

Tecnologías de Información

Tema 1. Introducción y lenguajes de consulta

Bibliografía:

Baeza-Yates y Ribeiro-Neto: “Modern Information Retrieval”
1999 (Capítulo 1).

Dr. Carlos Castillo

UPF – 2005

¿Sinónimos?

Datos

Información

Conocimiento

Introducción

- ♦ Datos
 - ♦ Hechos representados en forma legible
 - ♦ *Bases de datos relacionales*
- ♦ Información
 - ♦ Datos organizados de forma coherente
 - ♦ *Sistemas de recuperación de información*
- ♦ Conocimiento
 - ♦ Información que puede ser utilizada para un propósito

Temas de esta asignatura

- ♦ 1. Lenguajes de consulta
Formular necesidades de información
- ♦ 2. Procesamiento del texto
- ♦ 3. Modelos de recuperación
Transformar contenido en información
- ♦ 4. Evaluación de los resultados
- ♦ 5. Indexación
Acelerar búsqueda de información
- ♦ 6. Aplicaciones

Temas de esta clase

- ◆ Proceso de recuperación de información
- ◆ Punto de vista del usuario
 - ◆ Lenguajes de consulta
 - ◆ Interfaces

Motivación

- ♦ Colecciones de datos enormes y potencialmente valiosos
- ♦ Cifras
 - ♦ Una presentación con algunas fotos = 500 KB
 - ♦ 1 minuto de audio comprimido = 1 MB
 - ♦ 1 CD = 650 MB
 - ♦ Biblioteca grande (congreso US) = 10 Terabytes
 - ♦ Todas las páginas Web >> 170 Terabytes
 - ♦ Todos los e-mail > 400.000 Terabytes por año

¿Cuánta información?

- ♦ Kilobyte KB = 1.000 bytes
 - ♦ 100KB = una fotografía a baja resolución
- ♦ Megabyte MB = 1.000.000 bytes
 - ♦ 5MB = todos los libros de Shakespeare
- ♦ Gigabyte GB = 1.000.000.000 bytes
 - ♦ 1GB = una camioneta llena de libros
- ♦ Terabyte TB = 1.000.000.000.000 bytes
 - ♦ 1TB = 50.000 árboles hechos papel e impresos
 - ♦ 2TB = una biblioteca grande

Números grandes

- ◆ Petabyte PB 1.000.000.000.000.000 bytes
 - ◆ 70 PB = toda el cine y tv producido en 1 año
 - ◆ 200 PB = todo el material impreso en el mundo
- ◆ Exabytes EB 1.000.000.000.000.000.000 b
 - ◆ 17 EB = todas las conversaciones telefónicas en un año
 - ◆ 5EB = cantidad de información

Mayor acceso a la información

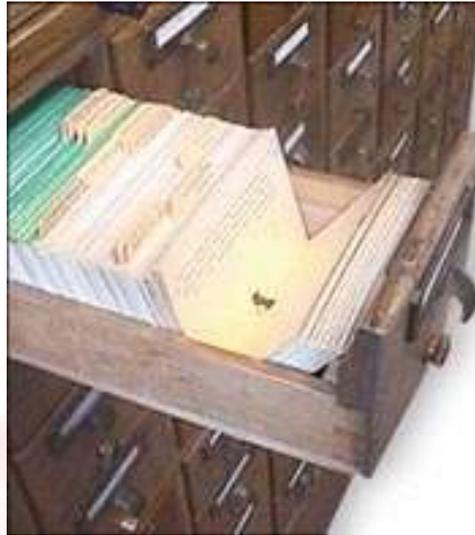
- ♦ Múltiples fuentes de información
 - ♦ Más colecciones disponibles on-line
 - ♦ Directorio de buscadores especializados:
<http://www.invisible-web.net/>
 - ♦ A veces desagregadas
- ♦ Auto-servicio de información
 - ♦ Usuarios no especializados

Algunas preguntas clave

- ♦ ¿Cómo representar contenidos?
 - ♦ Para búsquedas más efectivas
- ♦ ¿Cómo representar necesidades de información?
 - ♦ Ayudar al usuario a formular consultas
- ♦ ¿Cómo seleccionar lo relevante?
- ♦ ¿Cómo desplegar los resultados?

Desarrollos históricos

- ♦ Libros => explosión de información
- ♦ Índices en tarjetas (más versatilidad)



Sistemas de catalogación temáticos
Ejemplo: código Dewey

Código Decimal

- ♦ 000 Generalidades
- ♦ 100 Filosofía
- ♦ 200 Religión
- ♦ 300 Cs. Sociales
- ♦ 400 Lenguajes
- ♦ 500 Ciencias puras
- ♦ 600 Ciencias aplic.
- ♦ 700 Artes
- ♦ 800 Literatura
- ♦ 900 Hist. y geo.

Inventado en 1870 por Melvin Dewey, sistema jerárquico con divisiones de 10 en 10 ...

Ejemplo

- ◆ Tema: “Eclipses” en DEWEY
 - ◆ 500 Ciencias puras
 - ◆ 520 Astronomía
 - ◆ 523 Cuerpos celestes específicos
 - ◆ 523.7 Sol
 - ◆ 523.78 Eclipses



¿Por qué el esquema jerárquico no es suficiente?

- ♦ 1.- Desconocimiento de la jerarquía
- ♦ 2.- Clasificación manual
- ♦ 3.- ¿Todo se puede clasificar así?
- ♦ 4.- Múltiples jerarquías
- ♦ Posible solución:
 - ♦ Asociar palabras clave a documentos o temas
 - ♦ Búsqueda por palabras clave

Contenidos en un sentido amplio

- ♦ Texto sin etiquetar
 - ♦ Ej.: desde OCR
- ♦ Texto con campos
 - ♦ Ej.: base de datos
- ♦ Texto con estructura
 - ♦ Ej.: documentos legales, médicos
- ♦ Mapas, diagramas
- ♦ Fotografías
- ♦ Música, Videos

Contenidos multimedia

- ♦ Búsqueda por metadatos
 - ♦ Se traduce en búsqueda de texto
- ♦ Búsqueda por contenido
 - ♦ Audio: detección de voz, sonidos, etc.
 - ♦ Música: melodía, melodías similares, etc.
 - ♦ Imágenes: buscar caras, paisajes, etc.
 - ♦ Video: buscar gente, segmentación en escenas

Definición: metadatos

- ◆ “Datos acerca de otros datos”
- ◆ Ejemplos
 - ◆ Título
 - ◆ Fecha de creación
 - ◆ Autor
 - ◆ Tamaño
 - ◆ Formato
 - ◆ etc..

Proceso de recuperación de información

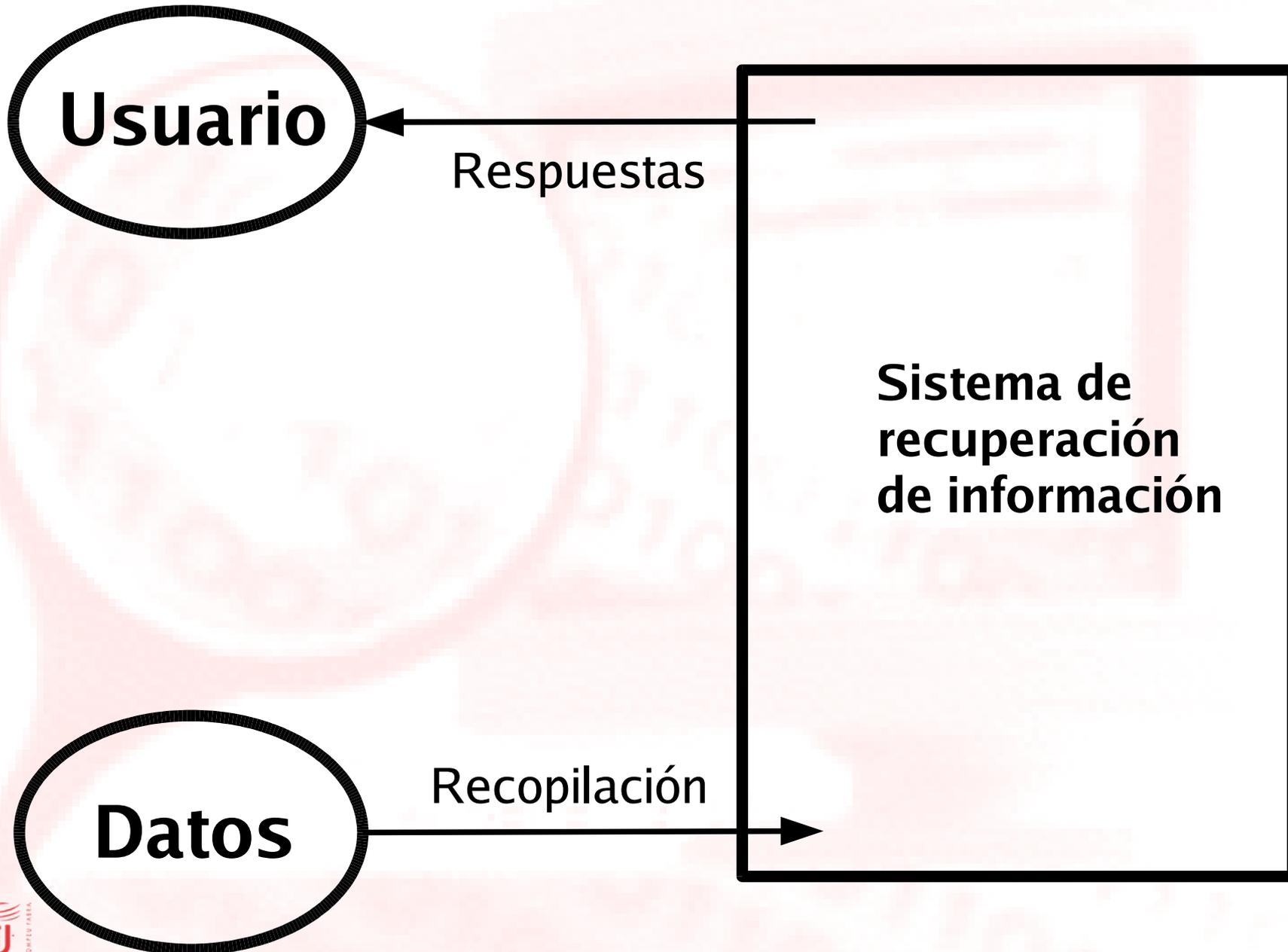
Recuperación de ...

- ♦ Recuperación de datos
 - ♦ ¿Qué documentos contienen un conjunto de palabras?
 - ♦ ¿Qué documentos responden a una cierta restricción estructural?
 - ♦ Cualquier diferencia => omitir documento
- ♦ Recuperación de información
 - ♦ Información sobre un tema o tópico
 - ♦ Semántica más relajada
 - ♦ Se toleran errores

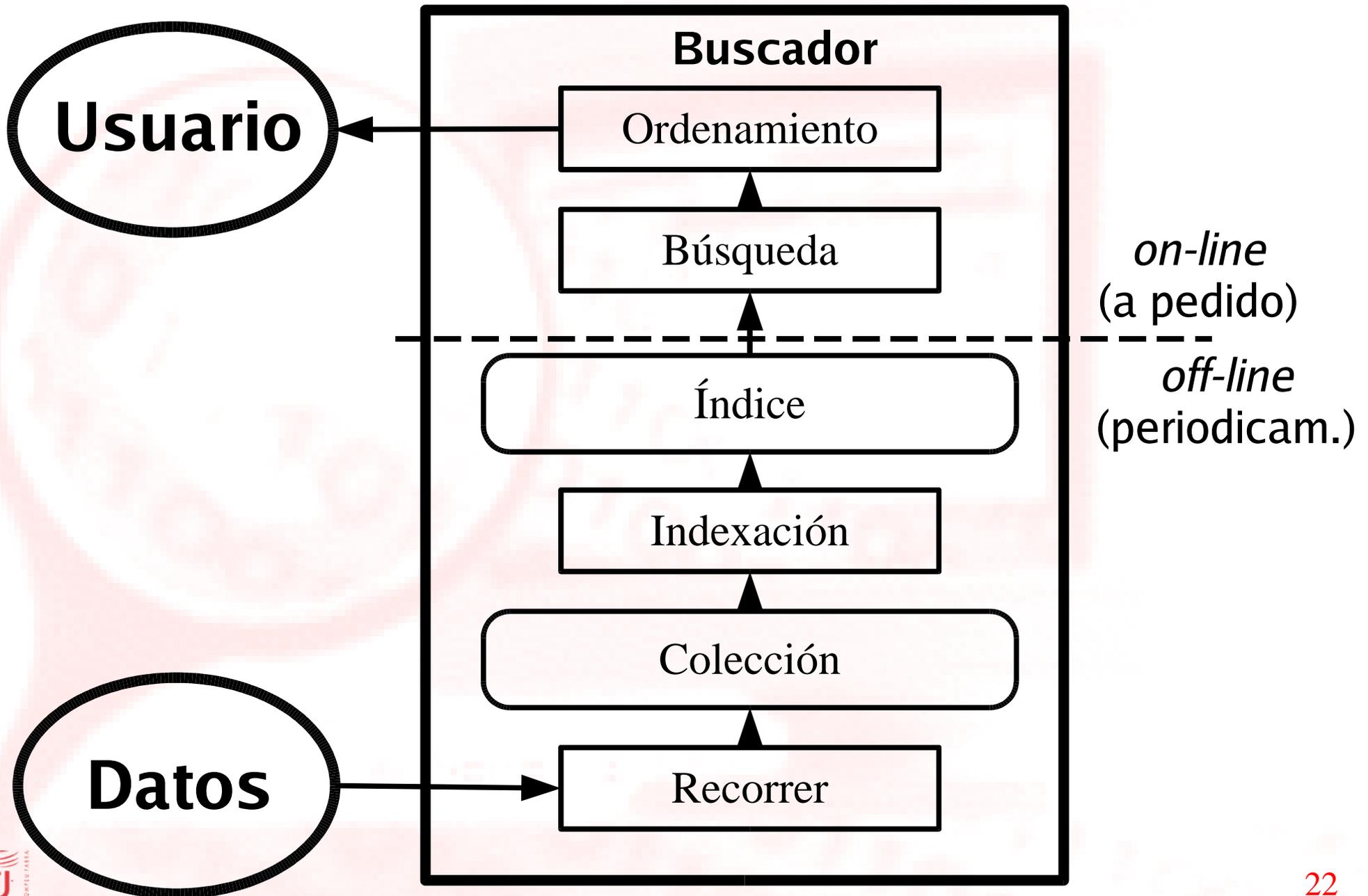
Recuperación de ...

	Datos	Información
Buscado	Registros	Documentos
Correspondencia	Exacta	Parcial
Pregunta	Completa	Difusa
Respuesta	Adecuada	Relevante
Modelo	Determinista	Probabilista
Lenguaje de consulta	Artificial	Natural
Clasificación	Un eje	Múltiples ejes

Proceso



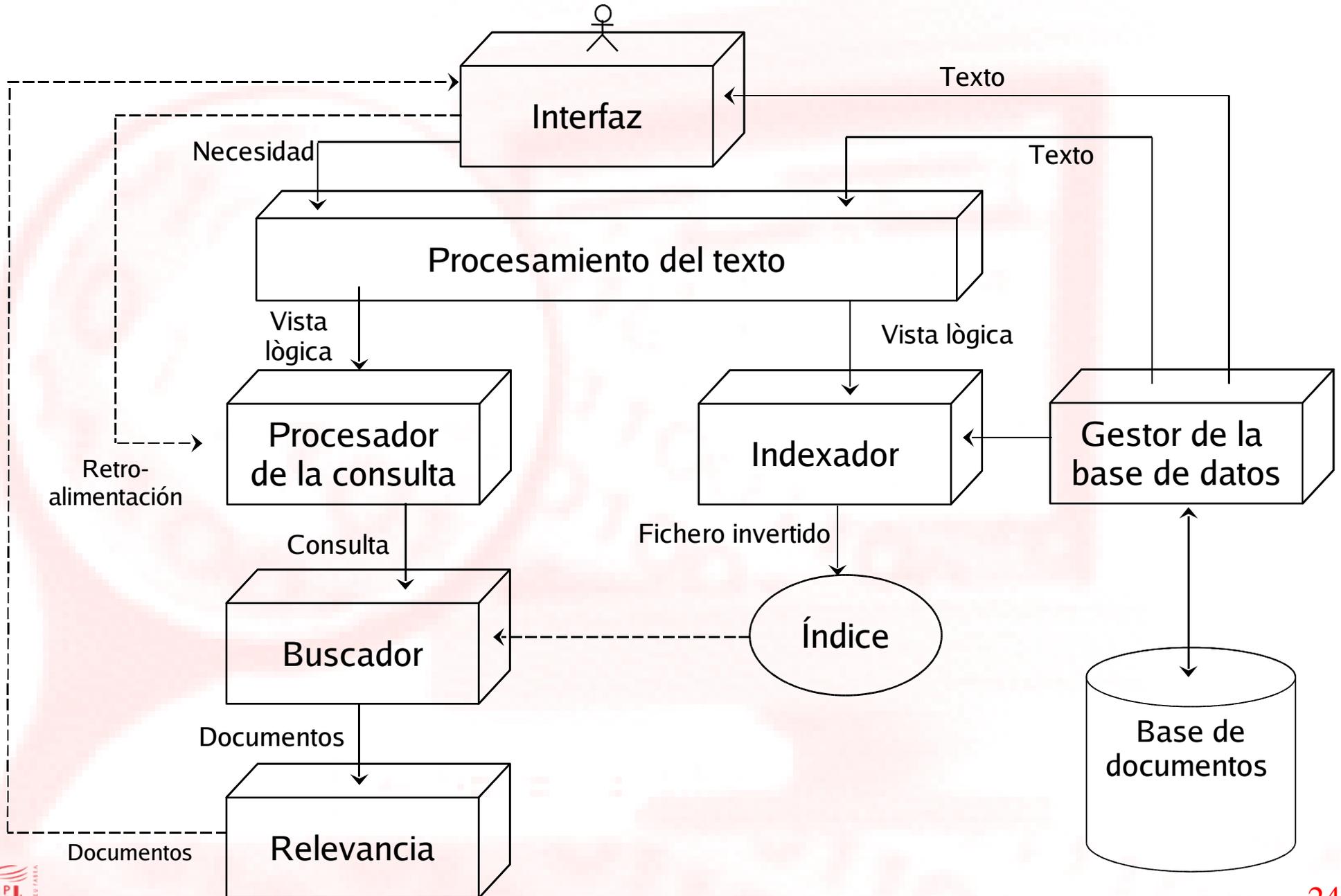
Proceso



Proceso

- ♦ Acceso a los documentos
- ♦ Procesamiento de los documentos
- ♦ Indexación de los documentos
- ♦ Necesidad de información del usuario
- ♦ Procesamiento de la petición del usuario
- ♦ Búsqueda de la respuesta
- ♦ Ordenación de los resultados
- ♦ Presentación de los resultados
- ♦ Retroalimentación

Proceso

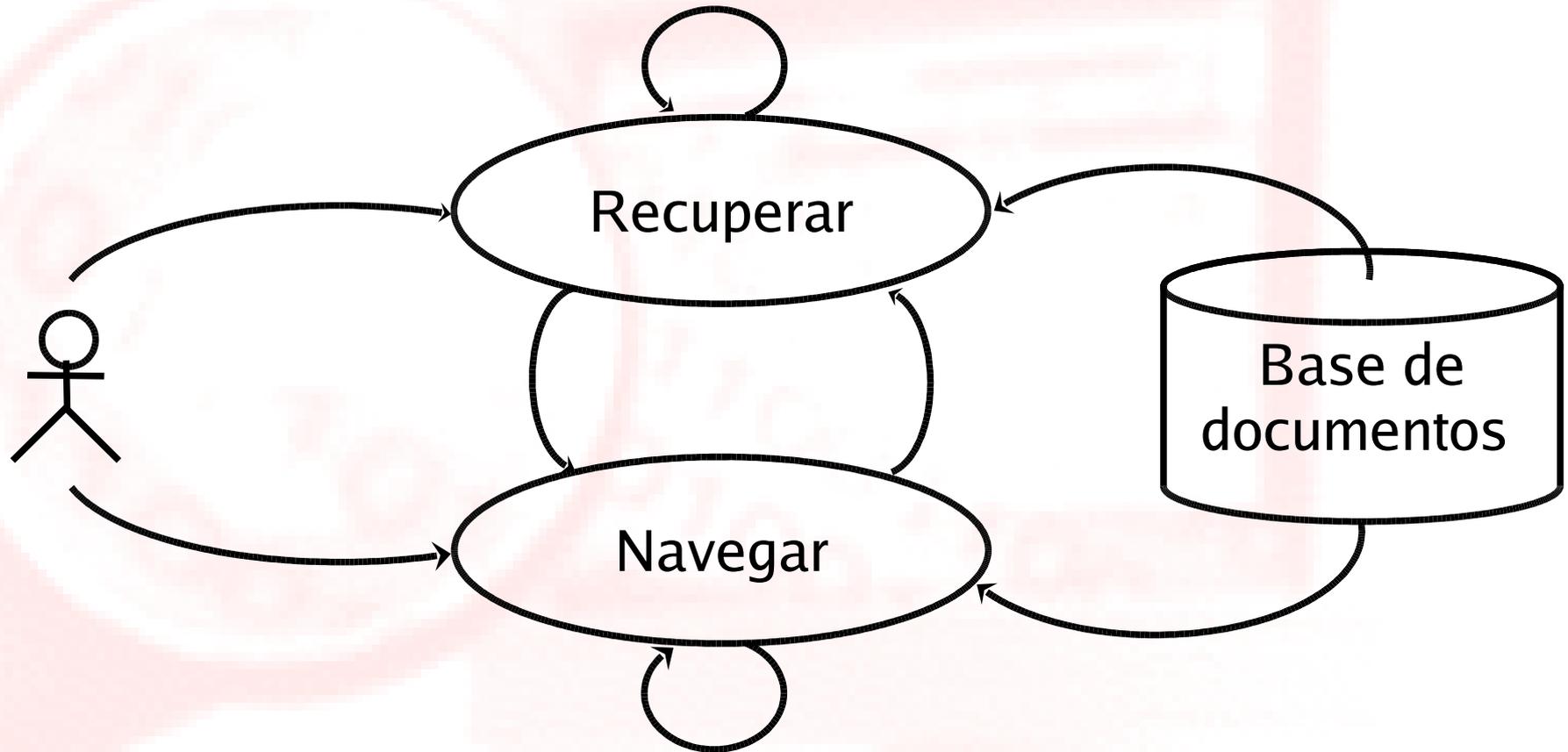


Sistema de recuperación de información

- ♦ Interpreta el contenido de los elementos de información
- ♦ Genera clasificaciones según relevancia de acuerdo a ciertos parámetros
- ♦ Si hay pocos elementos a buscar
 - ♦ Basta con retornar aquellos que son apropiados
- ♦ Si hay muchos elementos a buscar
 - ♦ La parte “fácil” es encontrar cuáles son apropiados
 - ♦ La parte “difícil” es seleccionar unos pocos

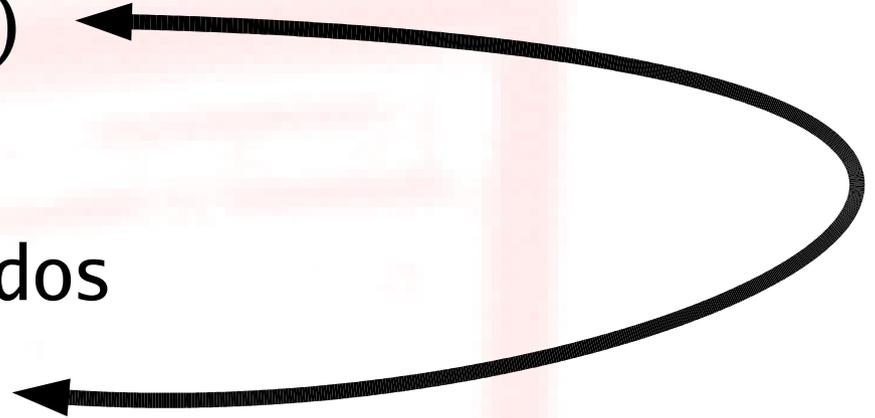
Punto de vista del usuario

Función del usuario



Función del usuario

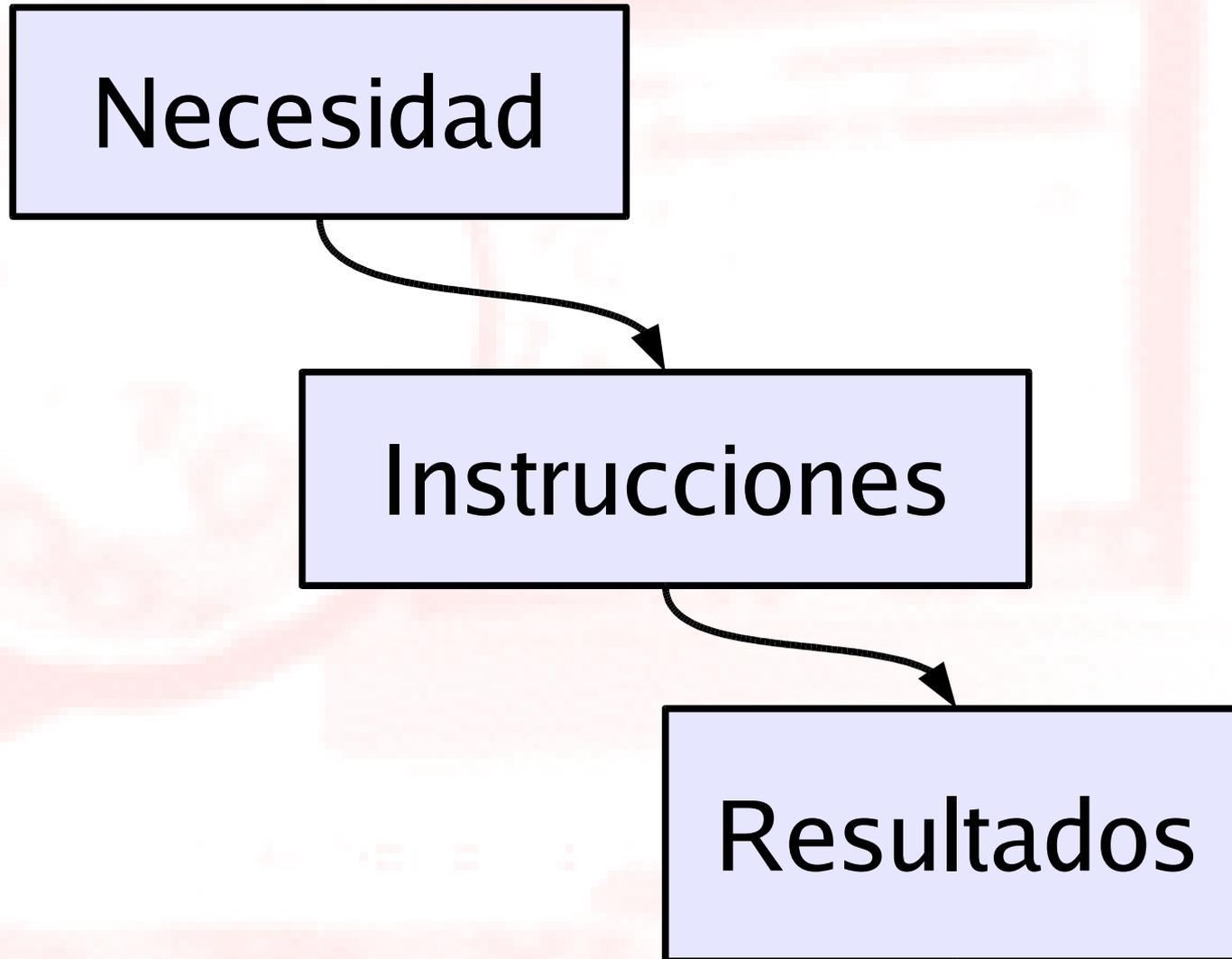
- ◆ Recuperación (retrieval)
 - ◆ Ingresar palabras clave
 - ◆ Revisar listas de resultados
- ◆ Navegación (browsing)
 - ◆ Navegar por sitios



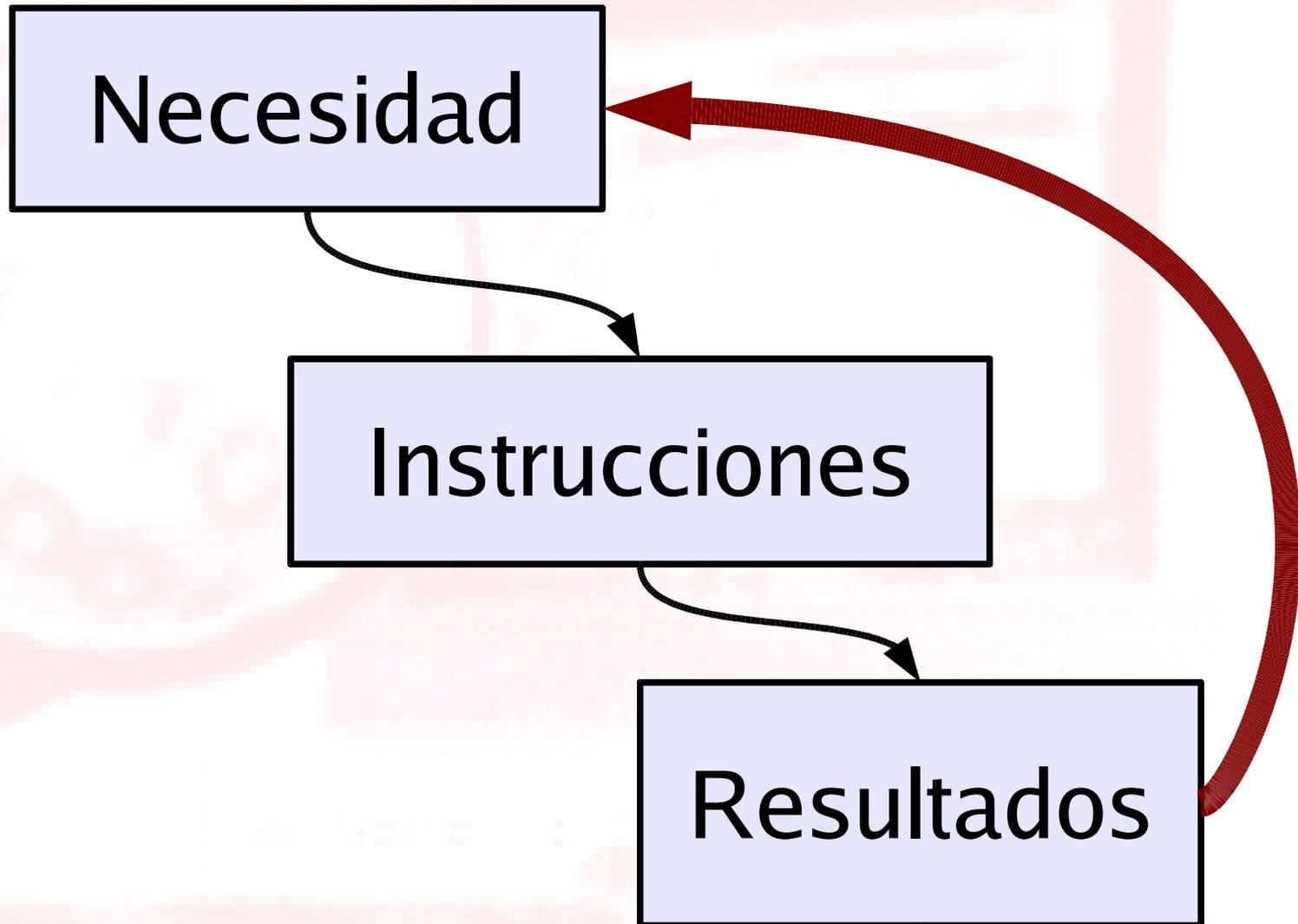
No es sólo filtrado (filtering)

- ♦ Filtering
 - ♦ Los usuarios deciden cuáles son sus preferencias, y cada vez que reciben un documento, deciden si les interesa o no
 - ♦ Las preferencias del usuario no cambian durante la sesión
- ♦ No es cierto, porque ...
 - ♦ Los usuarios no tienen claras sus preferencias a priori
 - ♦ Las preferencias van cambiando conforme se aprende durante la búsqueda

¿Comportamiento del usuario?



Comportamiento del usuario



Búsquedas típicas (zeitgeist)

Spain

Popular Sports Queries 2004

1. [fernando alonso](#)
2. [ronaldinho](#)
3. [valentino rossi](#)
4. [real madrid](#)
5. [barcelona](#)

Popular Queries 2004

1. [el mundo](#)
2. [renfe](#)
3. [fernando alonso](#)
4. [ronaldinho](#)
5. [valentino rossi](#)
6. [fernando torres](#)
7. [hilary duff](#)
8. [melendi](#)
9. [spanair](#)
10. [los serrano](#)

Popular Google News Queries - 2004

1. [real madrid](#)
2. [alejandra sanz](#)
3. [boda real](#)
4. [irak](#)
5. [eta](#)
6. [ronaldinho](#)
7. [atentado en madrid](#)
8. [la casa de tu vida](#)
9. [rajoy](#)
10. [princesa letizia](#)

Popular Travel Queries 2004

1. [renfe](#)
2. [spanair](#)
3. [iberia](#)
4. [vuelos baratos](#)
5. [air europa](#)

Necesidades del usuario

- ♦ Por tipo
 - ♦ Específica
 - ♦ Acción
 - ♦ No específica
- ♦ Por ocurrencia
 - ♦ Frecuente
 - ♦ Infrecuente
- ♦ Por formato buscado
 - ♦ Artículo largo, artículo breve, código fuente, sólo definición, lista de referencias, imágenes, etc.

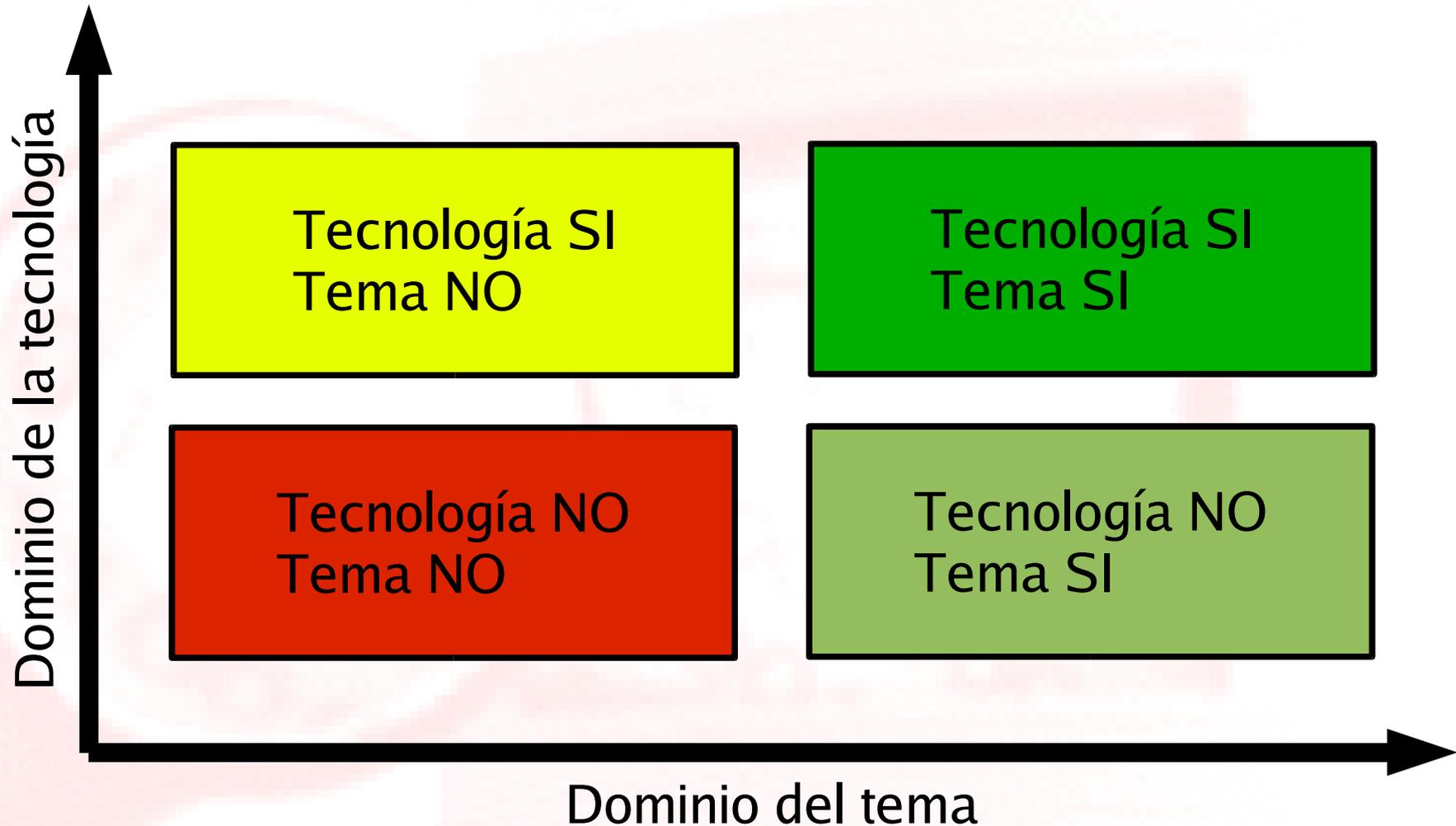
Evaluar un sistema de RI desde un usuario

- ◆ Tiempo de aprendizaje
- ◆ Sobrecarga cognitiva
- ◆ Expresividad
 - ◆ ¿Qué consultas se pueden hacer?
- ◆ Tiempo hasta primer resultado correcto
- ◆ Cantidad de *backtracking*

Usuarios

- ♦ Dominio tecnológico
 - ♦ Experto
 - ♦ No “alfabetizado” en términos informáticos => ¿informatizado?
- ♦ Dominio del tema consultado
 - ♦ Profundo
 - ♦ Superficial

Usuarios (cont...)



Para un usuario, es más importante saber qué palabras clave usar que conocer opciones avanzadas del buscador

Lenguajes de consulta e interfaces

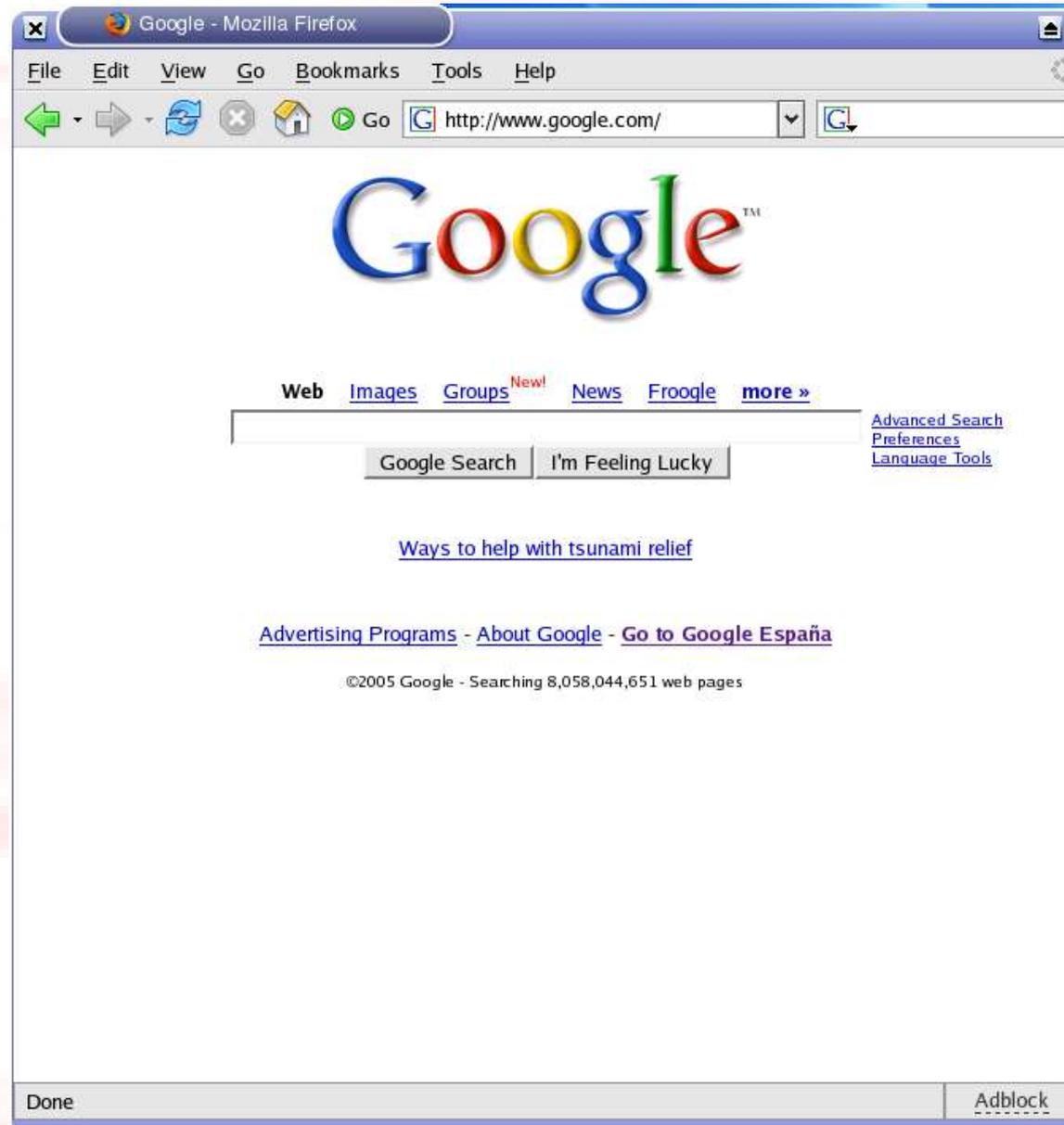
Definiciones

- ◆ Lenguaje de consultas
 - ◆ Recuperar información: admite ranking
 - ◆ Recuperar datos: no admite ranking
- ◆ Unidad mínima de recuperación
 - ◆ Sitio Web
 - ◆ Página
 - ◆ Extracto (abstract)

Información disponible

- ♦ Base de datos documental
 - ♦ Contenido
 - ♦ Propiedades
 - ♦ Estructura
 - ♦ Contexto de cada documento
- ♦ Base de datos de usuarios
 - ♦ Preferencias
 - ♦ Búsquedas anteriores
 - ♦ ... perfil

Ejemplo: buscador



Ejemplo: buscador + portal



Ejemplo: búsqueda avanzada

The screenshot shows the Google Advanced Search page in a Mozilla Firefox browser window. The browser's address bar displays the URL `http://www.google.com/advanced_s`. The page features the Google logo and the text "Advanced Search" with links for "Advanced Search Tips" and "About Google".

The main search area is titled "Find results" and includes four radio button options for search criteria: "with all of the words", "with the exact phrase", "with at least one of the words", and "without the words". Each option has an associated text input field. To the right of these options is a dropdown menu set to "10 results" and a "Google Search" button.

Below the search options are several filter sections:

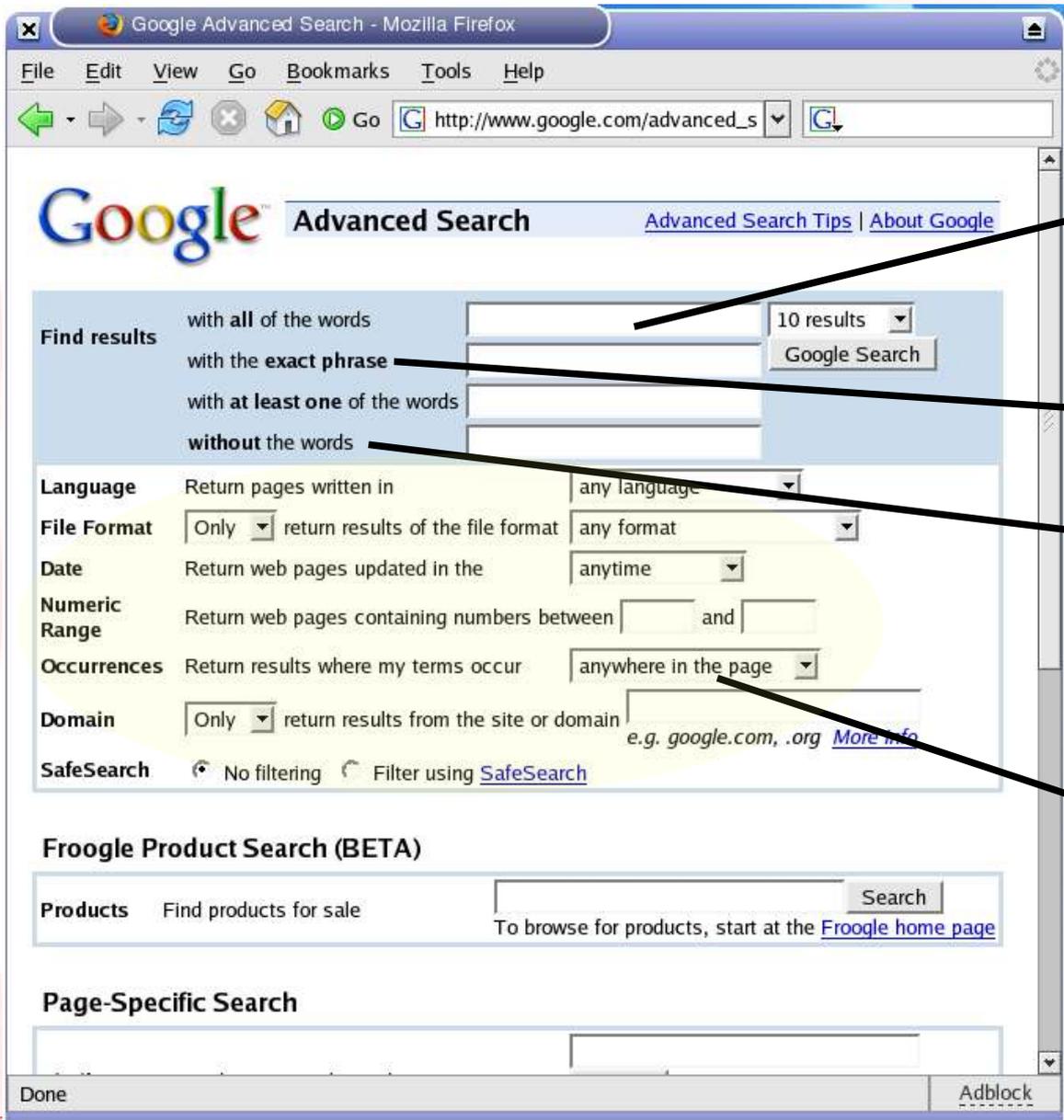
- Language:** "Return pages written in" with a dropdown menu set to "any language".
- File Format:** "Only" dropdown, "return results of the file format" with a dropdown menu set to "any format".
- Date:** "Return web pages updated in the" with a dropdown menu set to "anytime".
- Numeric Range:** "Return web pages containing numbers between" with two empty input fields and an "and" separator.
- Occurrences:** "Return results where my terms occur" with a dropdown menu set to "anywhere in the page".
- Domain:** "Only" dropdown, "return results from the site or domain" with an empty input field and a link to "More info" with the example "e.g. google.com, .org".
- SafeSearch:** Radio buttons for "No filtering" (selected) and "Filter using SafeSearch".

At the bottom of the page, there are two additional sections:

- Froogle Product Search (BETA):** A "Products" section with the text "Find products for sale", an empty input field, and a "Search" button. Below it is a link: "To browse for products, start at the Froogle home page".
- Page-Specific Search:** A section with an empty input field.

The browser's status bar at the bottom shows "Done" on the left and "Adblock" on the right.

Elementos



Keywords

Operadores proximidad

Operadores booleanos

Metadatos

Keywords o palabras clave

- ♦ Una palabra
 - ♦ Definición de qué es *keyword*
- ♦ Palabras en contexto: proximidad
 - ♦ Proximidad 1 (frase)
 - ♦ Con stopwords
 - ♦ Sin stopwords
 - ♦ Proximidad arbitraria
 - ♦ Proximidad y ordenamiento (más próximas o en mismo elemento estructural, mejor)

Consultas booleanas

- ♦ AND, OR, NOT
 - ♦ AND (conjunción)
 - ♦ OR (disyunción)
- ♦ Problema: en el lenguaje no se usa así
 - ♦ Ej.: busco pisos grandes y soleados
 - ♦ Deben tener ambas características ($y=AND$)
 - ♦ Ej.: mujeres y niños primero
 - ♦ Basta con una característica ($y=OR$)

Lenguaje natural

- ♦ En teoría
 - ♦ Representación del conocimiento
 - ♦ Meta final de la recuperación de información

Ejemplo lenguaje natural

YOU MAY FIND THESE OPTIONS USEFUL:

Other people with your search have also asked:

-  [Where can I find](#) government information for Nigeria ?
-  [Where can I find encyclopedic geographical resources on](#) Nigeria ?



Web | Pictures | News | Local **NEW!** | Products | More »

What time is it in Japan?

Search

Advanced

Web Search: What time is it in Japan?

1-10 results out of 12,740,000



It is 3:02:33 AM JST in Tokyo, Japan.

Email This | About

Related Topics

- › Japan
- › Timezones
- › Time Conversion

Web Results



Web | Pictures | News | Local **NEW!** | Products | More »

How old is Madonna?

Search

Advanced

Web Search: How old is Madonna?

1-10 results out of 876,400



Source

Madonna is 46 years old.

At first a streetwise bubblegum-pop ragamuffin, Madonna used a mixture of talent, cleavage, and relentless self-promotion to become one of the most famous recording artists of the 20th century. She released her self-titled first album in 1983; other albums include Like... [More»](#)

Find: [Pictures](#) | [News](#) | [Products](#)

Go To: [Official Web Site](#) | [Filmography](#) | [Discography](#)

Email This | About

Related Topics

- › Biography
- › Singer
- › When Madonna
- › Madonna Birth
- › Full Name Madonna

Consultas en lenguaje natural

- ♦ En la práctica
 - ♦ Traducidas a boolean y pesos
 - ♦ Uso de diccionario de consultas
 - ♦ Keywords y representante
 - ♦ Funcionan bien si hay pocas preguntas posibles
- ♦ Pueden ser muy demandantes computacionalmente

Satirewire.com (1999)

You asked: Thanks for being with us today, Jeeves. How are you?

Ask!

Ask! What day is it?

You asked: It's Monday.

Ask!

Ask! 9 matches by [About.com](#) Monday Again?

You asked: Yes, they do tend to recur. As often as once a week. What's wrong with Mondays?

Ask!

Ask! 10 matches by [AltaVista](#) What's Wrong with Garbage Disposals?

Interfaces con clustering: vivisimo

The screenshot displays the Vivísimo search engine interface in a Mozilla Firefox browser window. The search query is 'upf' and the results are clustered. The interface includes a navigation menu, a search bar, and a list of search results with their respective descriptions and source information.

Clustering Summary:

- upf (182)**
 - Universitat (54)**
 - Protection, Clothing (35)**
 - Universal Preservation Format (10)**
 - Foundation, Dedicated (5)**
 - IUA (5)**
 - Publicaciones (6)**
 - Technology (5)**
 - Upf Organization To Promote The Common Good (2)**
 - Skid Units (2)**
 - Applications (3)**

Search Results:

- Children's UV Swimwear** [new window] [preview] Sponsored Link
Maximum UV protection, maximum fashion. One and 2-piece swimsuits with matching hats, shoes and sunglasses. All for kids and all stylish and fun. UPF 50+ 97.8% UV blocking.
www.tugasunwear.com
- 1. United Pegasus Foundation** [new window] [frame] [preview]
Dedicated to providing race horses with dignified retirements, this association's page presents a mission statement, adoption information, and contacts.
URL: www.unitedpegasus.com - [show in clusters](#)
Sources: MSN 3, Lycos 4, Looksmart 5, Wisenut 5
- 2. University of Florida - University Press of Florida** [new window] [frame] [preview]
Check out manuscript submission criteria, the book of the month, recommended summer reading, and featured publication titles. Offers ordering and shipping options.
URL: www.upf.com - [show in clusters](#)
Sources: Looksmart 1, Open Directory 1, Lycos 7, MSN 45
- 3. Universitat Pompeu Fabra** [new window] [frame] [preview]
Segmented campus that spans all of Barcelona offers 12 majors and boasts a huge library. Read about its computer network and int'l programs.
URL: www.upf.es - [show in clusters](#)
Sources: Lycos 2, MSN 2, Looksmart 3
- 4. UPF** [new window] [frame] [preview]
UPF organization to promote the common good of all, to promote world peace thru

Interfaces con clustering: clusty

The screenshot shows the Clusty web interface in a Mozilla Firefox browser window. The address bar displays the URL <http://clusty.com/search?query=upf>. The page features a navigation bar with tabs for Web+, News, Images, Shopping, Encyclopedia, Gossip, and Customize!. Below this is a search bar containing the text 'upf' and a 'Cluster' button. A sidebar on the left allows users to 'Cluster by' Topics and lists various clusters for 'upf' (214 results), including 'Universitat, Pompeu Fabra' (58), 'Protection, Clothing' (33), 'Universal Preservation Format' (8), 'Foundation, Dedicated' (6), 'UPF 30' (7), 'Publicaciones' (6), 'Promote, Organization' (4), and 'Master' (6). The main content area shows 'Top 214 results of at least 40,253 retrieved for the query upf'. It includes sponsored results for 'Children's UV Swimwear' and 'Sun Protective Clothing and Accessories', and search results for 'University of Florida - University Press of Florida' and 'Universitat Pompeu Fabra'. The browser status bar at the bottom shows 'Done' and 'Adblock'.

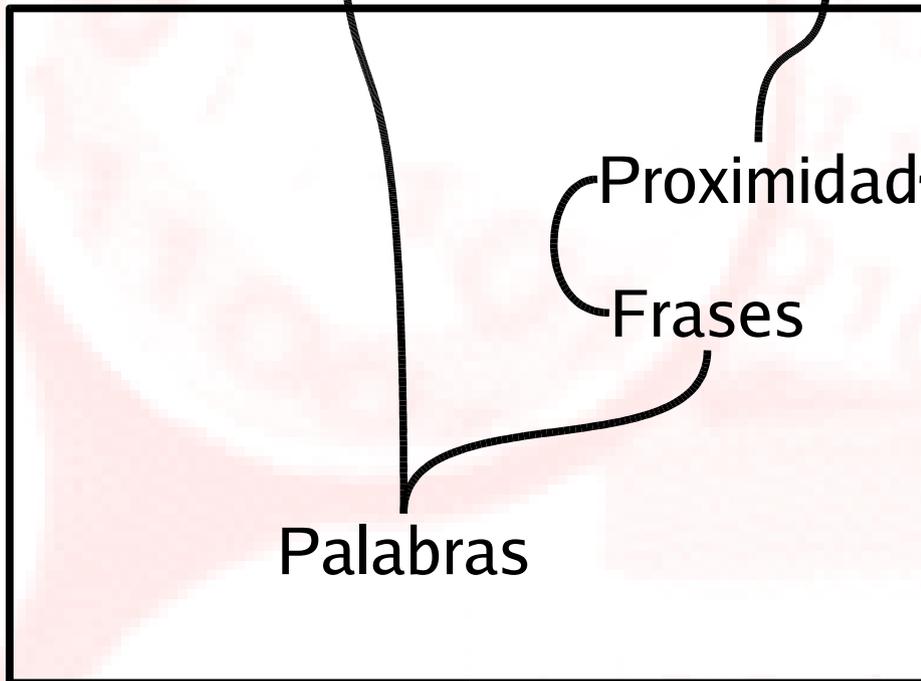
Interfaces con clustering, clustering gráfico: kartoo

The screenshot displays the KartOO web interface within a Mozilla Firefox browser window. The browser's address bar shows the URL `http://www.kartoo.com/flash04.php3`. The interface features a search bar with the query 'upf' and a 'Search' button. The main content area is a semantic map where nodes represent search results and their relationships. Nodes include domain names like `www.sun-togs.co.uk`, `www.upf.com.au`, `www.iaa.upf.es`, `www.unitedpegasus.com`, `info.wgbh.org`, `www.poloclubchantilly.com`, `www.upf.org`, `www.annuairefuneraire.com`, `psychogenicfugue.com`, `www.vcharite.univ-mis.fr`, and `www.presse-francophone.info`. Semantic terms such as 'range', 'produce', 'project', 'service', 'générale', 'help', 'organization', 'practices', 'mba.upf.es', 'universal', 'hands', 'presse', 'union', and 'asset' are placed between nodes to indicate relationships. A sidebar on the left lists 'Topics' and 'Popular Queries'. A 'Found sites' list is on the right, and a 'KartOO SITE BOX' provides information about the search solution. The bottom of the interface includes a 'next map' button and a status bar showing 'Transferring data from ww2.kartoo.com...'. The footer of the browser window contains the 'Adblock' logo.

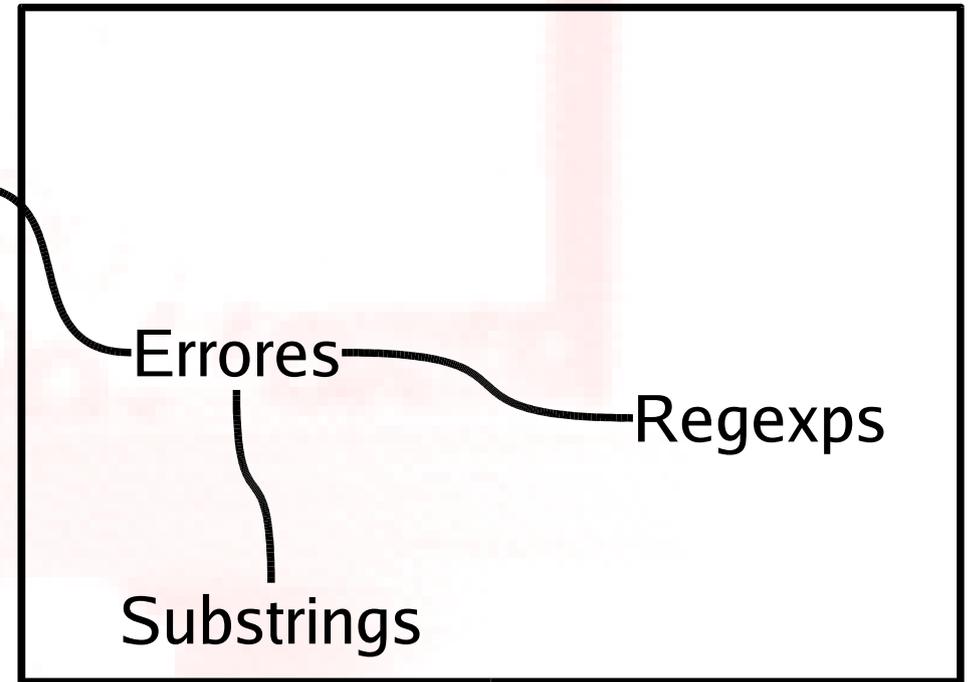
Lenguaje Natural

Fuzzy Boolean

Estructuradas



Palabras clave y contexto



Correspondencia patrones

Desafíos

- ♦ Cambiar la estructura y no las consultas
- ♦ Las consultas pueden ser complejas
 - ♦ Interfaz
- ♦ Lenguaje natural

Resumen

- ♦ Datos != Información
- ♦ Recuperación de información => relevancia
 - ♦ Admite errores
- ♦ Necesidad del usuario es cambiante
- ♦ Formular esta necesidad puede ser una tarea compleja