
TEMA
Recuperación de la Información en Internet
UNIVERSIDAD CARLOS III DE MADRID
2008-2009
Jorge Morato Lara
Sonia Sánchez-Cuadrado
UNIVERSIDAD CARLOS III DE MADRID


Recuperación de Información en Internet
1. Buscadores Web
Tipología
Componentes y funcionamiento
Ejemplos de Motores de Búsqueda
Buscadores Web especiales
2. Internet Invisible
3. Posicionamiento en Recuperación Web
4. Optimización de Páginas Web
5. Criterios de Evaluación de buscadores
UNIVERSIDAD CARLOS III DE MADRID

¿Qué es un Buscador Web?

Un **buscador (*search engine*)** es un software que recopila e indexa archivos almacenados en servidores web y recupera conforme a algunos criterios específicos.

Objetivos

- Recopilar la información de la red
- Indizar la red constantemente para permitir la consulta de sus índices
- Encontrar los documentos que contengan las palabras clave introducidas por el usuario.

Tipos:

- Directorios o índices temáticos
- Motores de búsqueda
- Metabuscadores

Tipología: Directorios

¿Qué es un Directorio?

- Sitio Web que gestiona una BBDD gestionada según criterios manuales
- Las URL están clasificadas en categorías temáticas
- Suele estar organizado por temas y categorías jerárquicas

Características

- Selección y clasificación manual de recursos
- Datos poco actualizados y poco exhaustivos
- Resultados relevantes y páginas de calidad
- Suelen ser temáticos

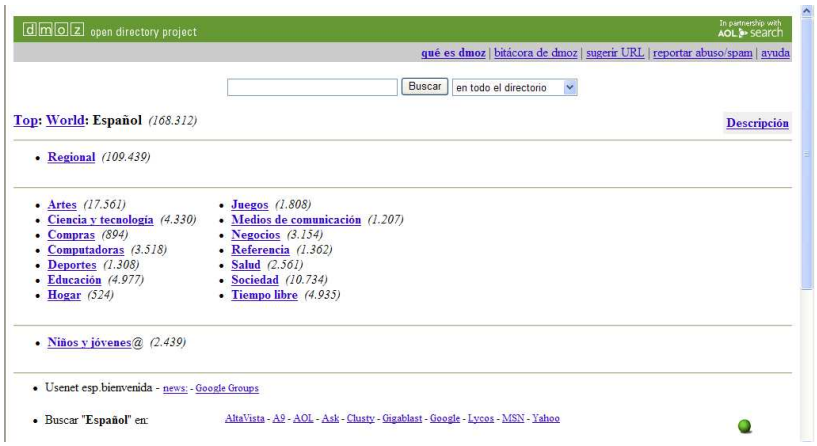
Tipología: Directorios

- Argus Clearinghouse
- DMOZ (The Open Directory Project -ODP- o Proyecto de Directorio Abierto)
- Yahoo!
- Google Directory
- Buscador.com http://www.buscador.com
- Lasonet.com www.lasonet.com es un directorio temático de enlaces a las páginas de Internet organizadas por temas, prensa, bancos, pueblos, buscadores, planos, etc.
- Web Directorio http://www.web-directorio.com
- Bubl http://bubl.ac.uk/ (Con la clasificación Dewey)
- Galaxy http://www.galaxy.com
- IPL the Internet Public Library http://ipl.org/ref
- Intute (RDN) http://www.intute.ac.uk

Directorios temáticos: Yahoo!

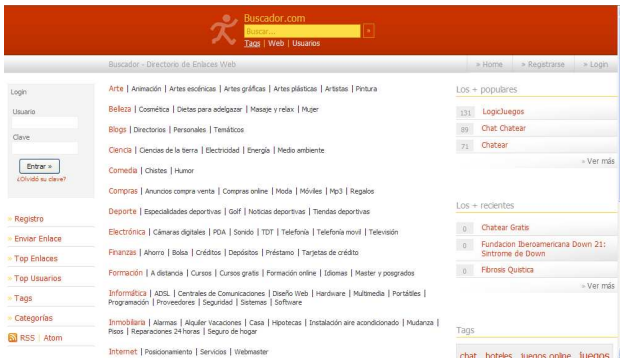
- Catorce materias subdivididas en un número similar de subtemas. Bueno para Usabilidad.
- Se puede hacer una búsqueda general en cualquier sección o nivel. Si no encuentra resultados "salta" Yahoo!Search
- Cada resultado consiste en un título o una breve descripción.

Directorios temáticos: Dmoz/Open Directory Project



UNIVERSIDAD CARLOS III DE MADRID

Directorios temáticos: Características de Buscador



UNIVERSIDAD CARLOS III DE MADRID

Directorios temáticos: Características de Google

Tipología: Motores de búsqueda

¿Características de los motores de búsqueda?

- Software con un sistema de recolección de URLs e indización automatizadas
- Generales o temáticos (buscadores verticales)

Características

- Muy exhaustivos
- Muy actualizados
- Manipulables
- Problemas con la calidad de los resultados y ambigüedad semántica

Tipología: Motores de búsqueda

Ejemplos

- Google
- MSN - Microsoft
- Teoma/AskJ
- Yahoo! Search
- Exalead
- Virgilio.it

- Searchenginewatch http://www.searchenginewatch.com/

Tipología: Motores de búsqueda

- **Generales** (Google, Alexa, Ask, Exalead, Wikia Search, Yahoo!, Msn, Yandex (ruso http://www.yandex.ru) Baidu (Chino http:www.baidu.com) Cuil (http://www.cuil.com))

- **Temáticos**
  - De alcance limitado geográficamente
  - Financieros
  - De negocios
  - Para la empresa
  - Dispositivos de búsqueda
  - Empleo
  - Legal
  - Médicos
  - Noticias
  - Personas
  - Propiedad inmobiliaria

Tipología: Motores de búsqueda

- Temáticos
  - Video juegos
  - Foros
  - Blogs
  - Multimedia
  - Código fuente
  - BitTorrent
  - Email
  - Mapas
  - Precio
  - Preguntas y respuestas (answers.com)
  - Motores de búsqueda de código abierto (Google Code Search, Jexamples, Koders, Krugle)
  - Motores de búsqueda sociales
  - Motores de búsqueda visuales
  - Usenet
  - Buscadores basados en otros buscadores

Tipología: Motores de búsqueda

Extintos

- BRS/ Search
- Google Answers
- Infoseek
- Inktomi
- Lotus Magellan
- Noxtrum
- Overture.com
- PubSub
- Singingfish
- WiseNut
- World Wide Web Worm

Transformaciones

- Web Crawler (ahora metabuscador)
- Excite (ahora metabuscador)
- Teoma /AskJ

Tipología: Metabuscadores

¿Qué son?

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

Características

- Sin base de datos propia
- Optimización por tiempos de respuesta
- Incertidumbres sobre métodos de combinación de buscadores, pesos, orden de resultados, …

Tipos:

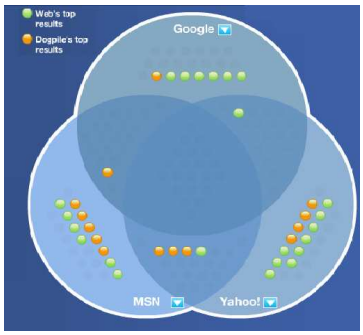
- Metabuscadores propiamente dichos
- Multibuscadores
- Agentes de búsqueda

Ejemplos

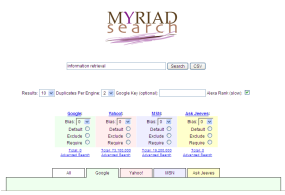
- MetaCrawler, Mibúsqueda, Copernic

UNIVERSIDAD CARLOS III DE MADRID

Tipología: Metabuscadores



Ejemplo: comprueba cuantas páginas únicas existen entre las primeras de cada buscador [Google, Yahoo, MSN y AJ] en promedio es un 1%
Comparar



UNIVERSIDAD CARLOS III DE MADRID

Tipos de metabuscadores

- Metabuscadores: combinan los resultados, lanzando una consulta en varios buscadores. Ejemplo: MetaCrawler
- Multibuscadores: no combinan los resultados, sólo lanzan la consulta en varios buscadores. Ejemplo: My Search (MiBúsqueda)
- Agentes de Búsqueda: metabuscadores instalado localmente. Ejemplo: WebFerrer, Copernic

Ejemplos Metabuscadores

- **MetaCrawler** www.metacrawler.com *Elimina los duplicados*
- **Dogpile** www.dogpile.com *motores distintos según categoría*
- **Vivisimo** www.vivisimo.com/ *con clusters y posición en cada buscador*
- **Kartoo** http://www.kartoo.com/ *mapas de **clusters** navegables y con **expansión de consultas***
- **Myway** http://myway.com/
- **SurfWax** http://www.surfwax.com/ (en la opción focus con **expansión de consultas en inglés mediante tesauro**)
- **Ixquick** http://www.ixquick.com/ buen metabuscador (cada estrella un buscador) con **refinamiento** de búsqueda
- **Beaucoup** http://www.beaucoup.com/ combina un metabuscador con un directorio y con términos específicos)

Recuperación de Información en Internet

1. Buscadores Web
   - Tipología
   - Componentes y funcionamiento
   - Ejemplos de Motores de Búsqueda
   - Buscadores Web especiales
2. Internet Invisible
3. Posicionamiento en Recuperación Web
4. Optimización de Páginas Web
5. Criterios de Evaluación de buscadores

Funcionamiento MB: Componentes

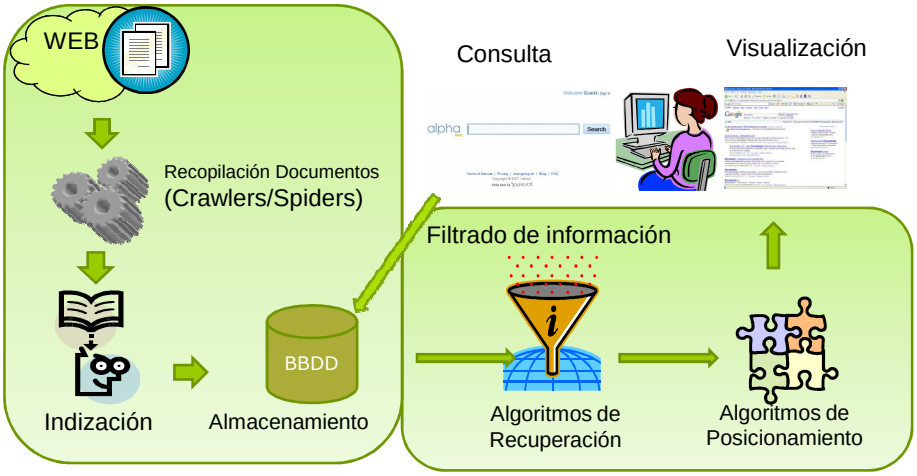
Componentes

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

Cada Buscador tiene su propio motor:

- Altavista-Scooter,
- Lycos-Tres,
- Excite-Architext,
- Infoseek-Sidewinder,
- Google-Googlebot,…

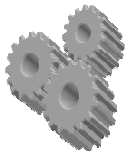
Motores de Búsqueda Web: Funcionamiento



UNIVERSIDAD CARLOS III DE MADRID

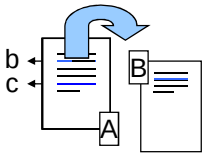
Motores de búsqueda: Recopilación de documentos

Spider, *crawlers*, *robots* o Agentes de Búsqueda son los nombres que recibe el software que recopila los documentos.



Funcionamiento

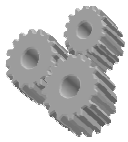
- Comienza en una página (A) y recopila todas sus URL
- Envía la página (A), comprueba que no está indizada y que no se tiene una versión menos actualizada, indiza la página (A)
- Recupera la página (B) que está primera en la lista
- Envía la página (B)...



UNIVERSIDAD CARLOS III DE MADRID

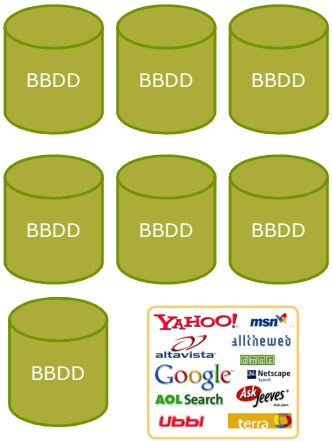
Motores de búsqueda: Recopilación de documentos

Criterios para organizar la lista a procesar:



- Puede tener en cuenta novedad o prestigio.
- 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.

Motores de búsqueda: Bases de datos

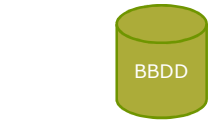


- Grandes **bases de datos:**
  - Google
  - Yahoo
  - MSN
  - Teoma
  - Wisenut/LookSmart
  - Gigablast
  - Exalead /Quaero

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



Término	#d	#frc
A	99	128
Arandela	1	2
Baraja	2	2
Casa	31	40
Tecla	1	1
…	..	..
…	..	..

LEXICON

URL	Posición
uc3m.es	34, 45, 78
uc3m.es	43, 67
cartas.com	45
www.o.org	33
…	..
…	..
…	..
…	..

POSICIÓN (POSTINGS)

Base de datos: Ficheros Inversos

Término	#d	#frc
A	99	128
Arandela	1	2
Baraja	2	2
Casa	31	40
Tecla	1	1
…	..	..
…	..	..

LEXICON

URL	Posición
uc3m.es	34, 45, 78
uc3m.es	43, 67
cartas.com	45
www.o.org	33
…	..
…	..
…	..
…	..

POSICIÓN (POSTINGS)

Ejemplo: La base de Google

Característica de la base de datos 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, es decir 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.

Motores de Búsqueda: opciones búsqueda por campos en Google

Opciones de búsqueda por campos

- 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)
- Calculadora
- Google Suggest, profiles y demás google labs
- Y demás Gmail (los indiza para adwords),telefonía IP,..

Recuperación de Información en Internet

1. Buscadores Web
   - Tipología
   - Componentes y funcionamiento
   - Ejemplos de Motores de Búsqueda
   - Buscadores Web especiales
2. Internet Invisible
3. Posicionamiento en Recuperación Web
4. Optimización de Páginas Web
5. Criterios de Evaluación de buscadores

Motores de búsqueda: 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**

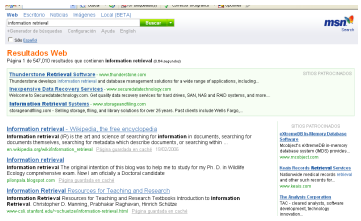
Comienza a utilizar **Yahoo Search**, con el motor de Inktomi

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

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

Motores de búsqueda: 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!



Ejemplos: A9 y 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

Recuperación de Información en Internet

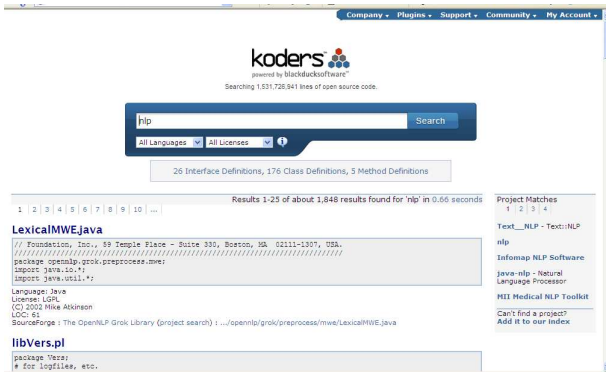
1. Buscadores Web
   - Tipología
   - Componentes y funcionamiento
   - Ejemplos de Motores de Búsqueda
   - Buscadores Web especiales
2. Internet Invisible
3. Posicionamiento en Recuperación Web
4. Optimización de Páginas Web
5. Criterios de Evaluación de buscadores

Buscadores Web especiales

1. Por su temática
2. Por su tecnología y recursos
3. Por los servicios que ofrece

Búsqueda de Código

- Koders http://www.koders.com/



Búsqueda de Código

- Google Code Search (http://www.google.com/codesearch)

Buscadores especializados

- Textos especializados y académicos → Digital.CSIC Acceso Abierto a Documentos Digitales http://digital.csic.es/
- Textos especializados y académicos →Citeseer
- Textos especializados y académicos →Google Scholar
- Videos →YouTube
- Imágenes

Scirus (for scientific information only)

- **Scirus** es un motor de búsqueda específico de contenido científico. Semejante a Citeseer y Google Scholar, está enfocado a la información científica.A diferencia de Citesser, Scirus no se centra sólo en ciciencias informáticas y no todo su contenido es de libre acceso (p.e. algunos resultados pueden estar en PubMed en cualquier revista de Elsevier, requieriendo suscripción para su acceso. www.scirus.com

Recuperación de imágenes

- Tradicionalmente con metadatos, texto asociado a la imagen por nombre del fichero, texto de la página, etc
- Con la Web social con descriptores de los usuarios, p.e. Flickr

Otros buscadores de este tipo son:

- Pixsy imágenes y videos de Buzznet, flickr, iStockphoto, Fotolia, YouTube entre otros
- Riya buscador que incluye a Google, Yahoo, MSN, y flickr.
- Tiltomo basado en flickr
- Liveplasma para buscar música y video, busca por artista, grupo, película, director o actor.
- Vdoogle para buscar videos

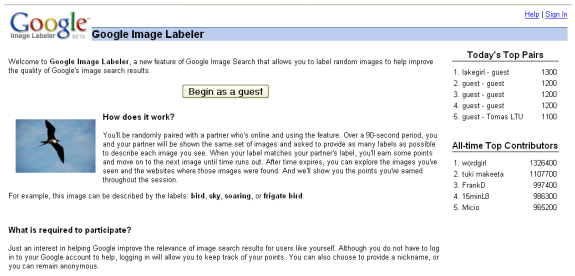
Recuperación de imágenes

- Para buscar imágenes similares se intenta actualmente utilizar inteligencia artificial, para ello se necesita una base de datos de imágenes y descriptores asociados, como experimente el más impresionante es http://images.google.com/imagelabeler/

- Existen buscadores para buscar este tipo de material como Retrievr capaz de buscar a partir de una imagen (lo hace con la transformación wavelet en vez de una red de neuronas). Quizás sea el buscador más prometedor de este dipo

Indización por competición: Imágenes

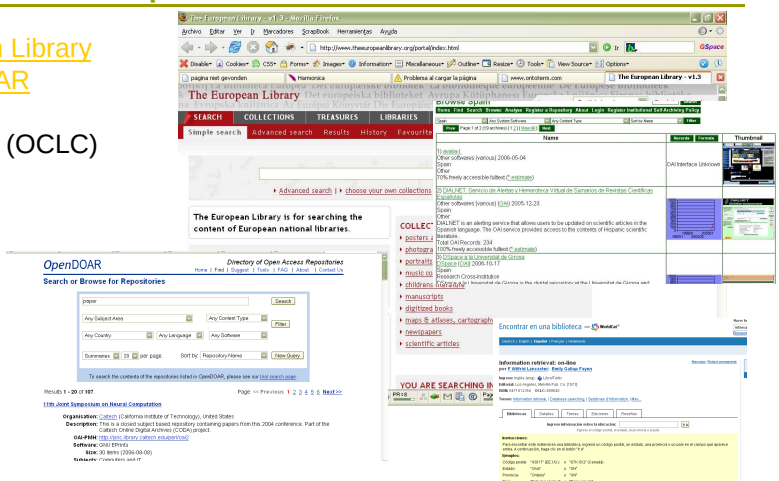
- Flickr
- http://www.espgame.org/
- Google Image Labeler



UNIVERSIDAD CARLOS III DE MADRID

Buscar en Repositorios de Bibliotecas

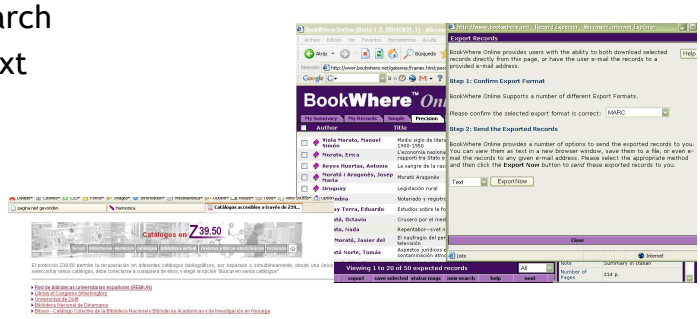
European Library
OpenDOAR
ROAR
WorldCat (OCLC)



UNIVERSIDAD CARLOS III DE MADRID

Búsqueda en Bibliotecas

- Z39.50
  http://www.csic.es/cbic/catalogosz.html
  http://www.loc.gov/z3950/gateway.html
- BookSearch
- Bibliotext
- ProCite



UNIVERSIDAD CARLOS III DE MADRID

Los Buscadores de la Web Social

En su mayoría vuelven a ser directorios:

- Ajaxwhois , para buscar nombres de dominios
- FundooWeb multibuscador, busca a la vez en Yahoo!, Flickr, Yahoo! News, Yahoo! Answers, Amazon, y Yahoo! Maps images.
- Keotag's multibuscador con Web 2.0. Google, Technorati, y Bloglines
- Whonu Multibuscador inteligente
- Similicio.us sitios similares segun delicious
- KwMap similar a kartoo con búsquedas próximas
- Mnemomap categoriza los resultados en "Token", "Tags", "Translations" y "Synonyms".
- PreFound, va contra unos pocos buscadores pero se pueden ajustar las preferencias si se esta registrado.
- Quintura, terminos para expandir la consulta, para obtener mas terminos solo pararse sobre un término
- **Ujiko**. Como un video, buen funcionamiento
- Topix buscador de noticias sobre un grafico cronologico

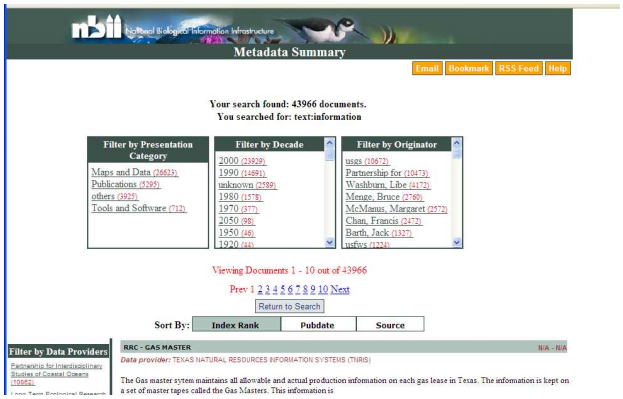
UNIVERSIDAD CARLOS III DE MADRID

Buscadores Web especiales

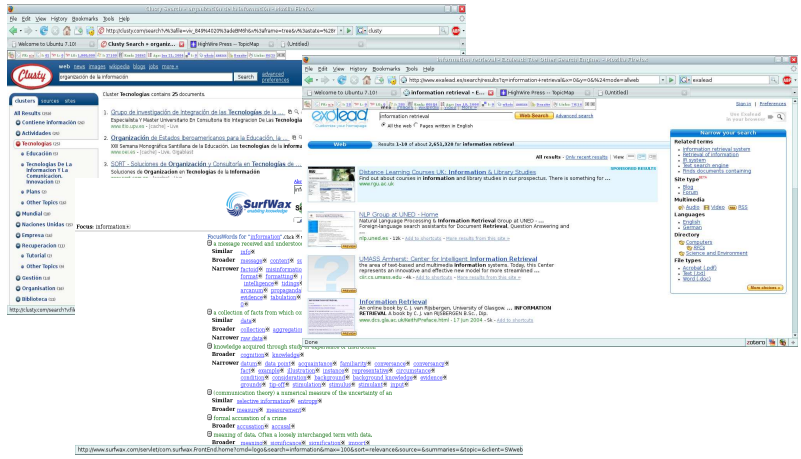
1. Por su temática
2. Por su tecnología y recursos
3. Por los servicios que ofrece

Búsqueda por Metadatos

Cleringhouse http://mercdev3.ornl.gov/nbii/

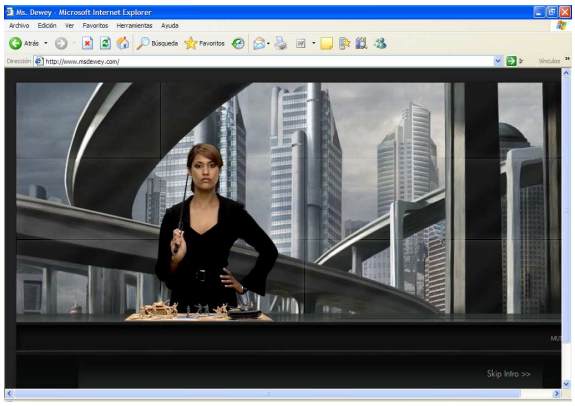


Recuperación con tesauros



Buscador – Sistema pregunta-respuesta

Mr. Dewey http://www.msdewey.com/



PLN y QA

- Procesamiento Lenguaje Natural
  - Natural Finder http://demos.bitext.com/LIVE/
  - Powerset http://www.powerset.com/
  - Hakia http://www.hakia.com
  - Google http://www.google.es
- *Question-Answer*
  - Brainboost→answers http://www.answers.com/bb/
  - Answers.com http://www.answers.com/
  - Ask Jeeves http://es.ask.com/
  - Start http://start.csail.mit.edu/

Asistentes Virtuales

- Se trata de un sistema pregunta-respuesta
- En inglés también se denomina Chatterbot agente conversacional (http://en.wikipedia.org/wiki/Chatterbot), este termino fue acuñado por Michael Mauldin en 1994 en las conferencias de Twelfth Nacional Conference on Artificial Intelligence
- Un chatterbot (o chatbot/ talk bots o chat bots o chatterboxes) es un tipo de agente conversacional, un programa de ordenador diseñado para simular un conversación inteligente con uno o más humanos.
- Se evaluan mediante los "Turing Test" (http://en.wikipedia.org/wiki/Turing_Test)

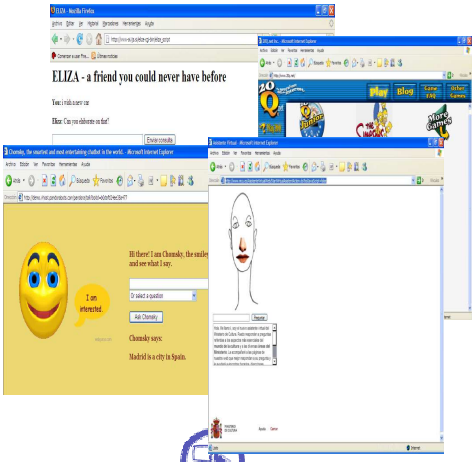
Asistentes Virtuales

- **Asistentes Virtuales Interactivos (*interactive virtual assistant*)** son básicamente personajes virtuales para mejorar la interacción con el usuario/cliente en entornos web, quioscos interactivos, cajeros, móviles, etc.
- En inteligencia artificial los avatares son utilizados por las organizaciones para interactuar con los consumidores. Algunos avatares son conocidos como "***bots***" (*robots*) y han adquirido relevancia por el procesamiento del lenguaje natural.
- Ejemplos: Algunos ejemplos famosos son los de **IKEA (Anna).** Con ellos se mantiene una conversación digital (digital conversation, http://en.wikipedia.org/wiki/Digital_conversation). Este tipo de avatar es conocido como ***Structured Language Processing or SLP Avatar***.

UNIVERSIDAD CARLOS III DE MADRID

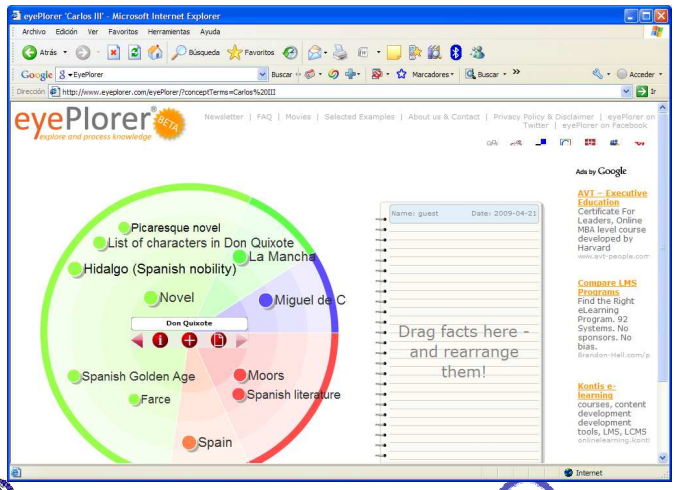
Asistentes Virtuales

- A.L.I.C.E (Artificial Linguistic Internet Computer Entity)
- Jabberwacky
- Ella
- Chomsky
- IKEA



UNIVERSIDAD CARLOS III DE MADRID

Buscadores semánticos



UNIVERSIDAD CARLOS III DE MADRID

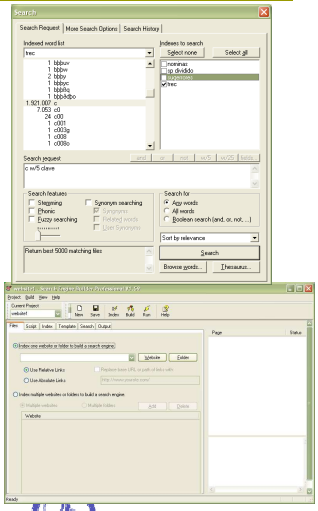
Buscadores Web especiales

1. Por su temática
2. Por su tecnología y recursos
3. Por los servicios que ofrece

UNIVERSIDAD CARLOS III DE MADRID

Buscadores en PC

- **Google Desktop**: rápido y gratuito pero sin facilidades de búsqueda
- **DTSearch**: buscador versatil, pero caro
- **SearchEngineBuilder**, constructor de buscador Web
- **Exalead Desktop**



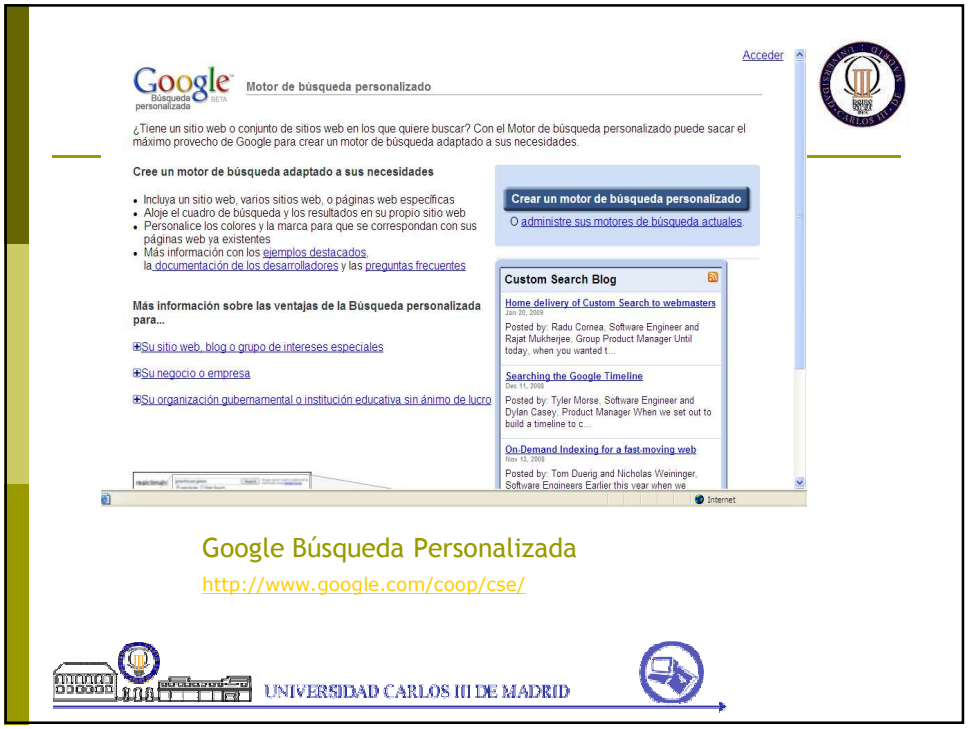
UNIVERSIDAD CARLOS III DE MADRID

Los Buscadores de la Web Social

En su mayoría vuelven a ser directorios:

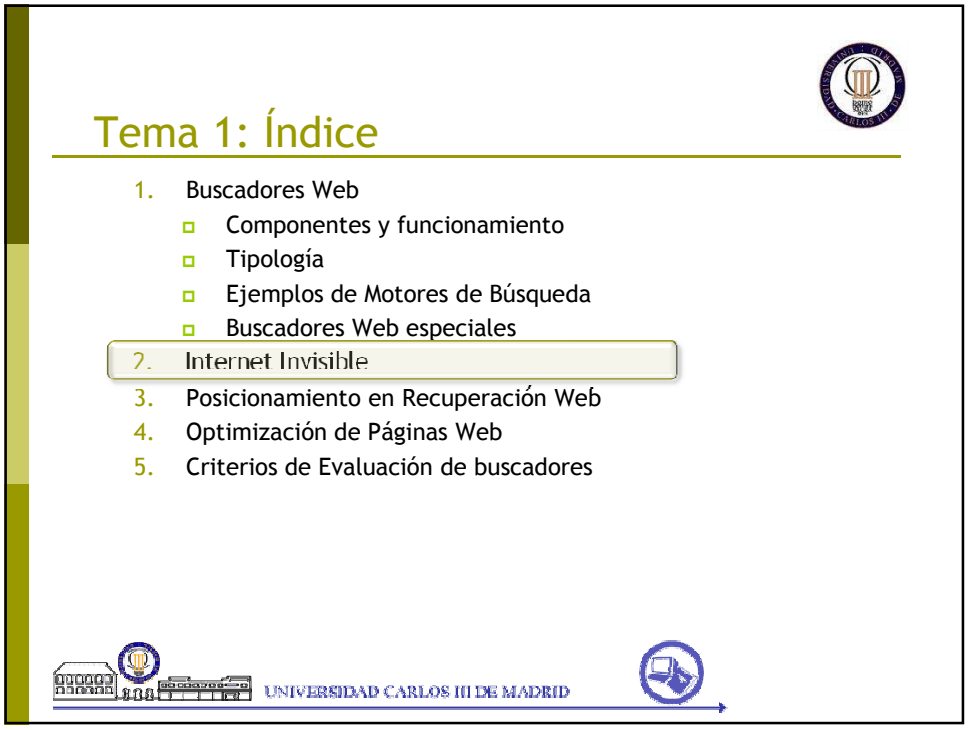
- Ajaxwhois , para buscar nombres de dominios
- FundooWeb multibuscador, busca a la vez en Yahoo!, Flickr, Yahoo! News, Yahoo! Answers, Amazon, y Yahoo! Maps images.
- Keotag's multibuscador con Web 2.0. Google, Technorati, y Bloglines
- Whonu Multibuscador inteligente
- Similicio.us sitios similares segun delicious
- KwMap similar a kartoo con búsquedas próximas
- Mnemomap categoriza los resultados en "Token", "Tags", "Translations" y "Synonyms".
- PreFound, va contra unos pocos buscadores pero se pueden ajustar las preferencias si se esta registrado.
- Quintura, terminos para expandir la consulta, para obtener mas terminos solo pararse sobre un término
- **Ujiko**. Como un video, buen funcionamiento
- Topix buscador de noticias sobre un grafico cronologico

UNIVERSIDAD CARLOS III DE MADRID



Google Búsqueda Personalizada

http://www.google.com/coop/cse/



Tema 1: Índice

1. Buscadores Web
   - Componentes y funcionamiento
   - Tipología
   - Ejemplos de Motores de Búsqueda
   - Buscadores Web especiales
2. Internet Invisible
3. Posicionamiento en Recuperación Web
4. Optimización de Páginas Web
5. Criterios de Evaluación de buscadores

2. 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 65% 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

Buscadores para Internet Invisible

- Direct Search www.freepint.com/gary/direct.htm
- Turbo10 http://turbo10.com
- Profusion http://www.profusion.com
- Internet Invisible http://www.internetinvisible.com http://www.internetinvisible.com/ii/
- Invisible Web http://www.invisible-web.net/
- Complete Planet http://www.completeplanet.com
- Librarian's Index to the Internet http://www.lii.org
- Infomine http://infomine.ucr.edu/
- Look Smart http://search.looksmart.com/
- Web Brain http://www.webbrain.com
- Easy searcher http://www.easysearcher.com

Tema 1: Índice

1. Buscadores Web
   - Componentes y funcionamiento
   - Tipología
   - Ejemplos de Motores de Búsqueda
   - Buscadores Web especiales
2. Internet Invisible
3. Posicionamiento en Recuperación Web
4. Optimización de Páginas Web
5. Criterios de Evaluación de buscadores

Posicionamiento

El posicionamiento es un conjunto de **criterios** que se aplican para construir un **algoritmo** para que un motor ordene por **relevancia las páginas** correspondientes al resultado de una consulta.

Características

- Secretos comerciales
- Cada buscador tiene algoritmos de posicionamiento diferentes
- Los algoritmos pueden ser más o menos complejos

Objetivos

- Seleccionar documentos relevantes como respuesta a una consulta
- Ordenar los resultados por orden de relevancia
- Encontrar los documentos que contengan las palabras clave introducidas por el usuario.

Posicionamiento: Factores que influyen en algoritmo

Factores directos

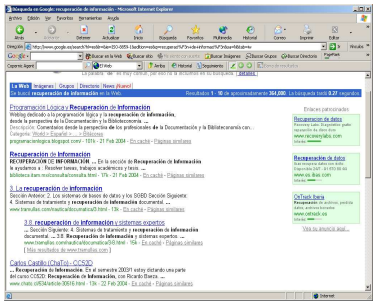
- Tipología de la búsqueda
- Popularidad de la página
- Formato (fuente, extensión)
- Perfiles de usuario
- Factor de interés económico

Factores indirectos

- Contenidos y estructura
- Credibilidad/fiabilidad
- Usabilidad
- Accesibilidad

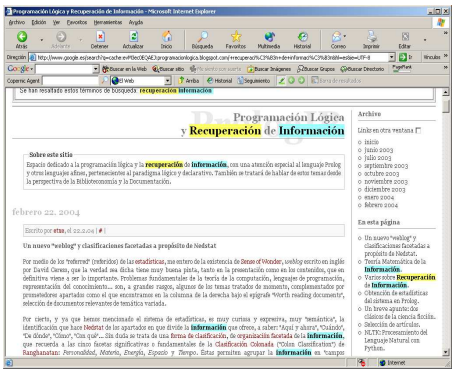
Factores directos: posicionamiento por similitud de búsqueda (i)

- Frases, presencia de frases de búsqueda en frases del documento (Infoseek)
- Palabras clave al principio y final de la página (Altavista, Inktomi)
- IDF – *Inverse Document Frequecy* (Altavista)
- Búsqueda booleana con OR, gana el que más palabras diferentes tiene (Altavista)
- Por *clusters* de conceptos asociados Excite, Intelligence Concept Extraction (tb en labgoogle.com/sets, clusty, vivisimo, exalead, kelkoo)
- Tamaño de la página (mejor si son pequeñas) (Google)



Factores directos: posicionamiento por similitud de búsqueda (ii)

- URL con el término de búsqueda (Altavista, Google, Hotbot, Infoseek, Excite y Lycos)
- Posición de los términos de búsqueda al principio (Altavista, Excite, Webcrawler)
- Proximidad de los términos de búsqueda en el texto (Altavista, Lycos, Webcrawler). Frecuencia del término de búsqueda en el texto (Altavista, Lycos)
- Título, presencia del término en búsqueda en el título (Lycos, Webcrawler, Google)



UNIVERSIDAD CARLOS III DE MADRID

Factores indirectos

Usabilidad. Ejemplos que bajan la popularidad:

- Poner una página de bienvenida
- Todo lo que tenga muchos colores, no sea claro en su unidad, o se parezca a un anuncio
- No poner las conclusiones al principio ni escribir de forma esquemática

Accesibilidad. Ejemplos:

- Animaciones flash (es una imagen y no la leen los motores), páginas dinámicas.
- Todo lo que afecte al tiempo de descarga

Otros: Contenido, Arquitectura hipertextual y Credibilidad

UNIVERSIDAD CARLOS III DE MADRID

Penalizaciones (i)

- **Por limitaciones del motor o limitaciones en su acceso**
  - Tamaño (muchas penalizan el tamaño, p.e. Google, sólo los primeros 100k)
  - Porcentaje alto de código respecto al texto, y todo lo que influye en web invisible (formularios, passwords, bases de datos, ficheros no procesables, ...)
- **Por limitaciones en la comunicación con el usuario**
  - Elementos que disminuyan la credibilidad, la usabilidad o accesibilidad

Penalizaciones (ii)

Por engaños al motor y protección contra herramientas SEO (Search Engine Optimization)

- Poner texto y fondo del mismo color o etiquetas o meta con contenidos engañosos (a Altavista se le engaño durante mucho tiempo. Google es la única que actualmente no lo penaliza)
- Creación de páginas falsas con la palabra clave repetida miles de veces y con enlaces a la página a promocionar (Google incluso no permite repetirla en la misma línea. También Infoseek, Stuffing)
- Poner texto en tamaño muy pequeño (Inktomi)
- Redireccionamientos automáticos (Altavista, Infoseek, Lycos, Excite)
- Mandar varias veces la misma página al motor, por ejemplo en la misma semana. (Evitar que nos lo envíe un software automático)
- Utilizar Cloaking (páginas falsas enmascaradas) y Doorway (páginas especiales para un búscador e ininteligibles para los usuarios)
- SandBox (cuarentenas) (motivo: migración BD o protección SEO para nuevos dominios)
- Revisad http://www.google.com/webmasters/seo.html

Tema 1: Índice

1. Buscadores Web
   - Componentes y funcionamiento
   - Tipología
   - Ejemplos de Motores de Búsqueda
   - Buscadores Web especiales
2. Internet Invisible
3. Posicionamiento en Recuperación Web
4. Optimización de Páginas Web
5. Criterios de Evaluación de buscadores

2. Optimización

- Directrices de diseño
  - Elementos a tener en cuenta (título, palabras clave, metadatos, descripción, texto)
  - Recomendaciones
- Optimización de páginas Web para un motor
  - Google: Descripción de elementos
  - Yahoo!Search
- Herramientas SEO (*Search Engine Optimization*)
  - Monitorización de páginas Web
  - Simuladores de motores (palabras clave y enlaces)
  - Soluciones integradas
- Factores indirectos
  - Accesibilidad
  - Usabilidad
  - Calidad de Uso
  - Credibilidad
  - Interoperabilidad

Directrices de diseño: Título de la página

- Las palabras clave se encuentran en las palabras del título y de la URL. Da muy buenos resultados, siendo lo más específicos posible.
- Poner una etiqueta META con TITLE (Con DC. Title mejor).
- Título sencillo acorde con el tema, palabras clave y con pocas palabras (entorno a 5).
- Algún buscador ordena alfabéticamente, que en ASCII es: "#$%&'()*+,-./012...89:;<=> ?@A...XYZ[\]^_`ab c...yz {|}~
- Dar de alta la página, con el mismo título. No es recomendable dar de alta la página con distintos títulos pero con la misma URL.

Directrices de diseño: Palabras Clave

- Pensad en el tema principal, pero también en como lo buscarían los usuarios para una necesidad de recuperación
- No poner en la misma línea la misma palabra clave
- Para seleccionar las palabras, se puede indizar el documento, quitando palabras vacías y realimentar con el estudio de los enlaces que apuntan a mi página

En meta keywords

- Unas 25 palabras o frases (no repetir en las etiquetas meta la misma más de tres veces para no ser penalizados)
- Evitar palabras muy comunes o muy raras o difíciles de escribir
- Palabras con/sin acento y en plural/singular
- Incluir errores comunes en la ortografía de las palabras (depende del idioma del público objetivo, p.e. *busines, bussines, ingeneering* ... son críticas en España)

Directrices de diseño: Descripción

- Sencilla explicación del contenido temático de su página Web
- Si el buscador exige que tenga un número determinado de palabras, se recomienda no superarlos, puede aparecer cortadas y sin ningún sentido
- Parece que en las meta, Google da mejores resultados si se usa el ***dc. description*** en vez de *description* a secas.

Directrices de diseño: Enlaces

Texto del enlace

- Enlazar sitios útiles con información y escribir términos que describan el contenido claramente y con exactitud
- Determinar las palabras de búsqueda que los usuarios escribirían para encontrar las páginas
- Ejemplos de malos textos para enlaces: "mas sobre mi", "mis enlaces", "mis amigos" "sobre mi"

- Sitios con jerarquía
- Si existen más de 100 enlaces dividir la página. Google podría marcarla como granja de enlaces y penalizarla
- Poner etiquetas ALT en imágenes y enlaces, el texto debe ser descriptivo, si no puede ser penalizado. (mejora la accesibilidad)
- Comprobar que no existen vínculos rotos o código HTML incorrecto
- Vínculos con texto claro (**Bombing**)
- En Google, poner URL absolutas. Se debe poder acceder a todas las páginas desde al menos un vínculo estático

Directrices de diseño: Recomendaciones

- **Abusos con Palabras clave**
  - Los robots las detectan y eliminan el sitio
  - No escribir más de 3 veces la misma palabra clave
  - Mantener separadas las frases y palabras que contengan palabras repetidas
- **Metaetiquetas**
  - Seleccionar diferentes palabras clave para cada página
  - Variaciones morfológicas de una palabras son palabras diferentes (p.e. casa, casas, casa de fin de semana)

Optimización de páginas Web: Google

Consideraciones

- El último gran cambio en el algoritmo es de octubre de 2004. Actualmente se cree que Google podría estar utilizando hasta tres algoritmos diferentes
- **PageRank**, cálculo basado en el número de páginas que apuntan a una página. Un enlace de una página vale más si está a su vez recibe muchos enlaces. Tiene en cuenta 100 factores, pero los enlaces son el principal dada la estabilidad de resultados a lo largo del tiempo.

Ejemplos de factores que valora Google

- Formato del texto
- Densidad de la palabra de consulta en título y body
- Presencia de la consulta de URL (tanto en el body como en URL, ignora los guiones bajos e interpreta los guiones normales como separadores)
- Texto de los enlaces que apuntan a la página
- Número de enlaces que recibe la página en estudio (PageRank)

Optimización de páginas Web: Google

PageRank de Google

- Estimación PageRank por vecindad. Profundidad en el directorio, si el sitio principal tiene un PR de 6 y la página está en una subcarpeta "nieta", se resta dos al PR

El PageRank tiene en cuenta:

- Número de páginas que enlazan a la página
- Los enlaces puntúan más si vienen de páginas que PR es alto o con pocos enlaces salientes (p.e. Estar dado de alta en DMOZ –directorio que incorpora Google- da de media un punto más de PR si el tenía un valor inferior a seis). Algo similar ocurre en Yahoo!

PR(A)=(1-d)+d(PR(t1)/C(t1)+PR(t2)/C(t2)....)

PR(A), es el PageRank de la página A
d, factor de amortiguación, probabilidad de que el usuario abandone la página. Por defecto 0.85
PR (t1)...t1, t2 son los Pr de las páginas que apuntan a A
C(t1), enlaces que salen de la página t1, para evitar división por cero, la página se considera autolenlace

Explicación PR:
http://www.webworkshop.net/pagerank.html
Calculadora de PR:
http://www.webworkshop.net/pagerank_calculator.php
Saber tu PR:
http://www.googlemania.com/pagerank.php

Optimización de páginas Web: Google

Directrices de Calidad -Principios básicos-

- Tiene en cuenta el texto del enlace con la consulta (*Bombing*)
- Tiene en cuenta enlaces internos, pero si sólo existen enlaces internos y no externos lo puede penalizar
- Google evita repeticiones en su lista de resultados
- Crear páginas para usuarios y no para motores de búsqueda.
- No tratar de engañar a los usuarios presentando a los motores de búsqueda contenido distinto al que se desea mostrar
- Penalizaciones: programas de comprobación de rankings, programas para remitir páginas, intercambio de links ...

Optimización de páginas Web: Recomendaciones

- Una página con *frames* depende de links externos y del título para posicionarse. Añadir información mediante la etiqueta **<no frame>**, poner títulos descriptivos y etiquetas meta.
- En las páginas flash, utilizar la etiqueta **<noembed>** para poner texto. Poner un buen título
- Javascript, usar la etiqueta **<noscript>** y poner buen título

Optimización de páginas Web: Recomendaciones para Yahoo!

- Incluirse en DMOZ y directorio Yahoo!. DMOZ (Open Project) está realizado a mano
- El texto de los enlaces que van en una página no son tan relevantes para esa página como en Google. Poner pocas palabras en el texto de los enlaces
- No tiene en cuenta palabras vacías
- Yahoo! Valora el texto no los enlaces que apuntan y la estructura del Web. Es muy permisivo con prácticas indeseables. Spam con keyworks abusivas (span interno) y el spam externo funciona bien, pues le da igual de quien recibe los enlaces
- Mucho peso al título, permitiendo repeticiones de palabras, se puede llegar a 100 caracteres
- Las keywords en la URL parecen tener más peso
- Utiliza etiquetas meta, pero las da poco peso
- Tiene en cuenta enlaces entrantes
- Tarda mucho en subir una página y penaliza sin sentido

Herramientas SEO (Search Engine Optimization)

Herramientas de análisis:

- Comprobar palabras clave y título con las mejores páginas posicionadas
- Estudio de enlaces
- Ver con los ojos de un motor
- Monitorizar nuestra página con la competencia

Herramientas de mejora:

- Elección del título, elección de palabras clave
- Políticas de intercambio de enlaces

Simuladores de motores

- http://www.1-hit.com/all-in-one/tool.search-engine-viewer.htm
- http://www.delorie.com/web/ses.cgi
- *Search Engine Spider Simulator* http://www.webconfs.com/search-engine-spider-simulator.php
- http://www.searchenginewold.com/cgi-bin/sim_spider.cgi

Cálculo de densidad de palabras

- Keyword Counter-Keyword Frequency Analyzer http://www.keywordcount.com/
- http://www.forobuscadores.com/recursos/index.php?m=c&c=27

Herramientas SEO: Monitorización

ALEXA http://www.alexa.com

- Ranking basado en popularidad, mide el tráfico (con usuarios con la barra de Alexa) media geométrica de
  - Reach (alcance): media del número de usuarios (visitas) que recibe un site (no URL) por millón y día
  - Page View: número de páginas vistas por diferentes URL que han solicitado determinada página [diferentes=mismo día]

Límites de ALEXA:

- sólo funciona con Explorer y Windows,
- sólo tiene en cuenta si hay más de mil visitas,
- no accede a https,
- solo mide la URI principal, no las páginas secundarias de un site.

Google Monitor
http://www.googlemania.com/monitor.php

Google Tracking mira los enlaces
http://www.googlemania.com/tracking.php

Soluciones Integradas

IBP

SEO Toolbox
http://www.webrankinginfo.com/english/tools/seo-toolbox.php

Los factores indirectos

- Accesibilidad
- Usabilidad
- Credibilidad
- Contenido
- Interoperabilidad

Todos estos factores tienen un equivalente en creación de aplicaciones

Los factores indirectos: Accesibilidad

No poner barreras al público objetivo (invidentes, daltónicos, "torpes" desmemoriados, propietarios de un hardware obsoleto), niños, conexiones lentas, países en desarrollo

La accesibilidad se mide con test con varios nivesl A, AA, AAA

- http:// www.w3c.org/TR/WCAG10/full-checklist.html
- Software: Bobby, TAW
  - http://www.tawdis.net
  - http://www.cynthiasays.com

Corrección HTML y CSS:

- http://validator.w3c.org Valid XHTML 1.0
- http://jigsaw.w3c.org/css-validator Valid CSS

Los factores indirectos: Usabilidad

Facilidad de uso y de aprendizaje.

Mejora si se elimina lo superfluo (no significa falta de estética)

Evalúa:

- Facilidad de aprendizaje
- Facilidad de recuerdo
- Eficiencia del usuario
- Tasa de errores
- Satisfacción

Son interdependientes, bajar una puede suponer aumentar otra

Los factores indirectos: Credibilidad

Estudia como se relaciona la credibilidad y el nivel de compromiso, enlaces recibidos y ofrecidos y diseño del Web

Depende de: mención de autoría, institución, política de confidencialidad, fecha de actualización

Estudios de usuarios revelan que también está en función de la **usabilidad y la calidad técnica del sitio**

Credibilidad positiva:

- El sitio te ha servido en el pasado
- Prestigio de la organización que avala el sitio
- El sitio responde rápido al cliente
- Se facilita la dirección física y el teléfono
- Se ha actualizado recientemente
- El sitio parece diseñado por profesionales
- La usabilidad es correcta (contenido, comprensión, arquitectura)

Credibilidad Negativa:

- Dificultad para distinguir anuncios de texto
- Falta de actualización
- Tiene pop-ups con anuncios
- Falta de usabilidad
- Enlaces erróneos o desacertados
- Enlaces o recursos no accesibles

Los factores indirectos: Credibilidad

Credibilidad positiva (factores de importancia)

- 1.5 dar la dirección e-mail
- 1.5 organización racional
- 1.3 listar autores, existencia de citas claras y referencias
- 1.3 sitio enlazado creíble
- 1.2 se muestra la política de privacidad
- 1.2 el sitio manda correos de confirmación tras las transacciones
- 1.2 existe un buscador interno
- 1.0 permite imprimir fácilmente
- 1.0 dar la información en más de un idioma

Credibilidad negativa (factores de importancia)

- -1.4 dificultad para navegar
- -1.4 enlaces no creíbles o erróneos
- -1.3 errores tipográficos
- -1.3 el sitio no está siempre accesible
- -1.1 no coincide nombre de empresa y nombre de URL
- -1.0 se tarda mucho tiempo en descargar la página

Los factores indirectos: Contenido

- **Redacción de texto breve** (50% menos texto que en otros medios). Evitar palabras no específicas del contenido (p.e. bienvenido), texto subjetivo, énfasis, metáforas, sin juegos de palabras, sin palabras ingeniosas ni símiles. No poner URL, sino texto para navegar a otra página mediante clic
- **Estructurado**: párrafos cortos, una idea un párrafo, con varios niveles de encabezados, viñetas anidadas, páginas cortas y vinculadas, subtítulos informativos, tipografía, colores en palabras clave (sin exagerar)
- Se citan menos páginas muchos temas distintos, hacer contenidos específicos
- No dividir un mismo texto en páginas con "continuación"

Legibilidad

- Selección de colores con contraste. Evitar fondo, mejor los fondos claros
- No usar marquesinas ni realizar cambios de tamaño. El cerebro lo asocia a publicidad (banners)
- Alinear a la izquierda
- Sans-serif (verdana) es la más legible a tamaño 10
- No usar tablas anidadas

Hoja de Estilo

- El formato es recomendable en hojas de estilo
- Dan coherencia, mejoran aprendizaje y memoria del sitio
- No usar más de dos fuentes
- La página debe ser legible si desaparece la hoja de estilo
- Tamaño de letra en tamaño relativo (%) a la que tiene por defecto el usuario

Los factores indirectos: Página de inicio y navegación

- Hacer visible logo y nombre de la compañía (superior izda)
- Ancho variable. Si se exige tamaño fijo poner 600*640

Navegación

- Accesible desde el resto de las páginas
- Navegación rápida y fácil. Poner directorio temático, interfaz de búsqueda y enlaces a noticias especiales
- Se debe saber dónde se está, dónde he estado y donde ir en relación al Web y al sitio
- Evitar navegación lineal
- Interfaz de navegación lo más estándar posible
- Evitar nuevas ventanas, impide ir hacia atrás y agobia al usuario

Factores indirectos: Interoperabilidad

- Metadatos
- Ontologías
  - RSS
  - RDF
- Mapeados (Alineaciones)

Tema 1: Índice

1. Buscadores Web
   - Componentes y funcionamiento
   - Tipología
   - Ejemplos de Motores de Búsqueda
   - Buscadores Web especiales
2. Internet Invisible
3. Posicionamiento en Recuperación Web
4. Optimización de Páginas Web
5. Criterios de Evaluación de buscadores

4. Criterios de evaluación de motores

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

- Frecuencia de reindización: Google (de 1-68 días), Hotbot, MSN y Alltheweb tardan poco, las más antiguas en Alltheweb todos pasa cada mes. MSN y hotbot los mejores. Altavista tardaba entre 12-51 días y WiseNut entre 247-286 días [abril 2002]. Google tiene el Google Dance cada 15 días y los mirrors nacionales cada mes.
- **Tamaño**: la mayor son Google 9500 millones (101 k por pg), MSN 5000 billones (150 kpag) y Yahoo!Search 19200 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% (9% 2002) mientras que Google tenía un 3%→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 el mundo google 46% yahoo!Search 24% MSN 11% (2005)
- **Tiempo** que se permanece en Google 29 m AOL 28 Netscape 13

4. Criterios de evaluación de motores

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

- Frecuencia de reindización
- **Tamaño:**
- **Enlaces muertos:**
- **Cobertura:**
- Los más **usados** en el mundo google 46% yahoo!Search 24% MSN 11% (2005)

Tema 1: Índice

1. Buscadores Web
   - Componentes y funcionamiento
   - Tipología
   - Ejemplos de Motores de Búsqueda
   - Buscadores Web especiales
2. Internet Invisible
3. Posicionamiento en Recuperación Web
4. Optimización de Páginas Web
5. Criterios de Evaluación de buscadores
6. Medidas de Recuperación de Información

Evaluación de un Sistema de Recuperación

CONTENIDO
- Cobertura
- ***Tamaño***
- Novedad
- Actualización

DISEÑO
- Interfaz de búsqueda
- Arquitectura:

Estrc.índices (árboles, hash, ...)
Tipo almacenamiento datos, etc
Eficacia almacenamiento
(Indices+reg.doc)/espac.docs
- ***Eficacia de ejecución***

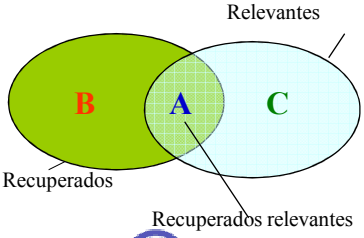
Tiempo en hacer una operación
- Visualización resultados
- Política de Indización

RECUPERACIÓN
- **Algoritmo Recuperación**
- **Algoritmo Posicionamiento**
  - Recall
  - Precisión

Ruido y Silencio

	Relevante	No Relevante
Recuperado	**A**	**B**
No Recuperado	**C**	*D*

- Ruido: Documentos no relevantes recuperados (B)
- Silencio: Documentos relevantes no recuperados (C)

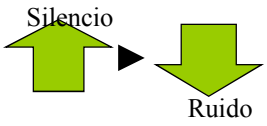


Relación Ruido/Silencio y Estrategias de búsqueda

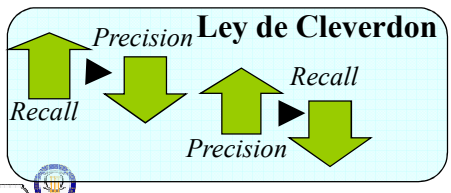
- Disminuir **Ruido**
  - **Consulta**
    - Utilizar términos específicos, añadir términos asociados
    - Operadores AND y NOT
    - Búsqueda por frases, campos, paréntesis, evitar términos polisémicos, usar términos poco frecuentes
  - **Medio**
    - Utilizar Directorios
- Disminuir **Silencio**
  - **Consulta:**
    - Emplear OR, variantes ortográficas (incluido acentos, mayúsculas, género, número, ..), idiomáticas y dialectales
    - Expansión de búsqueda: Términos genéricos y sinónimos
  - **Medio**
    - Metabuscadores y Motores

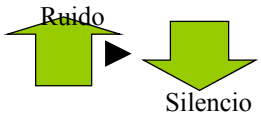
UNIVERSIDAD CARLOS III DE MADRID

Relación Ruido/Silencio



	Relevan.	No Relev.
Rec.	A	B
No Rec.	C	D





Recall =Exhaustividad= A/(A+C)

Mide como evita el sistema el silencio

Entre 0 y 1, mejor si próximo a 1

Precision= A/(A+B)

Mide como evitar el ruido

Entre 0 y 1, mejor si próximo a 1

UNIVERSIDAD CARLOS III DE MADRID

Ejercicio 1

Dos buscadores con misma consulta y misma BD

Buscador 1 { r, r, r, r, r, r }

Buscador 2 { r, nr, r, r, nr, r, r,nr, r, nr, r, r }

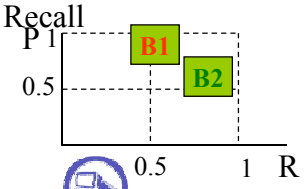
Donde nr es un documento no relevante y r es relevante

La base de datos tiene 10.000 documentos, 10 son relevantes a la consulta estudiada

Indicad que buscador evita mejor el ruido y el silencio según las tasas de Precision Recall

Pb1=6/6=1
Pb2=8/12=0.6

Rb1=6/10=0.6
Rb2=8/10=0.8



Ejercicio 2

Suponga los siguientes resultados de dos buscadores en Internet ante la misma consulta y la misma base de datos

Buscador 1 { 1, 2, 3, nr, 18, 12, nr, 4, 5, nr }

Buscador 2 { 1, 3, 2, 18, 9, 29, 6,nr, nr, nr }

Donde

- nr es un documento no relevante
- Los números son el orden de relevancia del documento
- El orden es en el que han ido apareciendo los documentos

Calcular las tasas de Precision/Recall

Solución Ejercicio 2

	Precision	Recall
Buscador1	7/10	7/x
Buscador2	7/10	7/x

¿son entonces iguales los dos buscadores?

Precision Recall- Problemas

- Una sola medida de precision recall mide la calidad del **algoritmo** de recuperación no del algoritmo de **posicionamiento**
  (el posicionamiento solo tiene sentido cuando el modelo de recuperación lo permite)
- En **Internet** es imposible saber **cuantos documentos relevantes** existen a una pregunta dada
- No se tiene en cuenta el ajuste a la medida **manual de la relevancia**
- No se tiene en cuenta la **interacción con el usuario**
- Son **dos medidas** de una misma cuestión, hay que decidir a cual se la quiere dar **preferencia**

Precision-Recall unificada

- Medida de la F
  - Unifica Precision-recall en una única medida utilizando la media armónica, cuanto más próximo a uno mejor (a cero peor). Se mide en el j documento recuperado.

    $F(j)=2/((1/r(j)+1/P(j))$
- Medida de Evaluación
  - Como la armónica pero configurable, si b>1 más peso a la precision, si b<1 a la recall

    $F(j)=1+b^2/((b^2 /r(j)+1/P(j))$

Otras medidas:

- Índice de irrelevancia

  Nº documentos no relevantes recuperados /

  nº documentos no relevantes en la colección
  - Da información aun cuando no hay documentos relevantes (¡para Recall division por cero!) o cuando no recupera documentos relevantes. Tiene en cuenta D el número de documentos irrelevantes recuperados. Cuanto más pequeña mejor
- Recall de documentos relevantes únicos (URR)
  - Sirve para comparar dos buscadores se tienen en cuenta sólo los relevantes no duplicados en los resultados de los dos buscadores
  - Nº de relevantes únicos/número total de relevantes

Gráficos de Precison Recall

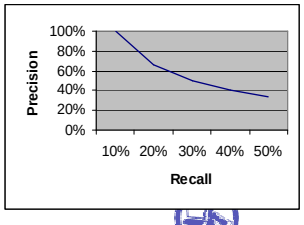
- Es el sistema **más utilizado** en la literatura para mostrar el funcionamiento de un motor o varios
- Sirve para **mostrar gráficamente**, de forma sencilla, la eficacia y eficiencia de un sistema de recuperación
- Se mide la Precision a **11 niveles de Recall**:
  0%, 10%, 20%, ...70%, 80%, 90%, 100%
- Si no se posee determinado valor de Precision se **interpola** con la Precision correspondiente al siguiente Recall conocido (incluido el caso del 0% de Recall)
- Opcionalmente se puede ver la **Precision en valores fijos**. P.e. Cuando se han recuperado 10, 20, 30... **documentos relevantes**

UNIVERSIDAD CARLOS III DE MADRID

Gráfico Precision Recall

Relevantes 10 / **Recuperados** 15																
										Buscador						
	r	nr	r	nr	nr	r	nr	nr	nr	r	nr	nr	nr	nr	r	
Documentos Rec	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Relev Rec	1	1	2	2	2	3	3	3	3	4	4	4	4	4	5	
			2/10						2/3							
Recall	10%	10%	20%	20%	20%	30%	30%	30%	30%	40%	40%	40%	40%	40%	50%	
Precision	100%	50%	67%	50%	40%	50%	43%	38%	33%	40%	36%	33%	31%	29%	33%	

Gráfico	10%	20%	30%	40%	50%
	100%	67%	50%	40%	33%



UNIVERSIDAD CARLOS III DE MADRID

Gráfico Precision Recall. Interpolación

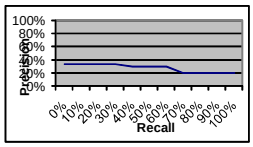
Relev	3
Recs	15

	nr	nr	r	nr	nr	nr	r	nr	nr	nr	nr	nr	nr	nr	r
Doc Rec	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Rel Rec	0	0	1	1	1	1	2	2	2	2	2	2	2	2	3

Recall	0%	0%	33%	33%	33%	33%	67%	67%	67%	67%	67%	67%	67%	67%	100%
Precision	0%	0%	33%	25%	20%	17%	29%	25%	22%	20%	18%	17%	15%	14%	20%

Gráfico

33%	67%	100%
33%	29%	20%

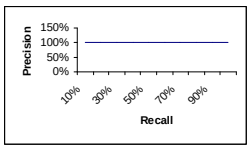


Interpolación

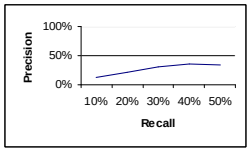
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
33%	33%	33%	33%	29%	29%	29%	20%	20%	20%	20%

UNIVERSIDAD CARLOS III DE MADRID

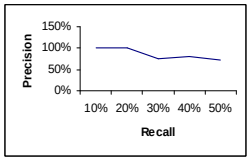
Gráficos de precisión recall



Recuperación idónea
Cada documento recuperado es relevante



Recuperación tardía
Los primeros docs no son relevantes
pero los últimos si



Recuperación temprana
Los primeros docs son relevantes
pero los últimos no

UNIVERSIDAD CARLOS III DE MADRID

Consultas agrupadas

- Los gráficos de precision recall no suelen contener una sola consulta, sino que **agrupan varias consultas**
- El método es calcular la **precision media** a cada uno de los 11 niveles de recall

Estimación Recuperación en Internet

- Problema:
  - Se desconoce el total de relevantes (Recall)
  - Difícil conocer el total de relevantes recuperados si la búsqueda tiene muchos docs
  - Dificultades añadidas por documentos no indizados por el motor y documentos no recuperados pero indizados por el motor
  - Para poder comparar motores en Internet deberíamos de poder utilizar la BD de un motor (p.e. Google) con los algoritmos de recuperación y posicionamiento de otro motor (p.e. Altavista)

Estimación Recuperación en Internet: Soluciones

- No calcular la Recall
- Limitarse a los n primeros resultados recuperados (20)
- Utilizar palabras de muy baja presencia para así poder evaluar todos los documentos
- Para Comparar motores: A veces se normaliza el número total de relevantes sumando los documentos relevantes de los 20 primeros resultados de varios motores
- Identificar documentos que deberían de estar (p.e.por estar en una revista electrónica o un dominio relevante), ver cuantos recupera
- Poner artículos relevantes en el motor y ver cuantos se recuperan
- Si se puede acceder a subcolecciones como newsgroups hacer muestreos de relevantes

Estimación Recuperación en Internet

- Algunos autores (Chignell) proponen modificar la medida de Precision de los 20 primeros resultados añadiendo información sobre el grado de Relevancia

  $P=\Sigma puntuación/20*4$

  La puntuación se asigna manualmente de 1 (mínimo) a 4 (máximo)

Consultas sin Agrupar

- Desventajas de Agrupar
  - No se puede saber como se comporta un tipo específico de consultas
  - No permite comparar dos algoritmos frente a consultas individuales

Tipos:
- Media de Precision en n valores de recuperación
- R-Precision
- Histogramas de Precision

Consultas sin agrupar

- Media precision: favorece los algoritmos que dan antes los docs relevantes

Relevantes	10
Recuperados	15

R-Precision = 40%
Valor de la precisión al recuperar el mismo nºde docs q el nº de documentos relevantes

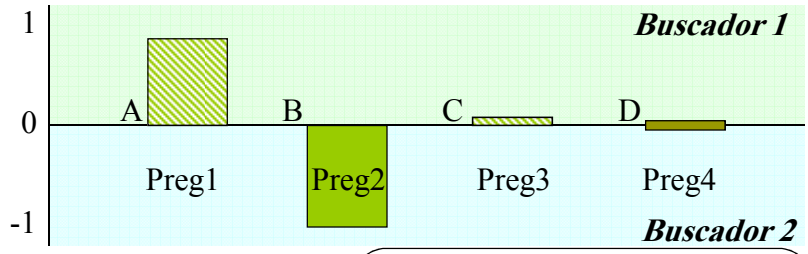
Documentos Recuperados	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Relevantes Recuperados	1	1	2	2	2	3	3	3	3	4	4	4	4	4	5
Recall	10%	10%	20%	20%	20%	30%	30%	30%	30%	40%	40%	40%	40%	40%	50%
Precision	100%	50%	67%	50%	40%	50%	43%	38%	33%	40%	36%	33%	31%	29%	33%

Precisión media a n documentos relevantes

10%	20%	30%	40%	50%
100%	67%	50%	40%	33%

=suma porcentajes dividido número de relevantes recuperados **58%**

Consultas sin agrupar: Histogramas de Precision



- Se representa R-precision de cada consulta en 2 buscadores distintos
- Se resta el valor de la R-precision en el buscador 1 al de la R-precision en el buscador 2

A-Buscador 1 mejor que el 2 en la primera pregunta

B-Buscador 2 mejor en la segunda pregunta

C-Buscador 1 un poco mejor

D-Los dos buscadores son idénticos

UNIVERSIDAD CARLOS III DE MADRID

Medidas orientadas al usuario

Para un usuario concreto	Conocidos	Desconocidos
Relevantes Recuperados	A	B
Relevantes (presentes o no en la BD)	C	D

- Cobertura=A/C
  - De los relevantes conocidos por el usuario cuantos se han recuperado
- Novedad=B/(A+B)
  - De los relevantes recuperados cuantos le eran desconocidos

UNIVERSIDAD CARLOS III DE MADRID

Medidas Centradas en el Usuario

- Recall Relativa

$$\frac{\text{Documentos relevantes recuperados}}{\text{Documentos relevantes esperados}}$$

- Esfuerzo en la Recuperación

$$\frac{\text{Documentos relevantes esperados}}{\text{Documentos relevantes examinados}}$$

Colecciones de Prueba: Test collections

- Las tasas de Precision Recall son solo ciertas para determinada colección y determinadas preguntas, no es extrapolable
- Colecciones predefinidas de documentos, preguntas y juicios de relevancia (ajuste de cada documento a cada pregunta)→Benchmarking
- Sirven para mejorar los algoritmos de recuperación y posicionamiento
- Tendencia a ajustarse a la realidad. En sus inicios eran documentos breves y las preguntas no eran las típicas de los usuarios
- En un principio con etiquetas propias, actualmente con DTDs de XML
- Existen competiciones en que varios motores muestran sus prestaciones:
  - TREC (Recuperación), Message Understanding Conferences (MUC), Document Understanding Conferences (DUC), Cross-Language Evaluation Forum (CLEF), Summarization evaluation effort (SUMMAC), SENSEVAL (Semántica), CLEF (Multilingüe)
  - Colecciones clásicas: ftp://ftp.cs.cornell.edu/pub/smart

Colecciones clásicas (Smart)

COLECCIÓN		DOCS	terms	PREG	termsTAMAÑO	
CACM	Informatica	3,204	10,446	64	11,4	1.5
CISI	Biblio.	1,460	7,392	112	8,1	1.3
CRAN	Aeronau.	1,400	258,771	225	4043	1.6
MED	Medicina	1,033		30		1.1
TIME	Articulos	425		83		1.5

Cranfield

- Ejemplo documento
  - **.I 250**
  - **.T** pressure distributions at zero lift for delta wings with rhombic cross sections .
  - **.A** eminton,e.
  - **.B** arc cp.525, 1960.
  - **.W** pressure distributions at zero lift for delta wings with rhombic cross sections ... calculation and some of the results are compared with those of slender thin wing theory .
- Ejemplo pregunta
  - **.I 029**
  - **.W** material properties of photoelastic materials .

Cranfield

- Evaluación

Pregunta ID	Documento ID	Grado Relevancia
29	225	3
29	250	2
29	464	4
29	513	-1

Campos en las colecciones clásicas

- Título, Autor, Fuente (casi todas)
- Resumen (Cranfield, CISI, Time, Medline)
- Fecha (Time,CACM)
- Raíces de palabras (CACM, CISI)
- Referencias (CACM)
- Categoria (CACM)
- Cocitaciones (CACM, CISI)
- Preguntas con autor y su perfil de trabajo (CACM)
- Glosario (Time, CACM)

TREC

- Antiguo TIPSTER, organizado por NIST y por DARPA
- Existen distintas modalidades, algunos son:
  - Ad hoc: Aparecen nuevas preguntas pero el corpus de documentos es fijo
  - Routing: Aparecen nuevos documentos pero el corpus de preguntas es fijo. Existe un corpus de entrenamiento
  - Grandes Corpus: de hasta 8 millones de documentos
- TREC tiene estadísticas propias de análisis que son las que la han dado su aceptación

Ejemplo Documento

<DOC>
<DOCNO> WSJ870324-0001 </DOCNO>
<HL> John Blair Is Near Accord To Sell Unit, Sources Say </HL>
<DD> 03/24/87</DD>
<SO> WALL STREET JOURNAL (J) </SO>
<IN> REL TENDER OFFERS, MERGERS, ACQUISITIONS (TNM) MARKETING, ADVERTISING (MKT)
TELECOMMUNICATIONS, BROADCASTING, TELEPHONE, TELEGRAPH (TEL) </IN>
<DATELINE> NEW YORK </DATELINE>
<TEXT>
John Blair & Co. is close to an agreement to sell its TV station advertising representation operation and program production unit to an investor group led by James H. Rosenfield, a former CBS Inc. executive, industry sources said. Industry sources put the value of the proposed acquisition at more than $100 million. ...
</TEXT>
</DOC>

TREC Consulta

<top> <head> Tipster Topic Description
<num> Number: 066
<dom> Domain: Science and Technology
<title> Topic: Natural Language Processing
<desc> Description: Document will identify a type of natural language processing technology which is being developed or marketed in the U.S.
<narr> Narrative: A relevant document will identify a company or institution developing or marketing a natural language processing technology, identify the technology, and identify one of more features of the company's product.
<con> Concept(s): 1. natural language processing ;2. translation, language, dictionary
<fac> Factor(s): <nat> Nationality: U.S.</nat></fac>
<def> Definitions(s): </top>

UNIVERSIDAD CARLOS III DE MADRID