

Índice

- 1 Tema 1: Internet
- 2 Tema 2: La web
- 3 Tema 3: Principios de Recuperación de Información
- 4 Tema 4: Búsqueda de información en la web**
- 5 Tema 5: Búsqueda avanzada en la web
- 6 Tema 6: Integración de conocimiento lingüístico
- 7 Tema 7: Más allá de la búsqueda textual

Buscadores

- Tamaño web (febrero 2007): $\pm 30,000,000,000$ páginas
- ¿Cómo encontrar algo?
- Sitios web especializados en buscar otros sitios web (**buscadores**):
 - **Directorios**: jerarquizados por temas y categorías
 - Google Directory (<http://directory.google.com/>)
 - Yahoo! Directory (<http://dir.yahoo.com>)
 - **Motores de búsqueda** (o buscadores): búsqueda por palabras clave
 - Google (<http://www.google.es>)
 - Yahoo! (<http://www.yahoo.es>)

Estructura de la web

PUBLICA

OCULTA

INDEXABLE

ESTATICA

DINAMICA

El buen buscador

- Manejo sencillo e intuitivo
- Rápido
- Actualización constante de contenidos
- Resultados claros y ordenados
- *Búsquedas avanzadas*

Breve historia de los buscadores

1990	Archie	Primer buscador de Internet (FTP)
1992 Dic	Veronica	Buscador de Gopher (menús jerarquizados)
1993 Jun	Wanderer	Primer buscador web
1993 Dic	RBSE	Primero en calcular medida relevancia
1994 Ene	Galaxy	Primer directorio
1994 Abr	Yahoo	Directorio revisado manualmente
1994 Abr	WebCrawler	Salto tecnológico: indexar texto completo
1994 Jul	Lycos	Índice masivo
1995 Feb	Infoseek	Netscape. Amigable, servicios adicionales
1995 Jun	Metacrawler	Primer metabuscador
1995 Dic	Altavista	Muy veloz. Lenguaje natural y ops. lógicos

Breve historia de los buscadores (cont.)

1996	Abr	Olé	Primer buscador hispano
1996	May	HotBot	Tecnología de búsqueda de alto rendimiento
1998		MSN Search	Buscador de Microsoft
1998	Sep	Google	Nuevo salto tecnológico: algoritmo <i>pagerank</i>
1999		Baidu	Buscador chino
2005	Nov	Live Search	Nueva plataforma <i>Windows Live</i> de Microsoft
2006		Quaero	"Buscador europeo"

<http://manuales.ojobuscador.com/historia>

Directorio

- Sitio web que contiene un índice o lista de páginas web estructuradas jerárquicamente en base a categorías y subcategorías temáticas

Yahoo! Directory (<http://dir.yahoo.com>)

Google Directory (<http://directory.google.com/>)

- Estructura navegable
- Generalmente creado/revisado a mano
 - Categorización automática
- Han ido perdiendo importancia frente a los motores de búsqueda
- Actualmente son un "complemento" a éstos
- Para búsquedas muy generales

Ejercicios

- Busca en Yahoo! Directory (<http://dir.yahoo.com>) documentos sobre *revistas de aikido*
- Ahora en Google Directory (<http://directory.google.com/>)
- Comprueba si están o no categorizados de igual forma en ambos directorios

Motor de búsqueda

- Sitio web que contiene una base de datos (índice) donde las páginas web han sido indexadas en base a palabras clave y sobre la cual podemos realizar búsquedas (consultas o *queries*)

Google (<http://www.google.es>)

Yahoo! (<http://www.yahoo.es>)

Live (MSN) Search (<http://www.live.com/>)

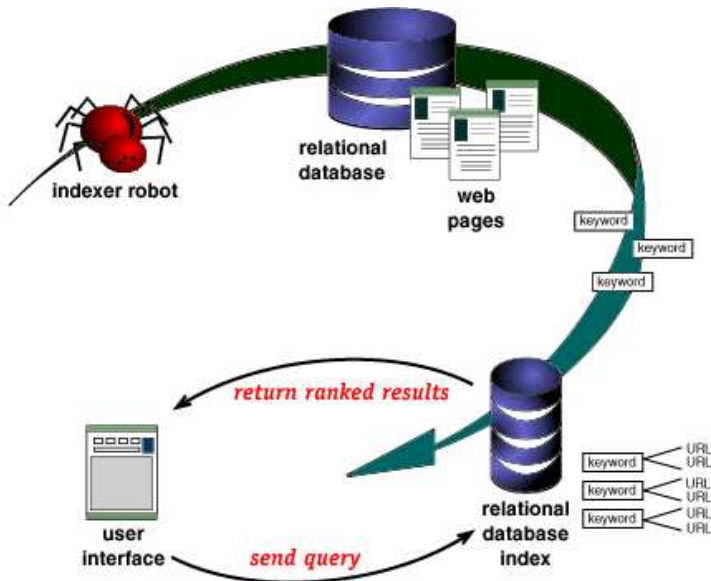
Altavista (<http://www.altavista.es>)

- Ante una consulta:
 - 1 Busca correspondencias en la base de datos
 - 2 Presenta las páginas web encontradas, por orden de relevancia
- Para búsquedas concretas

Ejercicios

- Lanzar en Yahoo! y en Google algunas de las siguientes consultas:
 - duelo Hamilton Alonso en Brasil
 - incidentes antimonárquicos en España
 - PhotoGalicia (ver diferencias con Photo Galicia)

Funcionamiento de un motor de búsqueda



Arquitectura de un motor de búsqueda

- **Robots:** Programas que recorren la red buscando documentos:
 - Analizan su contenido (total o parcial)
 - Devuelven las palabras clave o descriptores que lo describen (a indexar)
- **Base de datos:** índice de palabras clave o descriptores asociados a cada documento
 - Actualización periódica (robots)
- **Interfaz de consulta:** parte que ve el usuario
 - Introducir consulta
 - Presentar resultados

Palabras clave

- Representación del tema buscado
- Buena búsqueda = palabras clave adecuadas
- Evitar palabras demasiado generales, emplear términos específicos
sillas del siglo XIX vs. muebles antiguos

- Evitar palabras ambiguas:

reparaciones de pisos \rightarrow $\left\{ \begin{array}{l} \text{viviendas ?} \\ \text{suelos ?} \end{array} \right.$

- Evitar palabras "demasiado" comunes (*stopwords*): preposiciones, artículos, etc.
 - Eliminación automática

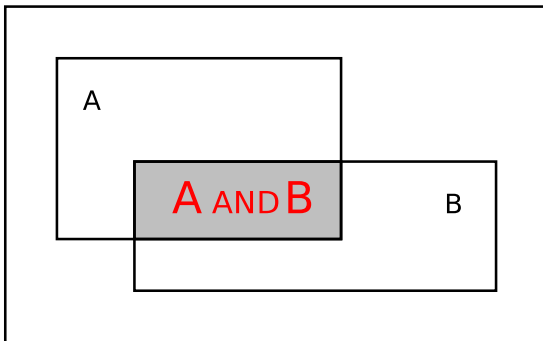
Palabras clave (cont.)

- Comprobar ortografía: coches≠cches
 - Autocorrección
- En minúsculas (salvo propios) y sin tildes
 - Algunos buscadores diferencian
- El orden puede influir
 - Primero las palabras más específicas y/o relevantes
- **Casi nunca se acierta a la primera**
 - Refinar la consulta sucesivamente

Operadores básicos

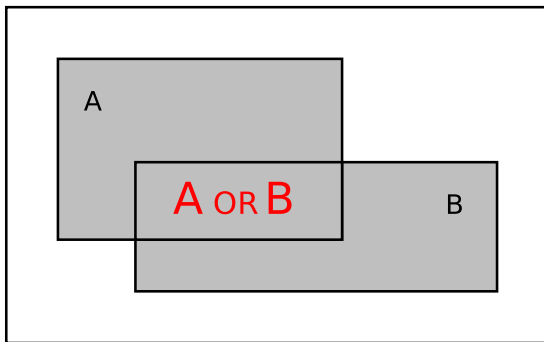
- Permiten ampliar, acotar y dirigir la búsqueda
- Interfaz disponible en Búsquedas avanzadas

Operadores básicos (cont.): AND (Y lógico)



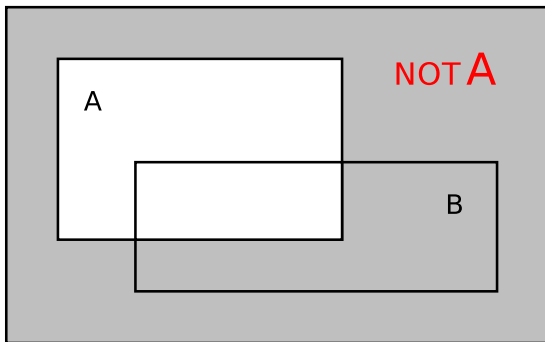
- Operador por defecto en Google y Yahoo (otros?)
- Correspondencia con TODAS las palabras
- Retringe la búsqueda
- Sintaxis:
 - **Google:** coche rojo coche AND rojo
 - **Yahoo:** coche rojo +coche+rojo

Operadores básicos (cont.): OR (O lógico)



- Correspondencia con ALGUNA (O TODAS!) las palabras
- Amplía la búsqueda
- Sintaxis:
 - **Google/Yahoo:** coche OR rojo

Operadores básicos (cont.): NOT (*NO lógico*)



- EXCLUYE la palabra
- Restringe la búsqueda
- Sintaxis:
 - **Google/Yahoo:** coche -rojo

Operadores básicos (cont.): frase

- Busca la SECUENCIA EXACTA (mismas palabras y mismo orden)
- Restringe la búsqueda
- Sintaxis:
 - **Google/Yahoo:** "coche rojo"

Presentación de resultados

- Google como ejemplo:

conceptos basicos internet

Calculando la relevancia: indicadores

- *Peso* de la palabra
- Frecuencia de la palabra en el documento y la consulta
- Número de palabras de la consulta con correspondencia
- Longitud del documento
 - Ponderar frecuencia
- Correspondencia exacta de la palabra frente a variantes
- Mismo orden

Calculando la relevancia: indicadores (cont.)

- Proximidad entre sí dentro del documento
- Posición en el texto
 - Mejor en título y encabezamientos
- Presencia en las etiquetas <META> (al crear página)
- Popularidad del documento (número de enlaces que lo referencian)
- Valoración de los usuarios
- Enlaces patrocinados
- "Trampas" y "castigos"