

*Sistemas de  
Recuperación de  
Información Web*

RAI 2005

# Índice

---

Qué es un motor de búsqueda

Tipos de buscadores Web

Motores de búsqueda

- i. Características de los motores de búsquedas
- ii. Componentes de un motor de búsqueda

Evaluación de Motores Web

## ¿Qué es un Buscador?

---

Un **buscador** es un software que busca en una base de datos o repositorio documental, conforme a algunos criterios específicos.

- Tipos:
  - Directorios o índices
  - Motores de búsqueda
  - Metabuscaadores

# Directorios

---

- Sitio Web que gestiona una BD creada manualmente.
- Las URL están clasificadas en categorías.
- Características:
  - Selección y clasificación manual de recursos
  - Datos poco actualizados y poco exhaustivos
  - Resultados relevantes y páginas de calidad
  - Suelen ser temáticos
- Ejemplos

## Características de Yahoo!

---

- Catorce materias subdivididas en un número similar de subtemas. Bueno para Usabilidad.
- Se puede hacer una búsqueda general en cualquier sección o nivel. Si no encuentra resultados “salta” Yahoo!Search
- Cada resultado consiste en un título o una breve descripción.

# Motores de búsqueda

---

- Recolección de URLs e indexación automatizadas
- Muy exhaustivos
- Muy actualizados
- Manipulables
- Problemas con la calidad de los resultados y ambigüedad terminológica.

Ejemplos

# Metabuscadores

---

Software que agrega los resultados de varios motores o directorios para encontrar las páginas más relevantes.

Sin base de datos propia

Optimización por tiempos de respuesta

Incertidumbres sobre métodos de combinación de buscadores, pesos, orden de resultados, ...

Distintos tipos:

- Metabuscadores propiamente dichos
- Multibuscadores
- Agentes de búsqueda

Hay tantos q existen buscadores [Searchenginecollosus.com](http://Searchenginecollosus.com)

## Tipos de metabuscadores

---

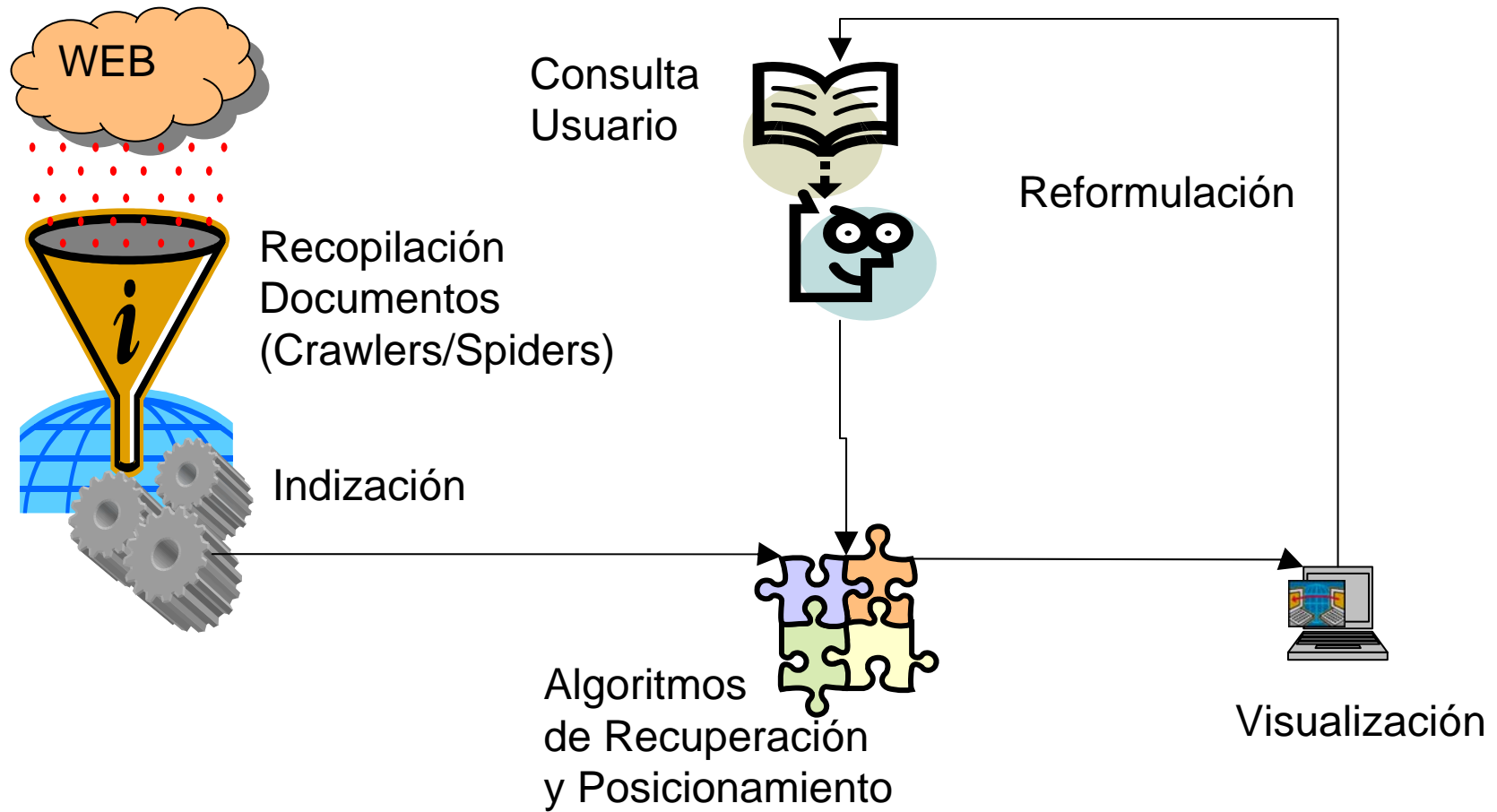
- **Multibuscadores:** no combinan los resultados, sólo lanzan la consulta en varios buscadores.
  - Ejemplos
- **Agentes de Búsqueda:** metabuscadores instalado localmente.
  - Ejemplos

# Ejemplos Metabuscadores

---

Profusion <http://www.profusion.com>

# Motores: Sistema de Recuperación Web



# Motor: Componentes

---

Cada Buscador tiene su propio motor: Altavista-Scooter, Lycos-Tres, Excite-Architext, Infoseek-Sidewinder, Google-Googlebot, ...

Componentes:

- Spider/Robot/Crawler. Robot.txt para ver directorios permitidos. Localizador y Recolector.
- Base de datos
- Indizador
- Interface de búsqueda

# Motores: Recopilación de documentos

---

- Lo realizan agentes de búsqueda (se les denomina spiders, crawlers, robots,...).
- Funcionamiento:
  1. Comienza en una página (A) y recopila todas sus URL
  2. Envía la página (A), comprueba que no está indizada y que no se tiene una versión menos actualizada, indiza la página (A)
  3. Recupera la página (B) que está primera en la lista
  4. Envía la página (B)...

# Motores: Recopilación de documentos

---

Criterios para organizar la lista a procesar:

Puede tener en cuenta novedad o prestigio.

También:

- Depth First Crawling: Hasta que no acaba con todas las páginas de un site no pasa a las del siguiente site.
- Breath Crawling: procesa primero las primeras páginas que ha encontrado en cada site, luego las segundas páginas de cada site, etc.

# Bases de datos

---

- Actualmente existen seis grandes bases de datos:
    - Google (8000 millones de páginas)
    - Yahoo
    - MSN
    - Teoma
    - Wisenut
    - Gigablast
    - Exalead (1000 millones pero mucha información)
- Los demás buscadores utilizan estas BD.

# Base de Datos: Ficheros Inversos

---

- Permite búsquedas rápidas de textos
- Cada término asociado a un conjunto de URLs y opcionalmente a su frecuencia y posición en cada URL.
  - Suele dividirse en glosario, con la frecuencia total y número de documentos
  - Y resto de datos: URL, posición o frecuencia en el documento.
- El glosario puede estar repetido y repartido en distintas máquinas.
- Una opción para acelerar es: las páginas más populares pueden estar en muchos servidores (y a ellos se acude primero), si no hay resultados se acude a unos pocos servidores que tienen las menos populares.

# Base de datos: Ficheros Inversos

Termino	#d	#frc	URL	Posición
A	99	128	uc3m.es	34, 45, 78
Arandela	1	2	uc3m.es	43, 67
Baraja	2	2	cartas.com	45
Casa	31	40	<a href="http://www.o.org">www.o.org</a>	33
...	..	..	...	..
...	..	..	...	..
...	..	..	...	..

**LEXICON**

**POSICIÓN (POSTINGS)**

# La base de Google

---

- “+”antes de una palabra no elimina aun siendo vacía, si se quiere buscar por frase poner comillas. “-” que no aparezca un término.
- No es lo mismo la ubicación geográfica desde donde hagamos la consulta (desde 2004)
- El orden de las palabras importa
- La misma consulta desde un mismo sitio con intervalo de segundos puede dar resultados distintos.
- No admite truncamiento, poner singular y plural
- No distingue mayúsculas, poner sin acentos
- Búsquedas por campos limitado
- Imposible combinar operadores booleanos de carácter distinto (todos AND todos OR pero no paréntesis)
- Aunque Google diga que existan 2000 resultados, jamás podrás pasar del resultado 1000.

# Búsqueda por campos en Google

---

- Descubrir vínculos que le apuntan link:www.google.com
- restricciones de búsqueda de un dominio "site:ejemplodedominio.com", "site:information"
- para encontrar información de artículos de prensa en el sitio de Google: press site:www.google.com
- Para que aparezca en el título: intitle, allintitle
- Para que aparezca en la url: inurl, allinurl
- Definition:"palabra"

## “Betas de siempre” de Google

- MoreGoogle, GoogleDesktop, Barra de Google
- Google Scholar, APIS (recuerda que van contra un servidor no actual), Gview
- Calculadora
- Google Suggest, profiles y demás google labs
- Y demás Gmail (los indiza para adwords), telefonía IP, ..

# Yahoo!Search

---

Yahoo fue el primer gran directorio hecho a mano, tardo cinco años en llegar al millón de páginas, mucha calidad, poca actualización, poco exhaustivo.

2004: compra AltaVista (por contenido de su BD), AlltheWeb (por contenido de su BD), Inktomi (por su algoritmo de posicionamiento y estructura de BD), Kelkoo (como comparador de precios) y Oberture

Comienza a utilizar Yahoo Search, con el motor de Inktomi

Posicionamiento: asignación de pesos denominada WebRank (parece relacionada con la barra de búsqueda), el interfaz y otros pesos de posicionamiento copiados de Google

Yahoo es el primer buscador en número de páginas

En 2004 muchos problemas con integración de BD, se quejan mucho de la calidad resultados...mejora mucho

# MSN

---

- Copia el algoritmo de posicionamiento de Google en 2004
- Problemas por instalación por defecto en las aplicaciones MS
- Actualmente el mejor valorado junto con Google y Yahoo

# A9 AskJeeves

---

- A9
  - Permite buscar a texto completo en libros de muchas editoriales
  - Pertenece a Amazon
- AskJeeves
  - Ha comprado a Excite, iWon y a Teoma
  - Siempre ha tenido algo de PLN y ha establecido comparación con news
  - Actualmente el motor de Teoma hace que AJ tenga los mejores rankings de precision. Teoma utiliza un criterio denominado autoridad basado en los enlaces de las páginas del mismo tema que apuntan a la página.
  - Es actualmente el cuarto buscador mundial

## Otros Raros

---

- Semantic Blogging Demonstrator (<http://www.semanticblogging.org/blojsom-hp/semnav.html>) que busca por un conjunto de metadatos más las cuestiones qué, quién, por qué, dónde.
- Exlead <http://beta.exalead.com/search/>
- NBII Clearinghouse <http://mercury.ornl.gov/nbii/> busca por metadatos.

## Internet invisible: conceptos y soluciones

---

**Internet invisible** sector de sitios y de páginas Web que no pueden indizar los motores de búsqueda de uso público

Aproximadamente el 70% del Web. Con un 50% más de tráfico que el visible (mayor calidad)

- P.e., OPACs , nombres de calles en mapas de ciudades, sitios que precisen de una password,...)
- "no indizable"
  - Formato de los documentos (no son html)
  - Formularios
  - Páginas generadas de forma dinámica, imágenes, ...
  - conjunto de sitios o de páginas web que, de forma expresa, se excluyen
- No todo es Web (p.e. P2P), no todo es realmente invisible.

## Internet invisible – Deep Net

---

Direct Search [www.freepint.com/gary/direct.htm](http://www.freepint.com/gary/direct.htm)

Turbo10 <http://turbo10.com>

Internet Invisible <http://www.internetinvisible.com>

Invisible Web <http://www.invisible-web.net/>

Librarian's Index to the Internet <http://www.lii.org>

Infomine <http://infomine.ucr.edu/>

Web Brain <http://www.webbrain.com>

Science.gov <http://science.gov/>

Easy searcher <http://www.easysearcher.com>

# Criterios de evaluación de motores

---

## Nielsen NetRankings

Actualización, tamaño, spam, enlaces muertos, cobertura según tema y área geográfica...

**Actualización** Google, Hotbot, MSN y Alltheweb tardan poco, las más antiguas en Alltheweb todos pasa cada mes. MSN y hotbot los mejores. Altavista tardaba tres meses. Google parece tener el Google Dance cada 15 días y los mirrors nacionales cada mes.

**Tamaño:** la mayor son Google 8100 millones (101 k por pg), MSN 5000 billones (150 kpag) y Yahoo!Search 4200 millones (500k) y por último AJ 2500 millones a 101 k . Directorios manuales DMOZ y looksmart 2, 5 millones Yahoo 1,8 millones

**Enlaces muertos:** Enlaces que conducen a enlaces muertos, en 2000 Altavista tenía un 14% mientras que Google tenía un 4% → se hunde Altavista

**Cobertura:** Páginas que aparecen en un único buscador: casi la mitad están en Google, pero tb destacan WiseNut y Yahoo

Los más **usados** en España msn 36% google 30% terra 20%. En el mundo google 41% yahoo!Search 31 MSN 27%

**Tiempo** que se permanece en Google 29 m AOL 28 Netscape 13