



Laboratorio de Redes,  
Recuperación de Información  
y Estudios de la Web

# Recuperación de Información en la Web y Motores de Búsqueda

Gabriel H. Tolosa  
tolosoft@unlu.edu.ar



# RI en Web: Motores de Búsqueda



# Motores de búsqueda

- Escenario/RI Web
- Arquitectura
- Recolección de páginas (Crawling)
- Ranking
- Queries y usuarios
- Escalabilidad (caching)

# Motores de búsqueda



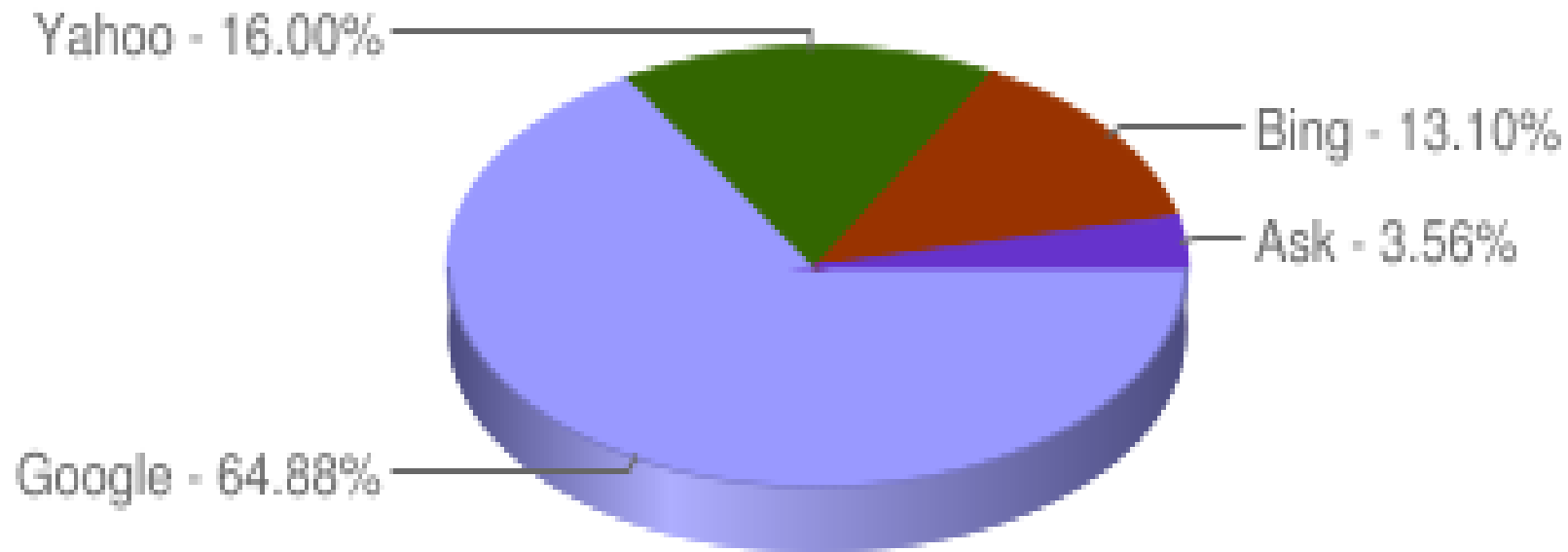
- **¿Son importantes?**
  - ~90% del tráfico a la mayoría de los sitios se encuentra mediante un motor de búsqueda
  - Son la primera interface entre los usuarios y la web
    - En el caso de sitios comerciales (productos) estar más allá de la posición 30 es ser “prácticamente” invisible.
  - Atraen la mayor diversidad de usuarios que cualquier sitio.
  - ~ 85% de las sesiones de usuario incluyen el uso de un MB
  - ~ 90% de los usuarios los usan para navegar la web

# Motores de búsqueda



- ¿Cuáles se usan?

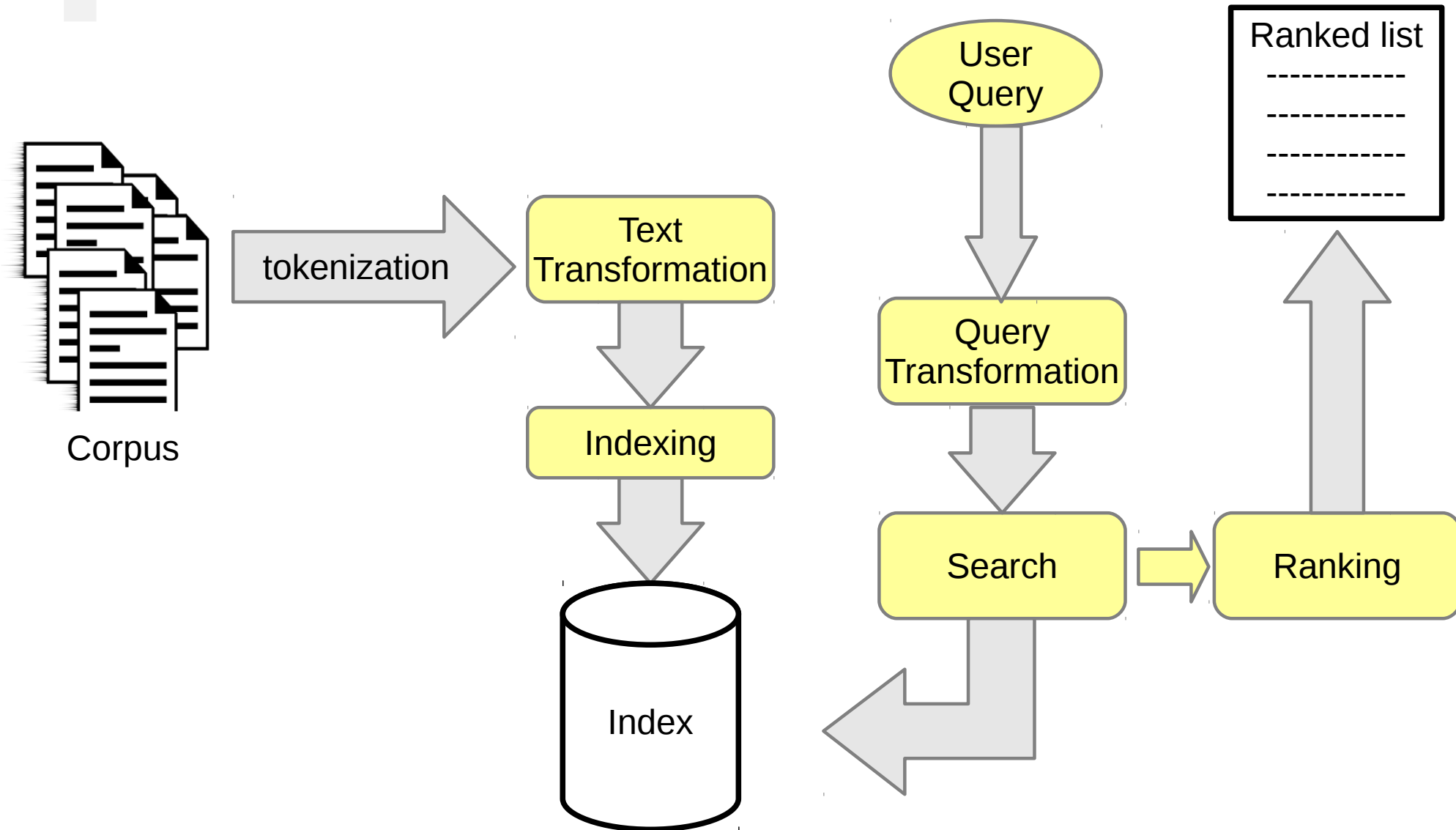
Search Engine Usage For August 2011



# RI tradicional vs web

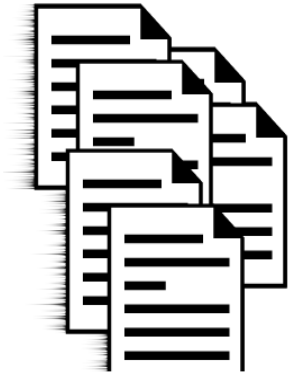
	RI Tradicional	RI en la Web
<b>Objetivo</b>	Recuperar documentos de texto con contenido relevante a la necesidad de información	Recuperar páginas web (y otros docs) de alta calidad relevantes a necesidad de información
<b>Colección</b>	Conjunto de documentos (generalmente homogénea)	La web pública (heterogénea)
<b>usuarios</b>	# proyectado intereses comunes	# impredecible intereses múltiples
<b>Contenido</b>	Relativamente pequeño y poco dinámico	Masivo y altamente dinámico
<b>Consultas</b>	Específicos	Cortos y poco descriptivos
<b>Ranking</b>	Según “grado” de relevancia	Relevancia + reputación + factores contextuales

# SRI tradicional

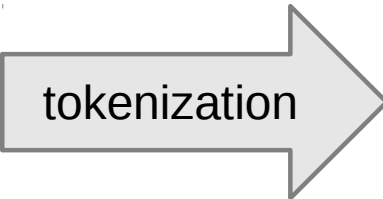


# Pero....

No lo tenemos



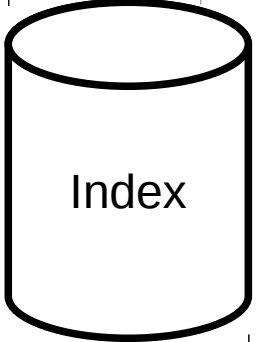
Corpus



Múltiples formatos

Text Transformation

Indexing



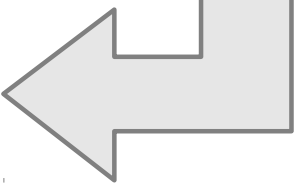
Index

Proceso dinámico

User Query

Query Transformation

Search



Usuarios de diferentes contextos



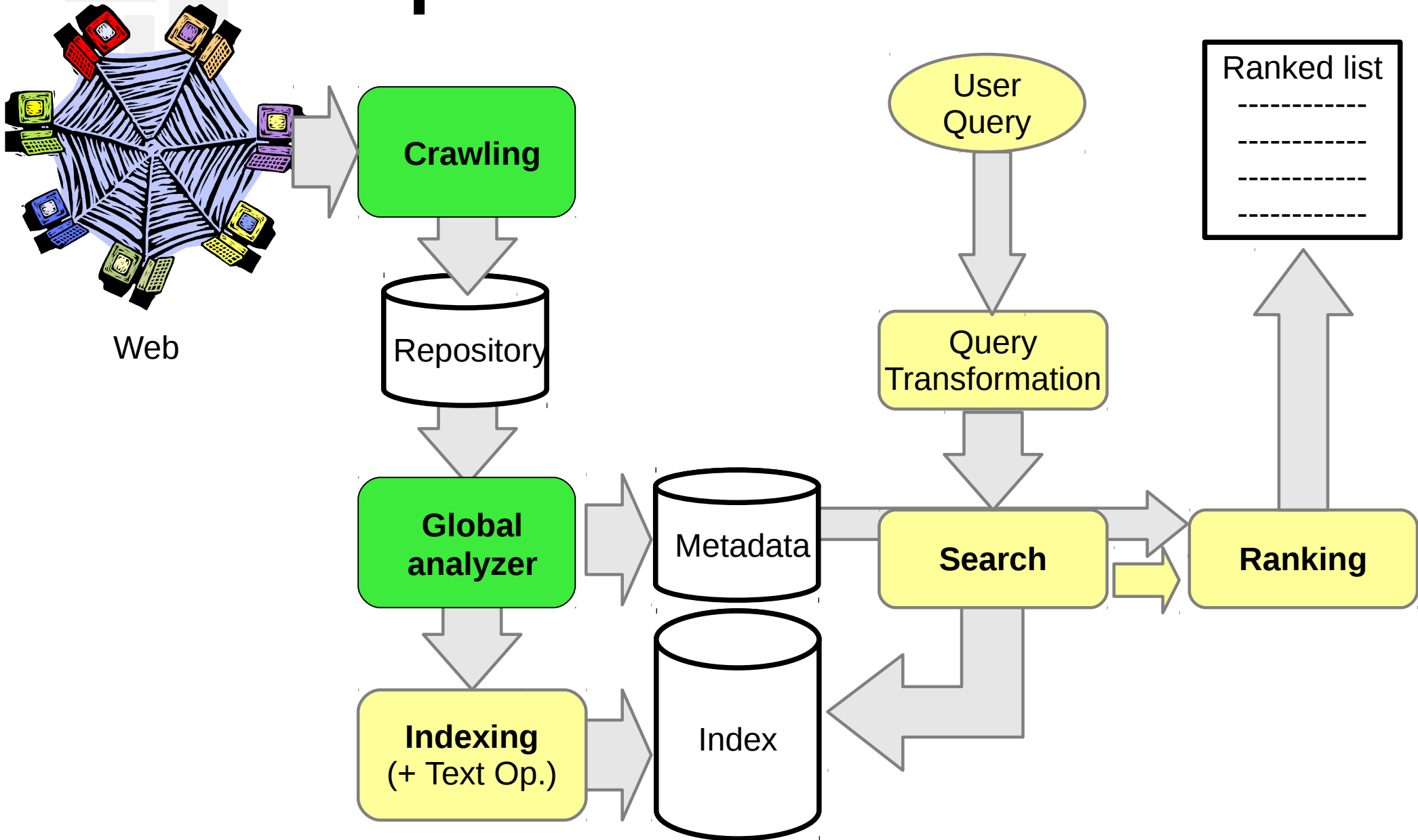
Ranked list

Tiene en cuenta la estructura

Ranking



# Arquitectura de un MB



# Evolución de los MB

## Primera generación

**Solo utilizaban el texto en las páginas**

Altavista, Exite, Lycos

## Segunda generación

**Analizan la estructura de enlaces de la web y los clicks**

“Anchor text”. Google y PageRank

## Tercera generación

**Tratan de resolver “*la necesidad detrás de la consulta*”.**

Ayudan al usuario: speell-checking, sugerencias, refinamiento

Integran múltiples fuentes (news, blogs, imágenes)

Análisis semántico básico. **Aún están evolucionando!**

## Cuarta generación

Incrementar el uso de contexto y la actividad del usuario!

(“*Information supply*”)



# Evolución de los MB

Cómo determinar:

**“la necesidad detrás de la consulta”**

## Determinación del contexto

- Espacial (ubicación del usuario o del objetivo)
- Stream del query (respecto de los anteriores)
- Información personal (perfil)
- Explícito (elige el usuario, por ej. un MB vertical)
- Implícito (uso de Google Argentina, google.com.ar)

## Uso del contexto

- Restricción de resultados (eliminar inapropiados)
- Modulación del ranking (genérico, personalizado)



# Evolución de los MB

## ¿Y los usuarios?

### Las consultas:

Las mayoría tienen de 1 a 3 términos (el 25% tiene 2)

Términos imprecisos

Uso subóptimo de la sintaxis (sólo ~10% con operadores)

### Mucha variación en:

Necesidades

Expectativas

Conocimiento

Recursos (ancho de banda)

### Comportamiento:

Sólo examinan unos pocos resultados (2-3 páginas), ~85% sólo la primera

Poco refinamiento (~80 no modifica la consulta original)

La interface de búsqueda avanzada es poco utilizada



**Crawling**

# El “corpus” web

- Creación no coordinada, distribuida (democrática)

- Ni de contenido ni de enlaces

- Diversidad

- No estructurado (txt, html)
  - Semi-estructurado (XML, objetos 'anotados')
  - Estructurado (BD), en menor medida.

- Tamaño: se duplica en pocos meses!

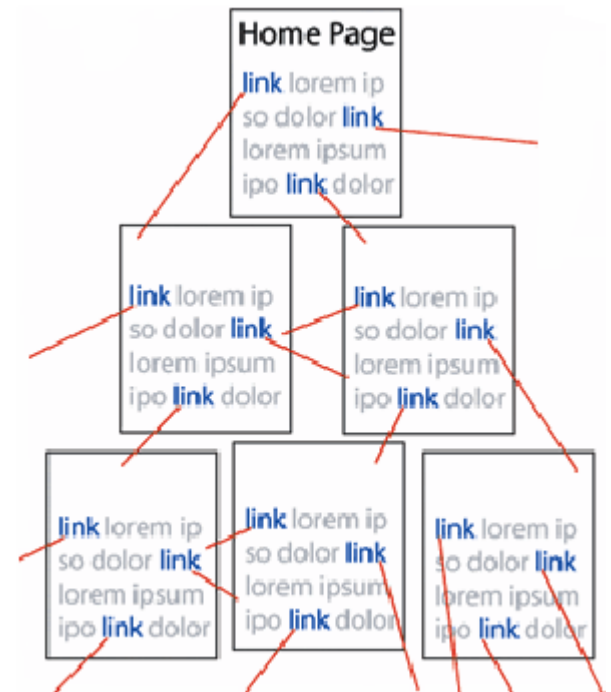
- Enlaces: 8/pág. en promedio

- Contenido dinámico

- 'On the fly'

- HTTP Get/Post

<http://www.google.com/search?hl=en&q=graph+structure+in+de+the+web+slides&btnG=Search>



- SPAM

# Crawling → Obtener la colección

- “Encontrar” y recuperar páginas automáticamente
- La web está constantemente cambiando
- Las páginas cambian
- La web no está bajo el control del propietario del motor de búsqueda
- Se basa solo en la URL:



**`http://www.unlu.edu.ar/academia/unidades.html`**

[proto] [hosts] [path] [objeto]

# Crawling → Obtener la colección

Web crawling  $\Leftrightarrow$  atravesar un grafo

```
S := {páginas iniciales}

mientras no-vacía (S)
{
    tomar s desde S

    si s no fue recuperada antes:
        recuperar s

    parsear s

    para cada link l en s:
        agregar l a S
}
```



# Crawling → Atravezar el grafo



---

**Algorithm 6.1** Simple Web-Crawler to save link structure

---

```
1: push(todo_list,initial_set_of_urls)
2: while todo_list[0]  $\neq \emptyset$  do
3:   page  $\leftarrow$  fetch_page(todo_list[0])
4:   if page downloaded then
5:     links  $\leftarrow$  parse(page)
6:     for all  $l$  in links do
7:       if  $l$  in done_list then
8:         push(todo_list[0].outlinks,done_list[l].id)
9:       else if  $l$  in todo_list then
10:        push(todo_list[0].outlinks,todo_list[l].id)
11:       else if  $l$  pass our filter then
12:         push(todo_list, $l$ )
13:         todo_list[l].id = no. of url's
14:         push(todo_list[0].outlinks,todo_list[l].id)
15:       end if
16:     end for
17:   end if
18: end while
```

---



# Crawling → Cuestiones

- ¿Cómo hacer el crawling?
  - Calidad (las mejores páginas primero)
  - Eficiencia (evitar duplicados)
  - Cortesía (con los servidores)
- ¿Cuánto recolectar?
  - Cobertura
  - Cobertura relativa
- ¿Con qué frecuencia?
  - “Frescura”

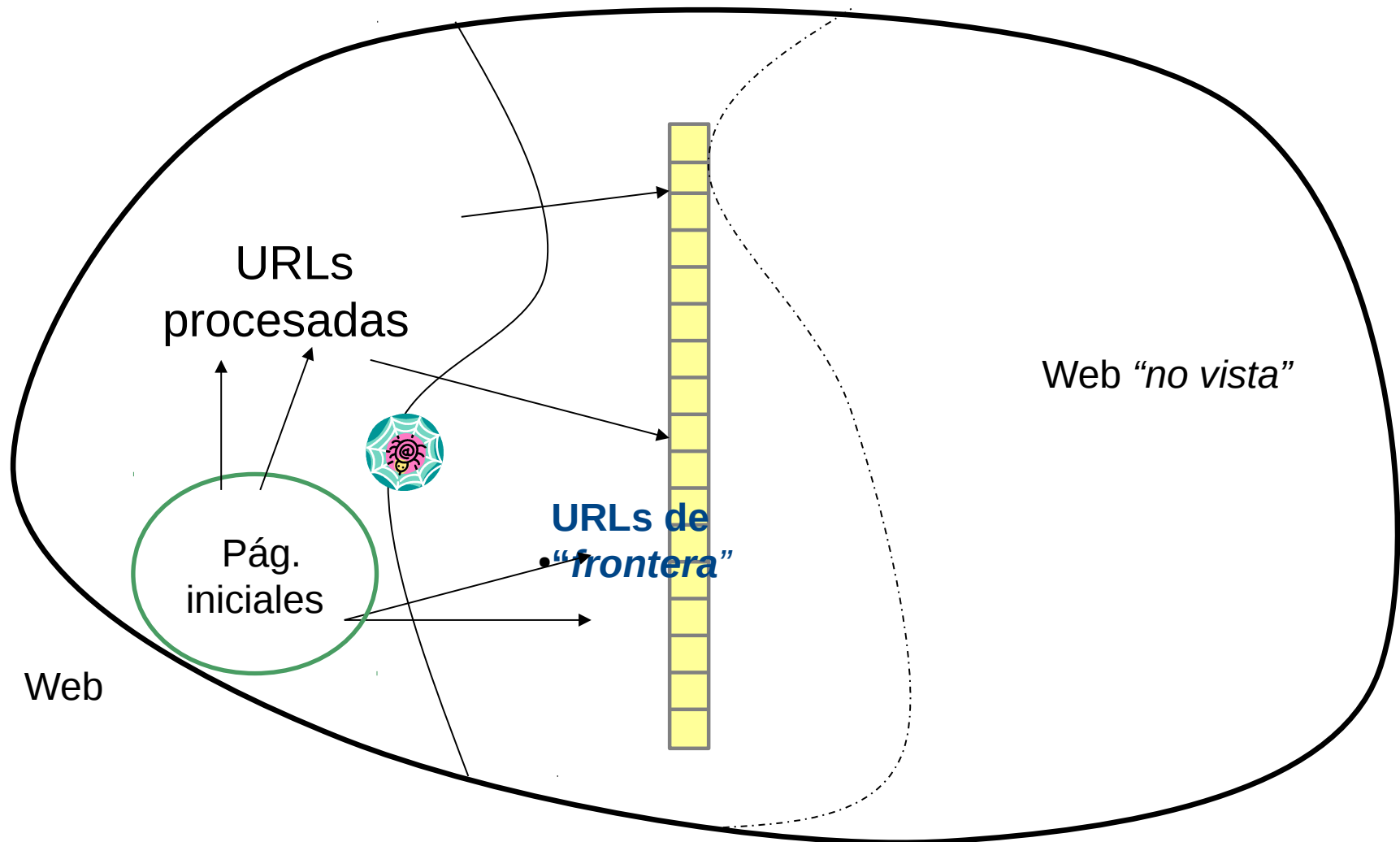


# Crawling → Más específicamente

Para cada URL, el crawler:

- Solicita la resolución del nombre a un servidor DNS
- Abre una conexión con el servidor (IP) en un puerto (usualmente 80)
- Envía una solicitud HTTP, generalmente usando la primitiva GET
- Recupera el objeto y se parsea
- Finalmente, actualiza la lista de URLs (frontera)

# Crawling → Frontera





# Crawling → Control

- Existe un delay hasta recibir las respuestas
  - Eficiencia → múltiples conexiones (hilos). Cientos de páginas en paralelo
  - Cuidado con sobrecargar servidores (políticas de cortesía)
  - Robots.txt
    - User-agent: \*
    - Disallow: /privado/
    - Disallow: /usuarios/
    - Allow: /varios/publico/
    - Sitemap: <http://www.misitio.com.ar/sitemap.xml.gz>

# Sitemap ejemplo

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/sitemap/0.90">
  <url>
    <loc>http://www.sitemappro.com/</loc>
    <lastmod>2011-01-27T23:55:42+01:00</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://www.sitemappro.com/download.html</loc>
    <lastmod>2011-01-26T17:24:27+01:00</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://www.sitemappro.com/order.html</loc>
    <lastmod>2011-01-26T15:35:07+01:00</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://www.sitemappro.com/examples.html</loc>
    <lastmod>2011-01-27T19:43:46+01:00</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.5</priority>
  </url>
  ...
</urlset>
```

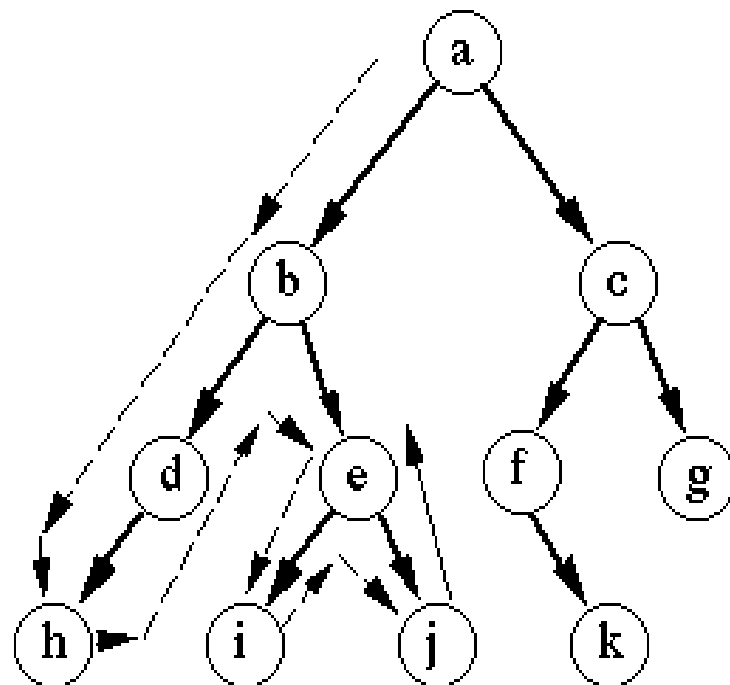


# Ejemplo [Manning]

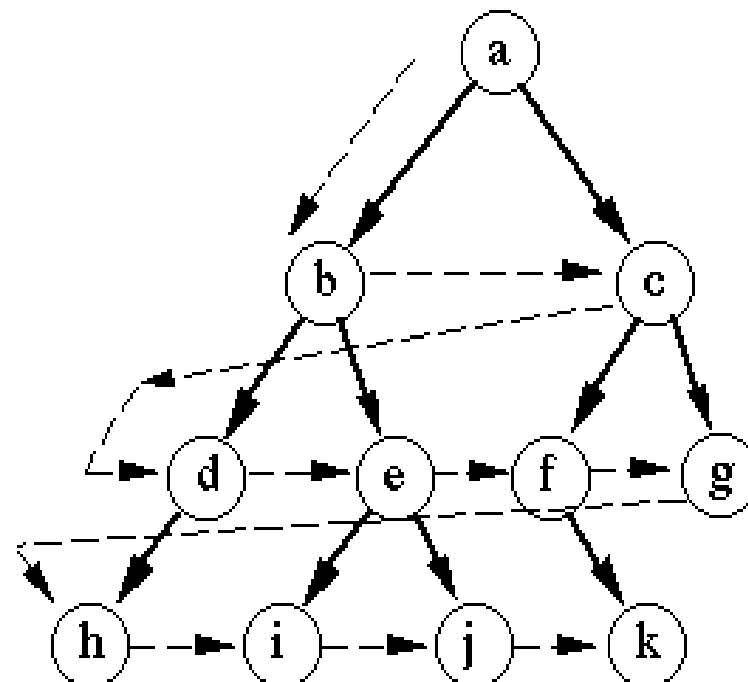
```
procedure CRAWLERTHREAD(frontier)
  while not frontier.done() do
    website ← frontier.nextSite()
    url ← website.nextURL()
    if website.permitsCrawl(url) then
      text ← retrieveURL(url)
      storeDocument(url, text)
      for each url in parse(text) do
        frontier.addURL(url)
      end for
    end if
    frontier.releaseSite(website)
  end while
end procedure
```

# Crawling → Estrategias

- Clásicas → Bread-First y Depth-First
- Otras → URL ordering



Depth-first search

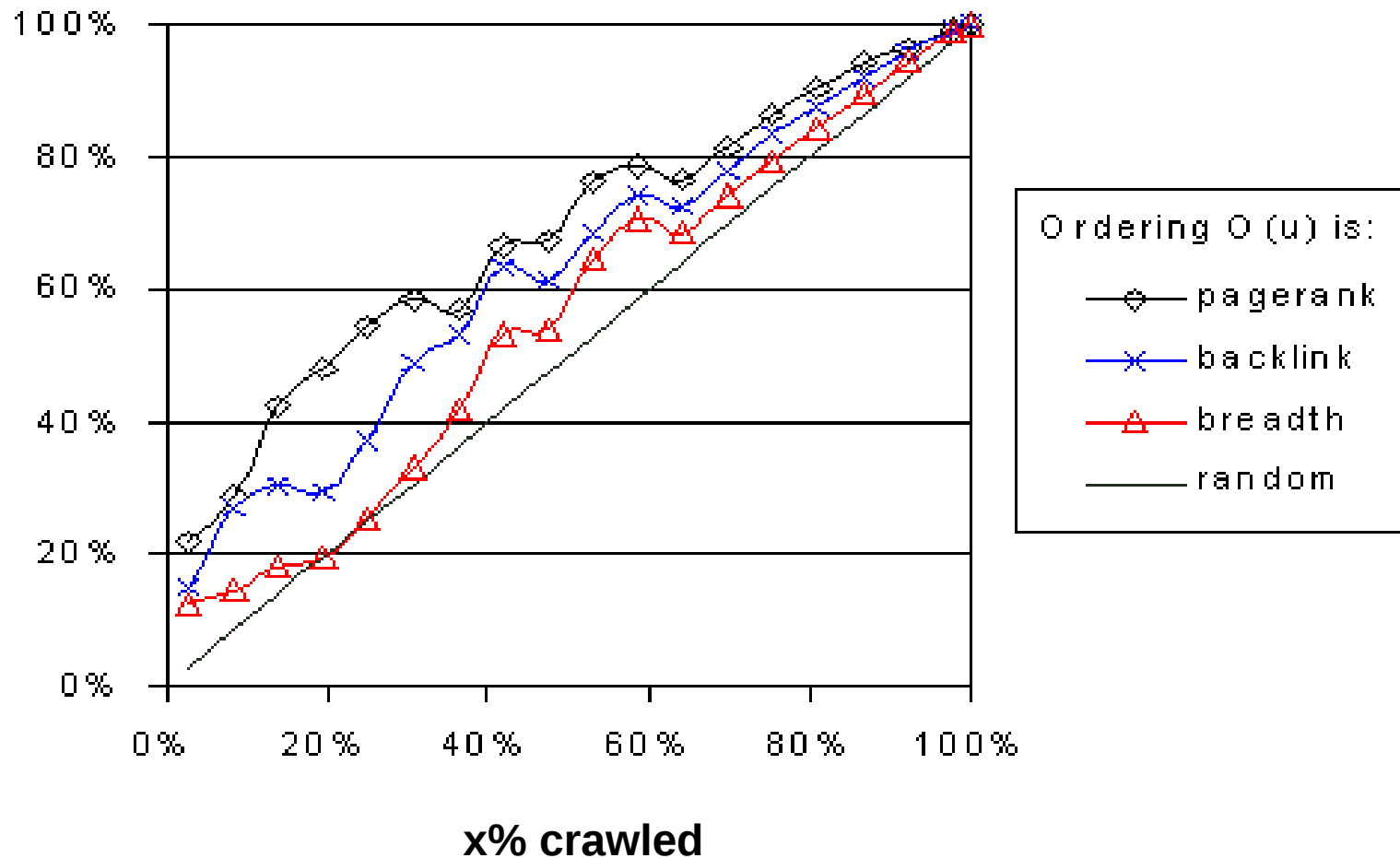


Breadth-first search



# Crawling → Estrategias

Overlap with  
best x% by  
indegree





# Crawling → Otras cuestiones

- Escalabilidad
- Crawling distribuido
- **Latencia/ancho de banda**
- **Profundidad**
- **Espejos/Duplicaciones**
- **Web SPAM → AIR**
- **DNS**
- **Robustez**
- **Cortesía/Estándares**
  - Explícita: robots.txt [[www.robotstxt.org/wc/norobots.html](http://www.robotstxt.org/wc/norobots.html)]
  - Implícita: No sobrecargar un servidor



# Queries



# Lenguajes de Queries

- No hay un lenguaje standard para queries web
  - No hay semántica explícita (c/ MB hace su interpretación)
  - Stemming, AND's... o no?
- Queries “Free-text” son el standard de facto
  - “Cualquier cosa” que el usuario escriba
  - No hay vocabulario controlado
  - Se aceptan errores de ortografía
  - Cuál es la diferencia con el lenguaje natural?
  - Cuál es la diferencia con una “pregunta”?

# Operadores comunes en MB

Operator Syntax	Details	Google	Yahoo! Search	Bing	Ask
".." double quotes surrounding a string	Phrase search	yes	yes	yes	yes
+ preceded by a space, operates on the term/phrase that immediately follows	This operator ensures that the associated term is included "as is" in the results	yes	yes	yes	yes
- preceded by a space, operates on the term/phrase that immediately follows, Bing uses NOT as well	This operator ensures that the associated terms do not appear in any result	yes	yes	yes	yes
OR (as well as  ) operates on preceding and succeeding terms or phrases	Equivalent to a Boolean OR	yes	yes	yes	yes
site: Followed by a site name	Returns results from the specific site only	yes	yes	yes	yes
hostname: Followed by a host name	Returns results from the specific host only	no	yes	no	yes
url: Followed by a URL	Checks that the following url exists in the engine index	no	yes	yes	no
inurl: Followed by a term	Returns results whose URL contains the specified term	no	yes	no	yes
intitle: Followed by a term	Returns results whose title contains the specific term	no	yes	yes	yes
inlink:/inanchor: Followed by a term	Returns results that contain the specific term in their link or anchor metadata	yes	no	yes	yes

# Consultas

- Existen diferentes motivaciones para usar un MB
- **Caracterización [Broder et al.]**
  - **Informational** – “saber” acerca de algo (~40% / 65%) **Algoritmos evolutivos**
  - **Navigational** – “ir” a algún lugar (~25% / 15%) **Aerolíneas Argentinas**
  - **Transactional** – “hacer algo” (web-mediante) (~35% / 20%)
    - Access a service **Clima en Luján**
    - Downloads **Imagen Ubuntu 11.04**
    - Shop **Canon S410**
  - **Areas “grises”**
    - Encontrar una buena páginas (HUB) **Alquiler auto Roma**
    - Explorar para “ver que hay allí”



# Consultas

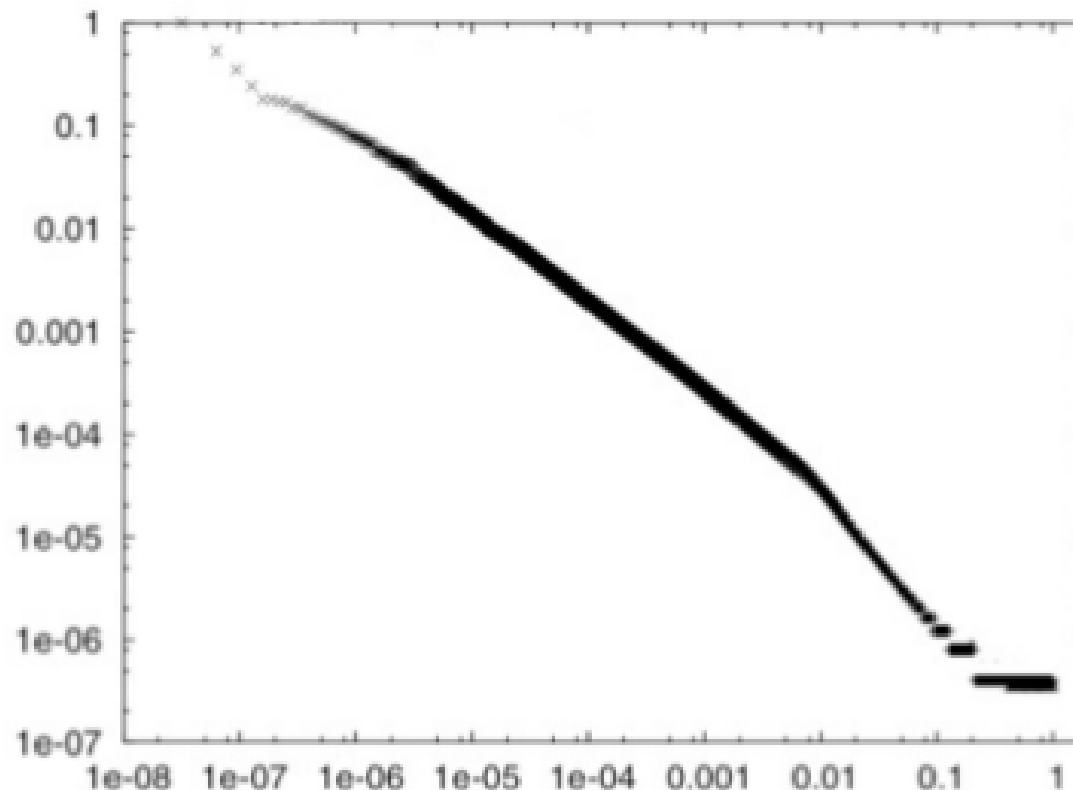
## Ejemplo:

- Juan quiere comprar una impresora – **Transactional Query**
- Encuentra 3 posibles impresoras pero quiere más info acerca de éstas – **Infomational Query**
- Luego, se decide por una Lexmark y necesita la URL donde comprar (Lexmark, eBay, Mercadolibre, etc.) – **Navigational Query**
- Juan necesita hacer la compra en línea de la elegida – **Transactional Query**

# Consultas

- La frecuencia de las consultas sigue una ley de Zipf con  $\beta = [0.6:1.4]$

Ejemplo: Yahoo! R. Baeza-Yates, A. Gionis, F. P. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri, "Design trade-offs for search engine caching," ACM Trans. Web, 2008.

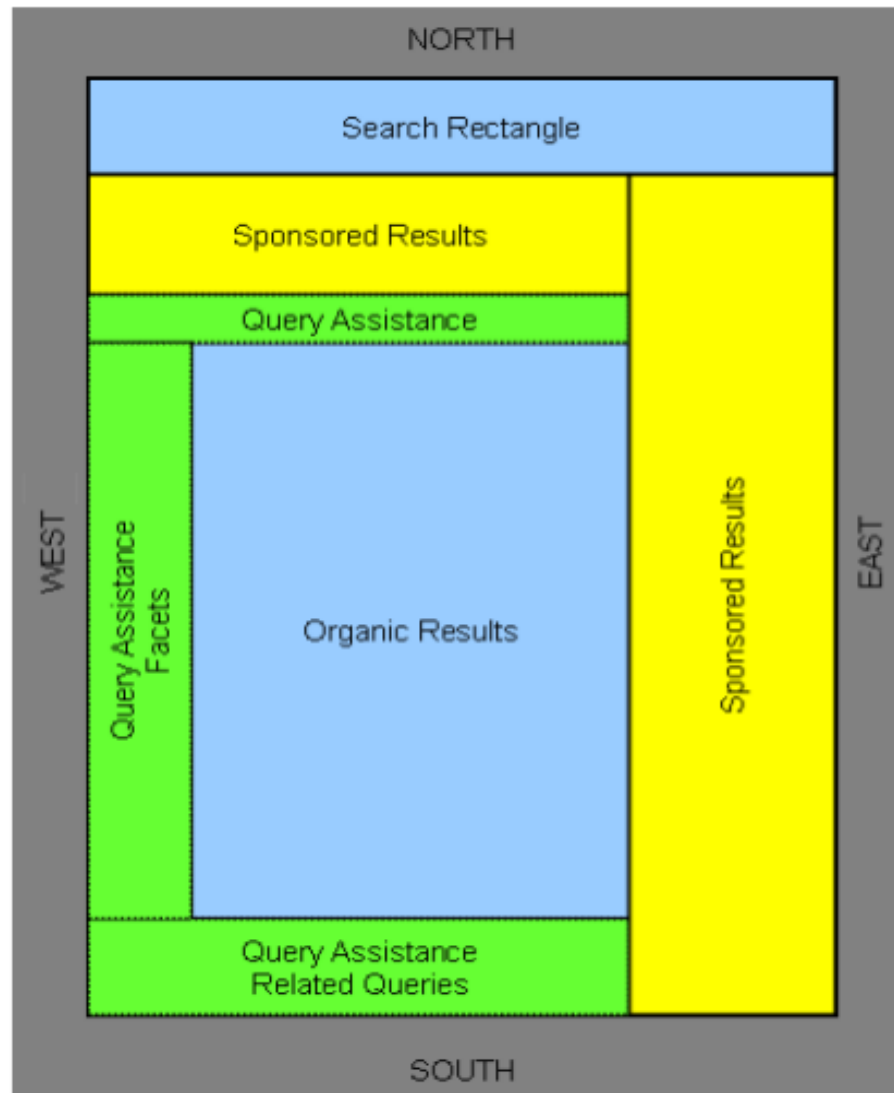






**Ranking**

# Resultados: SERP Layout



# Resultados



volkswagen voyage



Búsqueda avanzada

Búsqueda

Aproximadamente 1.620.000 resultados (0,14 segundos)

Todo

Imágenes

Videos

Noticias

Más

Chivilcoy, Buenos Aires

Cambiar ubicación

La Web

Páginas en español  
Páginas de Argentina  
Páginas extranjeras traducidas

Todos los resultados

Sitios con imágenes

Más herramientas

[Volkswagen Voyage 2011 - Estás Buscando Tu Nuevo 0Km?](http://www.volkswagen.com.ar/Voyage) Anuncios

[www.volkswagen.com.ar/Voyage](http://www.volkswagen.com.ar/Voyage) +1

Asesorate Con Expertos Acá!

Asesoramiento Comercial - Atención al Cliente - Amarak - Gol Trend

[Venta Autos Volkswagen | DeMotores.com.ar](http://www.demotores.com.ar/Concesionaria)

[www.demotores.com.ar/Concesionaria](http://www.demotores.com.ar/Concesionaria) +1

¿Buscás Autos Volkswagen? Todos los modelos en HausWagen

[Voyage > Modelos > Volkswagen Argentina](http://www.volkswagen.com.ar/ar/es/models/voyage0.html)

[www.volkswagen.com.ar/ar/es/models/voyage0.html](http://www.volkswagen.com.ar/ar/es/models/voyage0.html) +1

Sorprende por fuera. Sorprende por dentro., **Voyage**, Descubra un auto con excelente diseño, espacio, confort y versatilidad. Un sedan 4 puertas ...

[Imágenes de volkswagen voyage](#) - Informar sobre las imágenes



[Autos Volkswagen Voyage 0 km - DeMotores.com, compra y venta ...](http://autos.demotores.com.ar/vm-12-volkswagen-voyage)

[autos.demotores.com.ar/vm-12-volkswagen-voyage](http://autos.demotores.com.ar/vm-12-volkswagen-voyage) +1

Venta e información de Autos Volkswagen Voyage 0 km. Fichas técnicas, fotos, videos, reviews y vistas 360 de Volkswagen Voyage . Compra y venta de Autos ...

Anuncios

[Volkswagen Voyage 2011](http://www.espasavw.com.ar/voyage)

[www.espasavw.com.ar/voyage](http://www.espasavw.com.ar/voyage) +1

Conseguilo al mejor precio.

También financiación. Contactanos!

[Volkswagen Voyage 2011](http://www.concesionariasenred.com.ar)

[www.concesionariasenred.com.ar](http://www.concesionariasenred.com.ar) +1

Compra tu 0km - Representantes ofic

Llámanos 011-4762-0144

[Volkswagen en DeAutos](http://www.deautos.com/Volkswagen)

[www.deautos.com/Volkswagen](http://www.deautos.com/Volkswagen) +1

Venta de Volkswagen voyage Nuevos y Usados. Contratá el seguro online!

[Plan Volkswagen Retira Ya](http://www.modenamotorhaus.com.ar)

[www.modenamotorhaus.com.ar](http://www.modenamotorhaus.com.ar) +1

\$11000 de descuento y cuotas fijas. LLama ya:(011) 4343-0321/4343-0291.

[Mira tu anuncio aquí »](#)

# Resultados



Consulta los resultados traducidos de páginas web en inglés para:

[volkswagen voyage](#)

Búsquedas relacionadas con **volkswagen voyage**

[volkswagen voyage precio](#)

[volkswagen voyage diesel](#)

[volkswagen voyage colores](#)

[volkswagen voyage 2009](#)

[volkswagen voyage confortline plus](#)

[volkswagen voyage highline](#)

[test volkswagen voyage](#)

[volkswagen voyage ficha tecnica](#)



1 2 3 4 5 6 7 8 9 10

[Siguiente](#)

[Ayuda de búsqueda](#)

[Enviar comentarios](#)

[Google.com in English](#)

[Página principal de Google](#)

[Programas de publicidad](#)

[Soluciones Empresariales](#)

[Privacidad](#)

[Todo acerca de Google](#)



# Ranking

- Recuperación de Información
  - Términos incluir/excluir
  - Matching parcial → scoring
- **En la Web**
  - Frecuencia/ubicación de las palabras en el doc.
  - Metadatos
  - Existencia en directorio (si hay)
  - Tamaño/Edad del documento
  - Dominio
  - Y \$\$\$?



**+ Estructura  
de la WEB**



# Variables

De acuerdo a Matt Cutts [Ing. De Google] existen más de 200 variables que se tienen en cuenta para el ranking

- **Domain**
  - Age of Domain
  - History of domain
  - KWs in domain name
  - Sub domain or root domain?
  - TLD of Domain
  - IP address of domain
  - Location of IP address / Server
- **Architecture**
  - HTML structure
  - Use of Headers tags
  - URL path
  - Use of external CSS / JS files
- **Authority Link** (CNN, BBC, etc)
- **Content**
  - Keyword density of page
  - Keyword in Title Tag
  - Keyword in Meta Description
  - Keyword in KW in header tags (H1, etc.)
  - Keyword in body text
  - Freshness of Content
- **Per Inbound Link**
  - Quality of website linking in
  - Quality of web page linking in
  - Age of website
  - Age of web page
  - Relevancy of page's content
  - Location of link (footer, navig., body)
  - Anchor text if link
  - Title attribute of link
  - Alt tag of images linking
  - Country specific TLD domain
  - Authority TLD (.edu, .gov)
  - Location of server



# Variables

- **Cluster of Links**
  - Uniqueness of Class C address.
- **Internal Cross Linking**
  - No of internal links to page
  - Location of link on page
  - Anchor text of FIRST text link (Bruce Clay's point at PubCon)
- **Miscellaneous**
  - JavaScript Links
  - No Follow Links
- **Pending**
  - Performance / Load of a website
  - Speed of JS
- **Misconceptions**
  - XML Sitemap (Aids the crawler but doesn't help rankings)
  - PageRank (General Indicator of page's performance)
- **Penalties**
  - Over Optimisation
  - Purchasing Links
  - Selling Links
  - Comment Spamming
  - Cloaking
  - Hidden Text
  - Duplicate Content
  - Keyword stuffing
  - Manual penalties



# Variables

Los más importantes según Eric Smidt [CEO de Google]

- Uso de negrita alrededor del término
- Uso de “header-tags” alrededor del término
- Presencia del término en “Anchor-text” entrante
- Pagerank
- Pagerank / autoridad del sitio
- Velocidad del sitio
- Presencia del término en el título HTML (Title-Tag)



# Métricas complementarias

## Discounted Cumulative Gain

*“Mientras mas abajo se encuentre rankeado un documento relevante, menos útil es para el usuario (dada la probabilidad de que sea establecido)”*

**DCG:** *Es la ganancia acumulada en un ranking  $p$ :*

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i}$$

**Alternativa** (usada por algunas empresas)

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log(1+i)}$$



# DCG Ejemplo

Suponiendo 10 documentos rankeados en una escala 0-3

3, 2, 3, 0, 0, 1, 2, 2, 3, 0

Discounted gain:

3,  $2/1$ ,  $3/1.59$ , 0, 0,  $1/2.59$ ,  $2/2.81$ ,  $2/3$ ,  $3/3.17$ , 0

= 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0

• **DCG:** 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

**Normalized DCG:** *comparación con el ranking "perfecto"*

**Ejemplo:** Ranking perfecto: 3, 3, 3, 2, 2, 2, 1, 0, 0, 0

• **Valores ideales:** 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10

• **NDCG (dividir actual / ideal):** 1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88

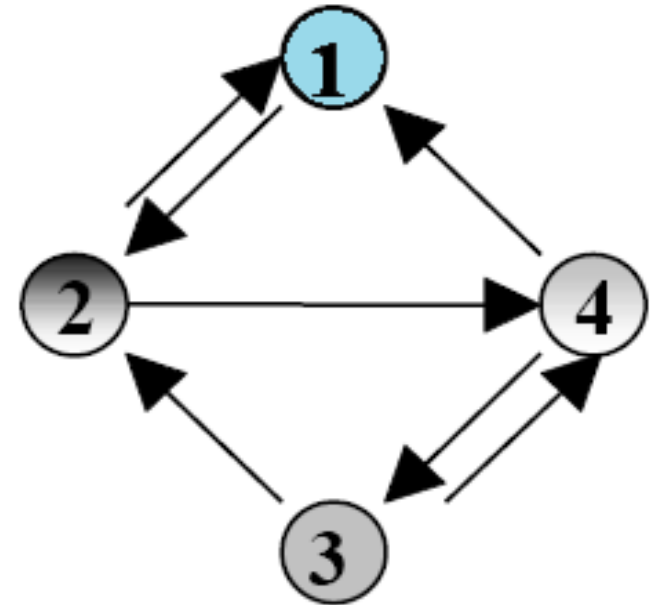


# Análisis de Enlaces

# 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)

X Representan una relación entre páginas conectadas

X Documento origen -> link

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

Doc. Destino

Anchor Text

X In-links → indegree

X Out-links → outdegree

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



# Estructura

## Suposiciones sobre la creación de enlaces

### X Recomendación

El autor recomienda la página destino

### X Localidad temática

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

### X “anchor text” descriptor

El texto del “ancla” describe el destino

### Para la indexación:

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



# Ranking

## Algoritmos de Ranking

### X Ranking basado en contenido

Modelos booleano, vectorial, etc.

### X Ranking basado en enlaces

Mediante el análisis de los enlaces se determina la calidad de la página

Clásicos: HITS [Kleinberg] y PageRank [Brin & Pag]

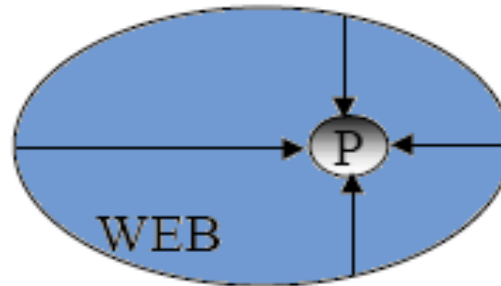
### X Combinación de los anteriores

# Análisis de enlaces

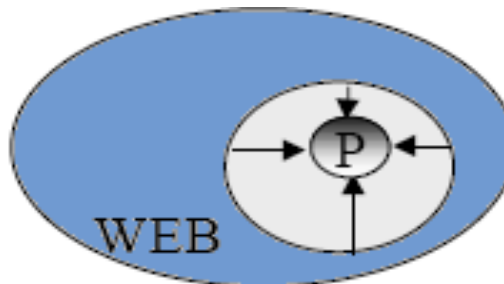
## Algoritmos

X Dos enfoques:

X Análisis global: la calidad de la página es **independiente** de la consulta



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





# Análisis de enlaces

## HITS – Hypertext Induced Topic Search

X Kleinberg, 1997.

X Identifica para un tema determinado (Query)

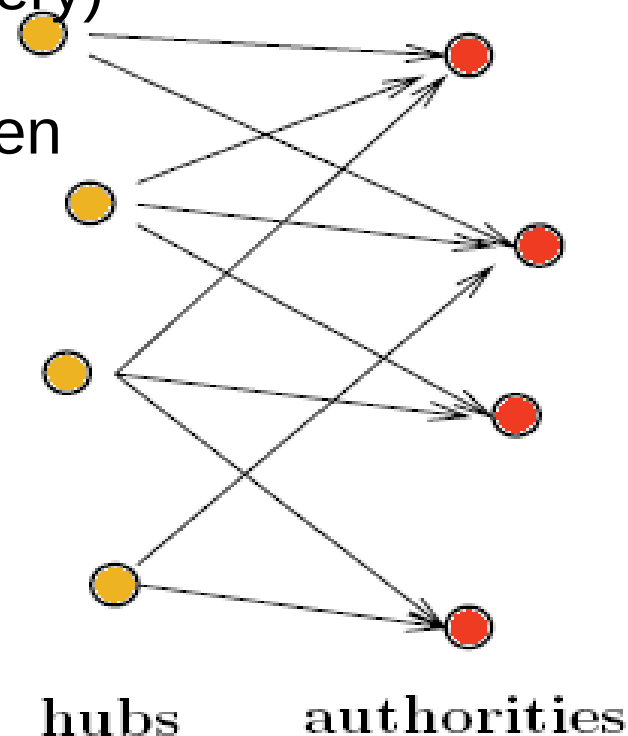
X Autoridades: Páginas que contienen información relevante respecto de Q.

X Hubs: Páginas que poseen links salientes ("apuntan") a páginas útiles.

X El valor de autoridad viene de los inlinks

X El valor de hub viene de los outlinks

X Refuerzo mutuo



# HITS

## Proceso

X Dado un query, identifica:

X Root set – top k relevantes

X Base set – vecinos-a-1

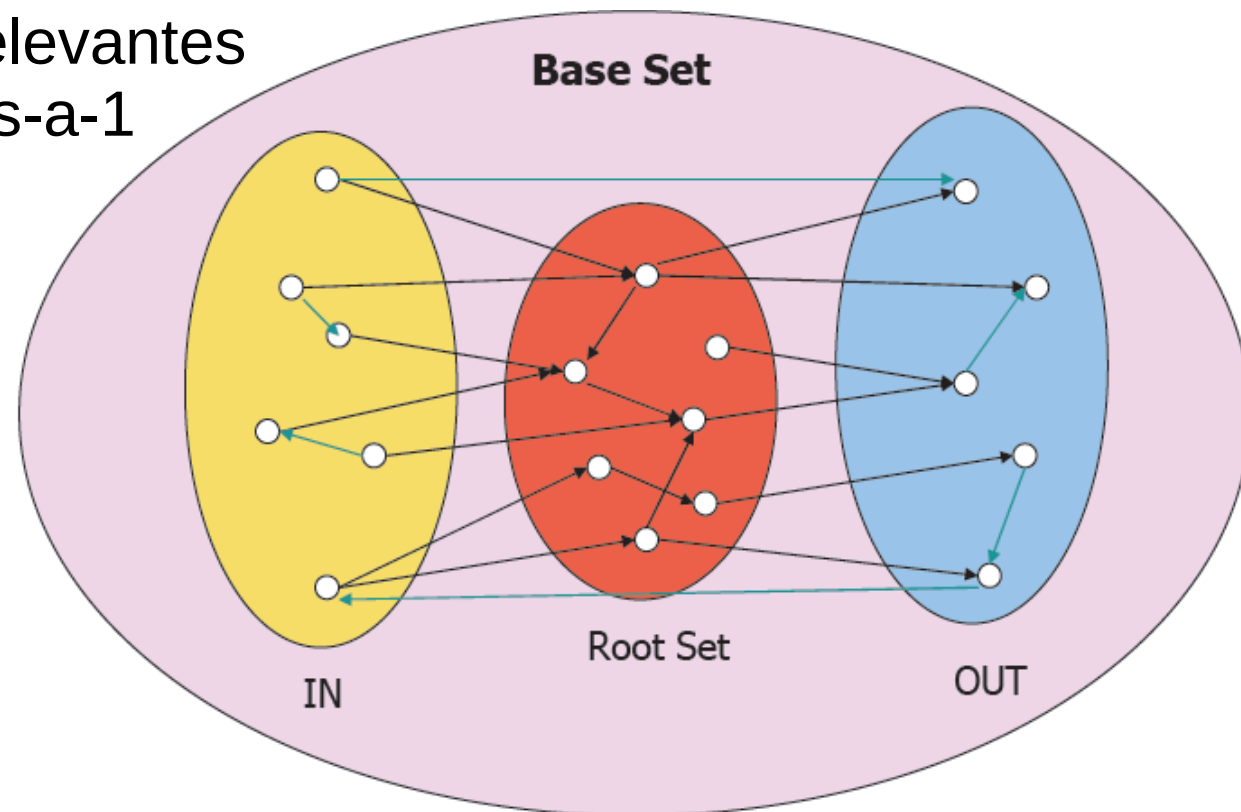
X Construye el grafo correspondiente

X Construye la matriz de adyacencia

X Calcula

X Hub-score

X 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

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

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

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

Normalizar los pesos

# Análisis de Enlaces

## PageRank

X Brin & Page, 1998.

X **Idea**: Una página es importante si otras páginas importantes apuntan a ésta (*Buenas autoridades apuntan a buenas autoridades*)

X Cada link entrante es un voto

X Entonces:

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

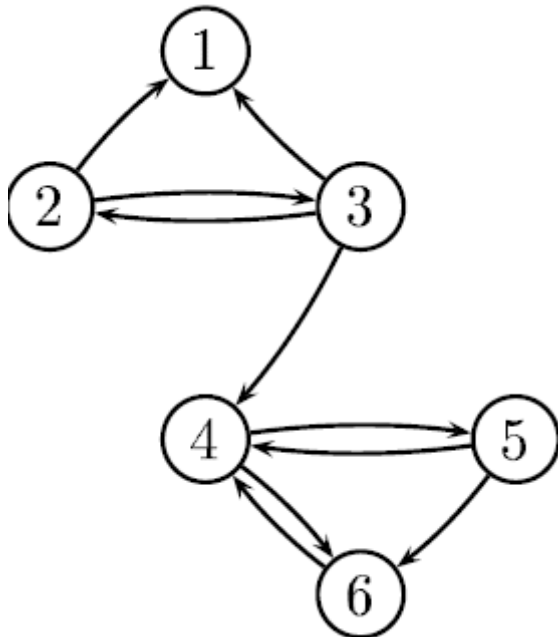
Donde:

$N_j$  → #de outlinks de  $P_j$

$L_i$  → Páginas que apuntan a  $P_i$

# Pagerank

- ✗ Random-Walk sobre el grafo web → Random Surfer
- ✗ Selecciona un página aleatoriamente
- ✗ Con probabilidad  $p$  "salta" a otra página



$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

✗ ¿Sink-Nodes?

# Pagerank

X Fórmula original:

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

Donde:

d: factor de damping (generalmente,  $d = 0.85$ )

$PR(T_i)$ : PageRank de la página  $i$

$C(T_i)$ : Outlinks de la página  $i$

Es decir:

$$PageRank(p) = (1 - d) + d \times \sum_{\substack{\text{all } q \text{ linking} \\ \text{to } p}} \left( \frac{PageRank(q)}{c(q)} \right)$$



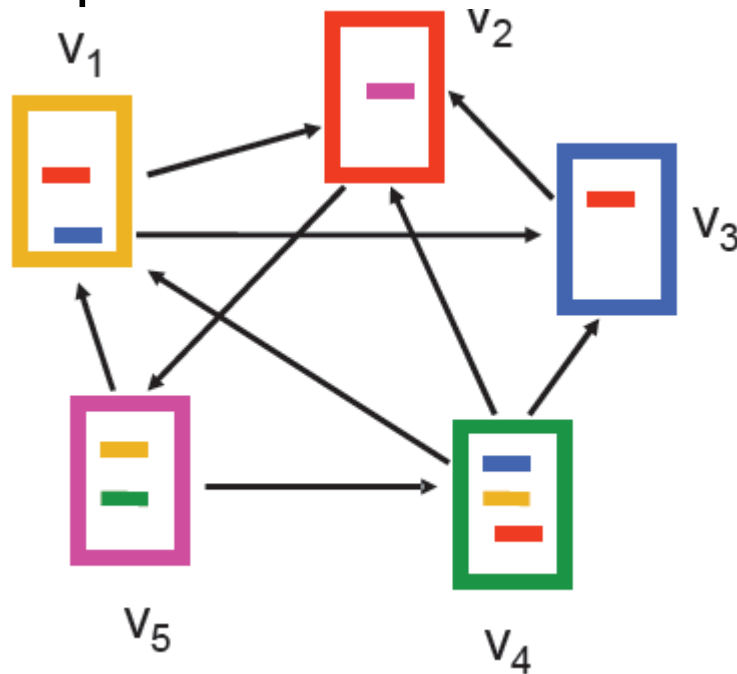
# Pagerank

Fórmula modificada:

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

# Pagerank

X Ejemplo



$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

$$q_1^{t+1} = 1/3 q_4^t + 1/2 q_5^t$$

$$q_2^{t+1} = 1/2 q_1^t + q_3^t + 1/3 q_4^t$$

$$q_3^{t+1} = 1/2 q_1^t + 1/3 q_4^t$$

$$q_4^{t+1} = 1/2 q_5^t$$

$$q_5^{t+1} = q_2^t$$

X Iterar hasta converger

X ¿Hay sesgo en el valor de PR?





# Pagerank

## X Algunas consideraciones:

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

# Comparación

<b>PageRank</b>	<b>HITS</b>
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 susceptible a ataques de Spam	Más susceptible a ataques de Spam
Más estable	Menos estable, la calidad depende del seed