Que documentos hay en la Web?

- Documentos
- Programas
- Revistas electrónicas
- Servicios de información on-line
- Bases de datos
- Catálogos de bibliotecas
- Listas de correo, listas de discusión etc.

Slide 1

Obtención de información

- Servidor
  Buscadores proporcionan un filtro de acceso a parte de la información.
- Usuario: búsqueda directa, o plantea una *Pregunta* a un buscador.

Slide 2

Buscador de información textual

Un programa de búsqueda que busca información en Internet y la organiza en el servidor, de manera que permite dar una respuesta rápida a un usuario.

El usuario puede hacer preguntas mediante la combinación booleana de palabras clave.

Como se obtiene información?

Slide 3

Como se pregunta?

Como se organiza la información?

Como se responde a la pregunta?

Obtención de información

Procedimientos automáticos **Robots o Crawlers**.

Consulta a expertos.

Utilización de recopilaciones de instituciones, bibliotecas, universidades, etc.

Slide 4

Robots, spiders, crawlers

Programas que atraviesan automáticamente la web y envian información al servidor.

Mecanismo

- Búsqueda textual en una página web
- Detección de links
- Seguir algún link

Slide 5

Residen en una m'aquina y hacen peticiones http de documentos.

Hay un standard de *exclusion de robots* mediante un fichero robots.txt

Más Información en The Web Robots Pages

http://www.robotstxt.org/wc/robots.html

Boolean queries

Se pueden combinar términos de búsqueda con operadores booleanos

Convenciones normales

AND, OR o NOT explícitos	cars AND trains NOT trucks
implícitos (+ −) por defecto OR	cars + trains - trucks
Exact match	"cars and trains but not trucks"

Slide 6

Proporciona un formalismo simple.

Puede ocasionar problemas con matching exactos.

A pesar de todo es crucial la clasificación de la respuesta.

Organización de la información

- Directorios clasificados (by subject)
- Buscadores genéricos
- Buscadores especializados
- Metabuscadores

Slide 7

Directorios clasificados (by subject)

Contienen una pequeña proporción de direcciones web, clasificadas por categorias.

Cada categoria se refina en subcategorias en forma jerárquica.

La clasificación ha sido realizada por expertos, al menos la definición de áreas.

En una segunda fase se decide la clasificación de un documento en relación a los tópicos categorizados.

Slide 8

Normalmente se puede buscar por palabra clave pero la búsqueda se hace sobre la clasificación.

Según los ultimos estudios los directorios de este tipo cubren (entre todos) menos de un 15% de la web.

`Ejemplos tipico`

Open Directory Project (ODP): www.dmoz.org

Técnicas algorítmicas

Mantenimiento de un índice.

Búsqueda en un índice.

Slide 9

Baeza Yates y Ribeiro Neto

Modern Information Retrieval

Addison Wesley

Buscadores genéricos: primera generación

Indices grandes obtenidos mediante la exploración independiente de uno o varios robots.

En vez de mantener una estructura categorizada, se almacenan los datos en una base de datos.

La respuesta es una lista de documentos (hits) ordenada por algún criterio.

Slide 10

Estudios de 1999 muestran que entre los 30 mejores cubren menos del 42% y el mejor de ellos no cubre más de un 16% de la web.

Todos combinan una categorización con la base de datos

Ejemplos buscadores genéricos

AltaVista: www.altavista.com

Excite: www.excite.org

HotBot: www.hotbot.lycos.com

Slide 11

Técnicas algorítmicas

Mantenimiento de una base de datos.

Recuperación de información en un base de datos.

Slide 12

Buscadores: Segunda generación

No intentan cubrir toda la web!

Pretenden extraer información relevante y útil.

Para ello deben clasificar páginas de acuerdo con algun criterio y seleccionar sólo las importantes.

Slide 13

Algunos criterios a ponderar

- relevancia
- popularidad
- authority

Utilizan técnicas algorítmicas sofisticadas para ponderar estos criterios.

Slide 14

Ejemplos buscadores

Google: www.google.com

DirectHit: directhit.com

FastSearch: www.alltheweb.com

HotLinks: www.hotlinks.com

Slide 15

Buscadores: Tercera generación

No intentan cubrir toda la web!

Pretenden extraer información relevante, útil.

Para ello deben clasificar páginas de acuerdo con algun criterio y seleccionar sólo las importantes nada nuevo!

Intentan extraer información semantica de la pregunta e incorporarla a la respuesta.

Slide 16

Search: Andrei Broder

Buscador detecta que se trata de una persona y localiza información personal.

Search: Star Wars

Buscador detecta que se trata de una pelicula, proporciona acceso a videoclips o fotografías.

Buscadores especializados y Metabuscadores

Buscadores especializados: Mantienen un índice de páginas de un tema.

Ejemplo típico: Notícias, Personas, Juegos, MP3.

Slide 17

Metabuscadores: pasan la pregunta a uno o más buscadores y reorganizan las respuestas.

All-In-One: www.allonesearch.com

Metaeureka: www.metaeureka.com