Recuperación de información.

Es una ciencia con metodologías de observación sistemática, medición, deducción e inducción, que buscan la ubicación y posicionamiento de la información.

El proceso de recuperación se lleva a cabo mediante consultas a la base de datos donde se almacena la información estructurada, mediante un lenguaje de interrogación adecuado. Es necesario tener en cuenta los elementos clave que permiten hacer la búsqueda, determinando un mayor grado de pertinencia y precisión, como son: los índices, palabras clave, tesauros y los fenómenos que se pueden dar en el proceso como son el ruido y silencio documental. Uno de los problemas que surgen en la búsqueda de información es si lo que recuperamos es «mucho o poco» es decir, dependiendo del tipo de búsqueda se pueden recuperar multitud de documentos o simplemente un número muy reducido. A este fenómeno se denomina Silencio o Ruido documental.

Silencio documental: Son aquellos documentos almacenados en la base de datos pero que no han sido recuperados, debido a que la estrategia de búsqueda ha sido demasiado específica o que las palabras clave utilizadas no son las adecuadas para definir la búsqueda.

Ruido documental: Son aquellos documentos recuperados por el sistema pero que no son relevantes. Esto suele ocurrir cuando la estrategia de búsqueda se ha definido demasiado genérica.

- Ya sabemos cómo se generan documentos digitales y conocemos los distintos formatos.
- Veremos cómo se recuperan centrándonos en un medio específico: Internet.
- Internet: red de ordenadores conectados, con una enorme cantidad de sitios Web, y por tanto de información.
- En la Web tenemos una gran base de datos con información de todo tipo: texto, imágenes, audio y vídeo, y en múltiples formatos.
- La Web tiene una serie de características específicas (problemas intrínsecos de los datos), como son:

– La información está distribuida en muchos ordenadores distintos.

– Hay un gran volumen de datos, que además son volátiles, ya que aparecen y desaparecen continuamente nuevas páginas.

– No se conoce a priori la estructura de la información, y gran parte se genera dinámicamente mediante consultas a bases de datos.

– Hay mucha redundancia de información (webs repetidas, webs con el mismo contenido).

– Los datos son heterogéneos, con diferentes tipos de formatos de ficheros.

– La calidad no es la misma en todas las fuentes de información.

Lista de algunos buscadores que se utiliza:

Los buscadores son herramientas que permiten localizar y recuperar la información almacenada en internet. El funcionamiento es parecido a las bases de datos, almacenan las páginas con determinadas características (metadatos) y que posteriormente tras utilizar unas palabras clave emiten un listado de las más relevantes.

Buscadores generales

- Google (http://www.google.com)
- Alltheweb (http://www.alltheweb.com)
- AltaVista (http://www.altavista.com)
- Excite (http://www.excite.com)
- Infoseek (http://www.infoseek.com)
- Lycos (http://www.lycos.com)
- Webcrawler (http://webcrawler.com)
- Hotboot (http://www.hotbot.com)

Directorios:

Los directorios son listas organizadas que nos permite acceder a la información de forma estructurada y jerárquica. Se clasifican en categorías y el usuario enlaza de lo más general a lo más específico

- Recomendados para las búsquedas en las que el usuario no sabe mucho sobre el tema en concreto
    - El directorio de Google (http://directory.google.com)
    - Ozú (http://categorias.ozu.es)
    - El índice (http://www.elindice.com)
    - Yahoo (http://www.yahoo.com)
- Directorio y motores especializados
    - Humbul http://www.humbul.ac.uk
    - Librarian Index to the Internet http://lii.org
    - Internet Public Library http://www.ipl.org
    - Scirus http://www.scirus.com
    - Search4Science http://www.search4science.com

Metabuscadores:

Son buscadores, con la cualidad de que no sólo buscan en una única base de datos, sino que al introducir los conceptos de búsqueda hace el barrido en distintas bases de datos, de esta forma la amplitud de resultados es mayor.

    - Vivisimo (http://www.vivisimo.com)
    - Dogpile (http://www.dogpile.com)
    - Kartoo (http://www.kartoo.com)
    - Qbsearch (http://www.qbsearch.com)
    - Metacrawler: (http://www.metacrawler.com)
- Buscadores selectivos. Utilizan una base de datos especializada en una materia.
    - Ask (http://www.ask.com)
    - Teoma (http://www.teoma.com)
    - Electric Library (http://www.elibrary.com)
    - Hieros Gamos http://www.hg.org/index.html
- Programa para buscar
    - Copernic (http://www.copernic.com)
- Agentes inteligentes. Los agentes inteligentes son herramientas que permiten localizar información de forma automática, sólo necesita que se le definan un perfil de búsqueda y donde debe lanzarla (bases de datos, sitios web, etc.) y, automáticamente va presentando un informe sobre la nueva información que va surgiendo.
    - BookWhere http://www.bookwhere.com
    - BullsEye Pro http://www.intelliseek.com
    - WebSeeker 5 http://www.bluesquirrel.com/
    - WebFerret http://www.ferretsoft.com

Técnicas de recuperación de la información con motores de Búsqueda

Procedimiento a seguir:

1. Definir bien el objetivo de la búsqueda

2. Utilizar estrategias de búsqueda de acuerdo al objetivo

3. Ordenarlas según su eficacia y eficiencia.

4. Replanteamiento de estrategia y/o buscadores de recursos (directorios, motores de búsqueda conceptuales) si no ha obtenido los resultados esperados.

Técnicas de Búsqueda Avanzadas (Google)

• Documentos en un cierto formato:

Sistemas operativos **filetype:** ppt

• Páginas que apuntan a otras:

link: wikipedia.com

link: microsoft .com -**inurl:** microsoft.com

• Búsqueda de palabras cercanas:

explosion * super nova

• Búsqueda de Definiciones

define: computer

• Búsqueda de Sinónimos

lenguaje c **intitle:** ~curso

• Información sobre un sito

info: www.fcharte.com

• Búsqueda de Sitios relacionados

related: www.astroseti.org

• Búsqueda dentro de un dominio:

site: www.astroseti.org supernova

Técnicas de Búsqueda Avanzadas (Bing)

• Enlaces a documentos de tipo específico

tenis **contains:** pdf

tenis **filetype:** pdf

• Encuentra páginas que están alojadas en un determinado host que tienen la dirección **ip** que tu busca

ip:207.241.148.80

• Buscar en un determinado idioma:

tenis (**language:** fr)

• Encuentra páginas que contengan una determinada palabra en el “body” de una página. **inbody:** tennis

• Encuentra páginas que contengan una determinada palabra en el “title” de una página. **intitle:** tennis

• Limita tu búsqueda a un dominio específico:

site: .org

site: .gov

site: .edu

• Encontrar páginas que en el url contengan unos determinados términos

url: about.com

• Obtener sitios web que contenga que cuentan con un sistema de subscripcion (RSS o ATOM) astronomia **hasfeed:** tennis

Técnicas de recuperación de información.

Sistemas de recuperación de lógica difusa

Esta técnica permite establecer consultas con frases normales, de forma que la máquina al realizar la búsqueda elimina signos de puntuación, artículos, conjunciones, plurales, tiempos verbales, palabras comunes (que suelen aparecer en todos los documentos), dejando sólo aquellas palabras que el sistema considera relevantes. La recuperación se basa en proposiciones lógicas con valores de verdadero y falso, teniendo en cuenta la localización de la palabra en el documento

Técnicas de ponderación de términos

Es común que unos criterios en la búsqueda tengan más valor que otros, por tanto, la ponderación pretende darle un valor adecuado a la búsqueda dependiendo de los intereses del usuario. Los documentos recuperados se encuentran en función del valor obtenido en la ponderación. El valor depende de los términos pertinentes que contenga el documento y la frecuencia con que se repita. De forma que, el documento más pertinente de búsqueda sería aquel que tenga representado todos los términos de búsqueda y además el que más valor tenga repetidos más veces, independientemente de donde se localice en el documento.

Técnica de clustering

Es un modelo probabilístico que permite las frecuencias de los términos de búsqueda en los documentos recuperados. Se atribuyen unos valores (pesos) que actúan como agentes para agrupar los documentos por orden de importancia, mediante algoritmos ranking.

Algoritmos utilizados para realizar la categorización (cluster):

- Algoritmo K-means
- COBWEB
- Algoritmo EM

Técnicas de retroalimentación por relevancia

Esta técnica pretende obtener el mayor número de documentos relevantes tras establecer varias estrategias de búsqueda. La idea es que, tras determinar unos criterios de búsqueda y observar los documentos recuperados se vuelva a repetir nuevamente la consulta, pero esta vez con los elementos interesantes, seleccionados de los documentos primeramente recuperados.

Algoritmo Genético: es el que se ha utilizado para llevar a cabo este tipo de técnicas de recuperación http://www.pmsi.fr/gainits.htm

Técnicas de stemming

Morfológicamente las palabras están estructuradas en prefijos, sufijos y la raíz. La técnica de Stemming lo que pretende es eliminar las posibles confusiones semánticas que se puedan dar en la búsqueda de un concepto, para ello trunca la palabra y busca solo por la raíz.

Algoritmos utilizados para desechar prefijos y sufijos:

- Paice/Husk
- S-stemmer / n-gramas
- Técnicas lingüísticas

Pretenden acotar de una manera eficaz los documentos relevantes. Por esta razón, esta técnica lo consigue mediante una correcta indización en el proceso de tratamiento de los documentos con ayuda de índices, tesauros, etc.; evitando las ambigüedades léxicas y semánticas a la hora de establecer las consultas.

Fuentes consultado.

http://ocw.uv.es/ingenieria-y-arquitectura/infoirmatica-2/tema_e.pdf

https://onretrieval.com/recuperacion-de-informacion-y-de-datos/

http://glossarium.bitrum.unileon.es/Home/recuperacion-de-informacion

http://www.mariapinto.es/e-coms/busqueda-y-recuperacion-de-informacion/