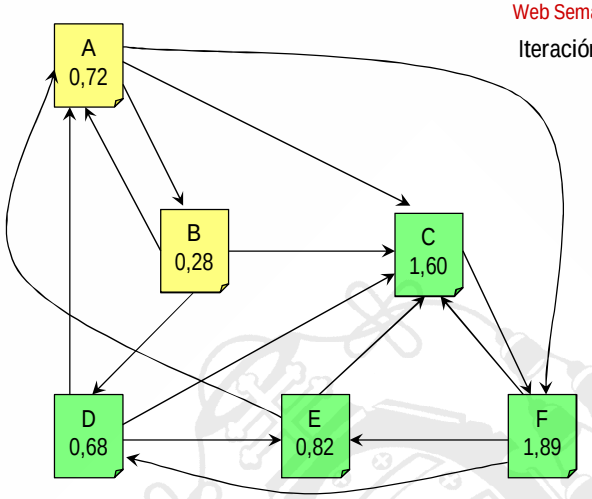
Iteración 3

PageRank



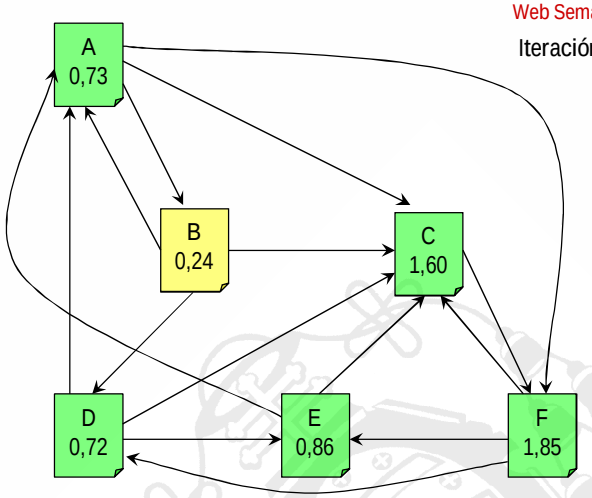
Iteración 4

PageRank



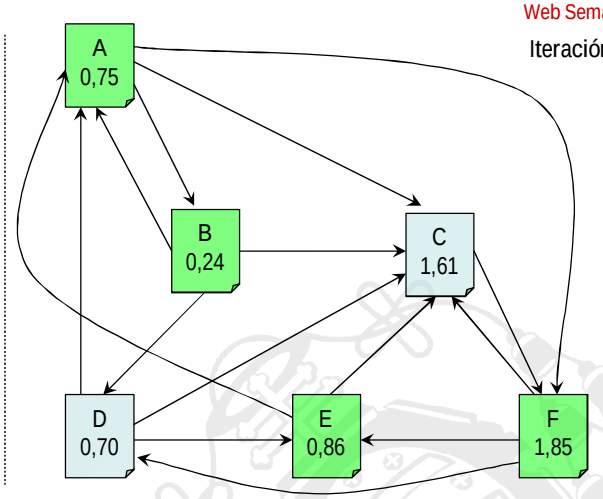
Iteración 5

PageRank



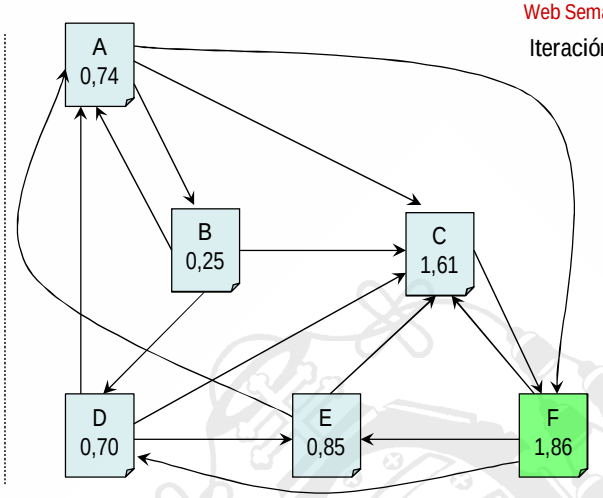
Iteración 6

PageRank



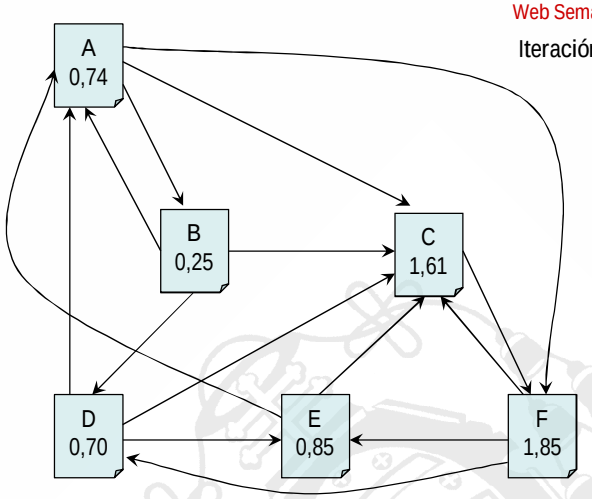
Iteración 7

PageRank



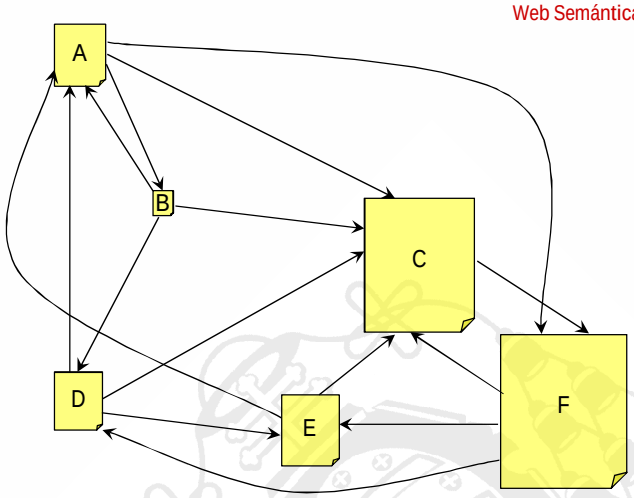
Iteración 8

PageRank



Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

PageRank



PageRank

Algunas características interesantes de *PageRank*

Los valores de *PageRank* calculados para los nodos se "estabilizan" con rapidez (p.ej. 52 iteraciones son suficientes para obtener valores razonables para 322 millones de enlaces)

Es relativamente insensible a los valores de "partida", afectaría al número de iteraciones necesarias y a los valores finales (obviamente) pero no al *ranking* obtenido

El *PageRank* total en la Web es constante

Si el valor inicial asignado a cada documento es *1/N* (número de documentos) el valor de *PageRank* equivale a la probabilidad de que un usuario llegue a dicho documento siguiendo enlaces al azar (*random surfer model*)

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

PageRank

Suposiciones sobre la Web...

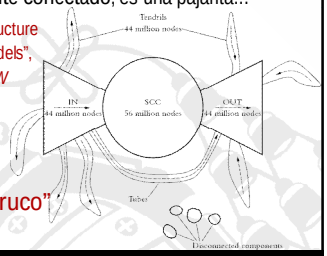
Es un grafo fuertemente conectado (desde cualquier nodo *v* se puede llegar a cualquier nodo *w*)

Todos los nodos tienen enlaces salientes

... que son falsas...

La Web no es un grafo fuertemente conectado, es una pajarita...

★ Broder, A. et al. 2000, "Graph structure in the web: experiments and models", en *Proceedings of the ninth WWW Conference*



Sólo el 90% de la Web está fuertemente conectada

... así que habrá que usar algún "truco"

PageRank

A vueltas de nuevo con el *random surfer*...

El modelo descrito hasta ahora se correspondería con esta ecuación

$$PR(p_i) = \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

que modela a un usuario que va de página en página siguiendo enlaces aleatoriamente, *ad infinitum*…

Problema: Aquellas páginas que no forman parte del núcleo fuertemente conectado de la Web siempre tendrán *PR* nulo…

Sistemas y servicios informáticos para Internet (2007/08)
Oviedo, 3, 4 y 5 de Marzo de 2008
Departamento de Informática
Web Semántica

PageRank

A vueltas de nuevo con el *random surfer*...

Lo que hay que conseguir es que, de vez en cuando, el navegante "salte" a una página aleatoriamente. Es decir, en cada página el usuario toma una "decisión"

- Saltar a una página aleatoria con probabilidad *d*
- Continuar con un enlace al azar de la página actual con probabilidad *1-d*

Este modelo puede representarse según esta ecuación (un valor habitual para *d* es 0,15)

$$PR(p_i) = (1-d) \cdot \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} + \frac{d}{N}$$

PageRank

¿Y los nodos sin enlaces salientes?

- Se eliminan para después calcular el *PageRank* del resto del grafo
- Una vez calculado éste se determina el de los nodos sin enlaces salientes en base al *PageRank* de sus enlaces entrantes

Búsquedas en la Web con *PageRank*

Recordemos lo que dijo Marchiori

> *[...] focusing separately on the "textual" and "hyper" components.*

PageRank no tiene en cuento el contenido de los textos para determinar el prestigio/autoridad/relevancia de un nodo, sólo los enlaces

¿Cómo se realizan las búsquedas entonces? (Versión simplificada)

- Se extraen los términos (palabras) de la consulta
- Se localizan documentos que contengan todos los términos
- Se ordenan los documentos obtenidos por *PageRank* decreciente

Es decir, *Google* proporciona a los usuarios aquellos documentos que satisfacen la consulta y tienen más prestigio en la Web

STOP!

Por hoy estuvo bien…

¿Preguntas?

Para mañana…

Lawrence, S. y Giles, C.E. 1998, "Searching the World Wide Web", *Science*, vol. 280, no. 3, pp. 98-100.

Brin, S. y Page, L. 1998, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117.

Kleinberg, J.M. 1998, "Authoritative sources in a hyperlinked environment", en *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pp. 668-677.

Jansen, B.J. y Spink, A. 2003, "An Analysis of Web Documents Retrieved and Viewed", *The 4th International Conference on Internet Computing*, pp. 65-69.

¿Para qué tipo de consultas son adecuados los buscadores actuales?