



## CENTRO DE ESTUDIOS FINANCIEROS

VIRIATO, 52	28010 MADRID	914 44 49 20
PONZANO, 15	28010 MADRID	914 44 49 20
G. DE GRÀCIA, 171	08012 BARCELONA	934 15 09 88
ALBORAYA, 23	46010 VALENCIA	963 61 41 99

[www.cef.es](http://www.cef.es)

[info@cef.es](mailto:info@cef.es)

## Índice Tema 13

---

1. Introducción.
  - 1.1. Introducción a la gestión documental.
2. Captura y archivo electrónico de documentos.
  - 2.1. Captura de la fuente de información.
  - 2.2. Almacenamiento de la información.
  - 2.3. Salidas de la información.
3. Organización funcional de un sistema documático.
  - 3.1. Análisis documental.
  - 3.2. Almacenamiento en la base de datos documental.
  - 3.3. Consulta y recuperación de la información.
  - 3.4. Técnicas y lenguajes específicos de interrogación.
4. Parámetros básicos de evaluación de resultados obtenidos.





## CENTRO DE ESTUDIOS FINANCIEROS

VIRIATO, 52	28010 MADRID	914 44 49 20
PONZANO, 15	28010 MADRID	914 44 49 20
G. DE GRÀCIA, 171	08012 BARCELONA	934 15 09 88
ALBORAYA, 23	46010 VALENCIA	963 61 41 99

www.cef.es

info@cef.es

### TEMA 13

**Documática. Archivo electrónico de documentos. Organización funcional de los sistemas documentáticos. Optimización de consultas y recuperación de la información.**

#### 1. INTRODUCCIÓN.

Actualmente, con el avance de los sistemas basados en web se ha producido una gran mejora respecto al manejo de la información. Podemos acceder a la información a través de los navegadores, tener la información distribuida y gestionada de forma electrónica. Esto es, una «oficina sin papeles». Así podremos publicar nuestros contenidos de manera más eficiente y automática, el almacenamiento y recuperación será más rápido y eficiente, utilizando, por supuesto, las herramientas adecuadas.

En este tema vamos a tratar precisamente de todas las operaciones que vamos a realizar para tener un sistema documental, desde el momento en el que se recibe el documento y se transforma a través del reconocimiento electrónico de documentos, pasando por su almacén en la base de datos hasta el momento de su recuperación producida por una petición de consulta efectuada por el usuario.

##### 1.1. INTRODUCCIÓN A LA GESTIÓN DOCUMENTAL.

La gestión documental sirve para facilitar y gestionar un buen almacenamiento, indización y recuperación de documentos, generalmente de tipo texto.

Los sistemas de gestión documental (SGD) te permiten:

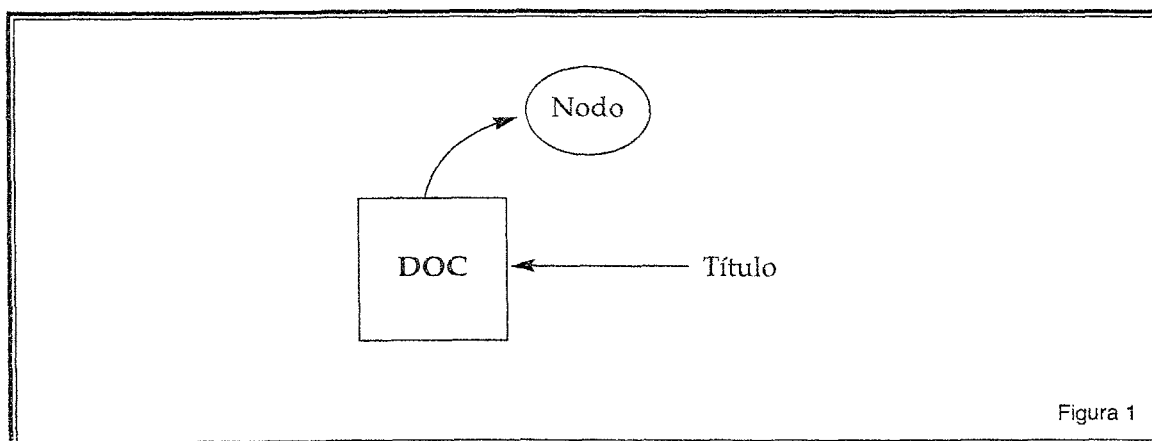
- Gestionar millones de registros, pudiéndolos recuperar en pocos segundos.
- Compartir documentos con otras oficinas, colaboradores...
- Enviar documentos a través de fax o e-mail de forma directa.
- Establecer métodos seguros de distribución de la documentación.



• ETAPAS EN LA GESTIÓN DE DATOS:

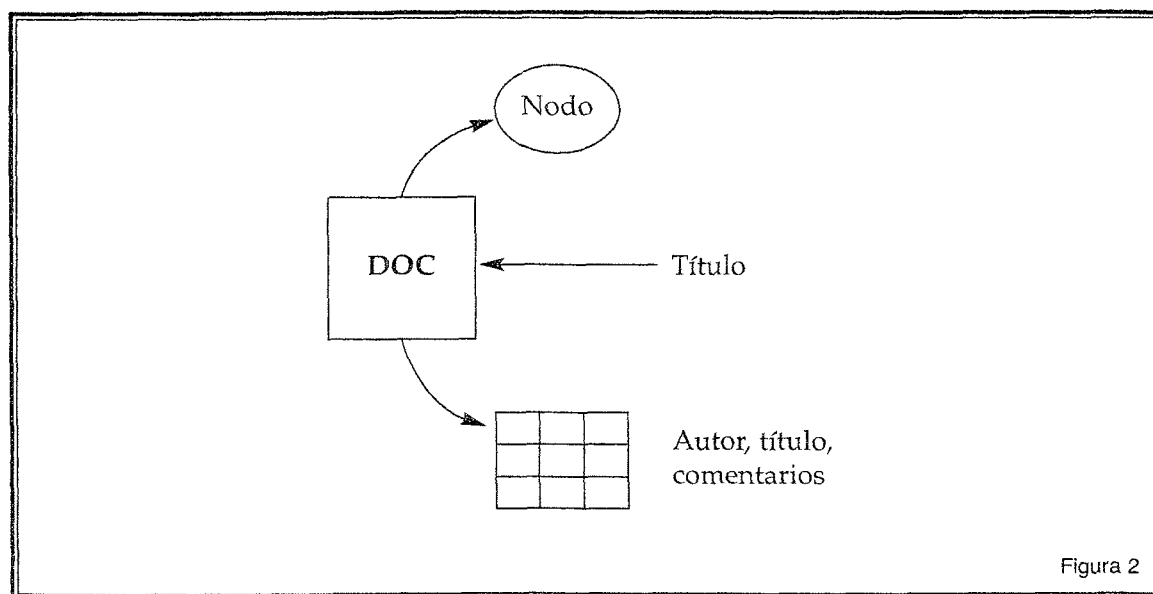
1. Monádica.

La información se refería con un único identificador o clave que determinaba unívocamente el documento. En esta etapa no se hace referencia a ninguna otra información para obtener lo que se quiere. La información conseguida podía no ser lo que se quería.



2. Estructural.

Cada documento tiene unos datos extra estructurados, que permiten ampliar la capacidad de búsqueda, almacenamiento y selección. De esta forma se va ajustando más a lo que el usuario necesita.



### 3. Contextual.

También se tiene en cuenta el contexto o entorno de la consulta. Cada persona, según sus circunstancias, puede pretender obtener distintos aspectos (vistas) de una misma búsqueda. Por ejemplo: libro de JAVA para un principiante o un experto. A ambos les interesan cosas distintas.

### 4. Cognitiva.

El sistema se adapta totalmente a la necesidad de conocimiento de las personas.

Es difícil definir la necesidad de información. Por ejemplo: quiero ver apartamentos en San Sebastián. El sistema puede interpretar que puede ser debido a que se quiere realizar unas vacaciones allí con lo que la información que te presenta son: apartamentos, trenes, actividades...

## 2. CAPTURA Y ARCHIVO ELECTRÓNICO DE DOCUMENTOS.

Se refiere a los sistemas digitales que existen para el tratamiento de la información, tanto desde su captura y procesamiento como en su salida o explotación.

La edición electrónica se realiza sobre un formato digital de la información. Los documentos originales que se pretende archivar pueden tener:

- Forma digital, por lo que directamente darían lugar a los documentos digitales.
- Forma analógica, por lo que requerirían de una conversión analógica/digital para dar lugar a los documentos digitalizados.

Hoy en día los documentos que queremos incorporar a la base de datos pueden ser de distintos tipos:

- Sonido: existen distintos formatos estandarizados:
  - .wav: es típico de Microsoft Windows. Genera ficheros muy grandes, difíciles de portar, mucha calidad (16 bits de calidad por sonido, 44 KHz).
  - .midi: tabla de ondas, se envían instrucciones para la tarjeta de sonido donde se va a escuchar. Son de poco peso, es decir, poco tamaño y de poca calidad.
  - .mp3: es similar a wav, eliminando las frecuencias fuera del rango (20 Hz, 16 KHz), las menos audibles por el hombre. Es pequeña y de calidad aceptable.
  - .mod: mezcla midi y wav, integrando ambas. No está estandarizado para Windows, es más conocido en el mundo UNIX.
  - .rmp o .ra: son de Real Audio, de calidad media pero necesitan un plugin de Real Audio para ejecutarse.

- Imagen:
  - .gif: son obligatorias si se le quiere dar movimiento a la imagen. Si varía poco de color tiene mucha compresión. Es exacto al original.
  - .jpeg, .jpg: de cada celda se obtiene un color representante, se sustituyen los elementos de ésta por el representante. Cuanto más pequeña la celda más calidad.
- Vídeo: es difícil garantizar la eficiencia:
  - AVI.
  - RTV: Real VideoPlayer.
- Animación:
  - Flash: es el estándar de facto. Ahora es el Flash Mx.

## 2.1. CAPTURA DE LA FUENTE DE INFORMACIÓN.

Si la información que vamos a almacenar es analógica, esto es proviene de una fuente, debemos realizar un proceso de conversión de esa información a digital para poder utilizarla, sino no haría falta.

Entre los distintos métodos para capturar la fuente nos encontramos:

- Escaneado.

Uno de los métodos más empleados para su captura suele ser el escáner. Es la forma más empleada para obtener documentos digitalizados a partir de señales analógicas (fotos, mapas, texto impreso, gráficos...). Los parámetros más significativos son:

- Resolución de la imagen escaneada R.
- Profundidad del píxel PP.
- Tamaño de la imagen T.
- Volumen de información  $VI = (R^2 \times T \times PP)/8.000$  Kbytes.

- Transformación.

Si la información origen proviene de un procesador de textos o de una hoja de cálculo se convierte a imagen (formato TIFF), creándose una imagen inalterable del documento electrónico original.

- Importación.

También pueden importarse los ficheros gráficos, de audio o vídeo. En este caso se mantiene el formato original y para visualizarlos se requiere de un programa compatible con ese formato.

## 2.2. ALMACENAMIENTO DE LA INFORMACIÓN.

Una vez que los archivos se han capturado se han de almacenar de una forma apropiada en elementos que sean perdurables, abiertos (no sean propietarios) y compatibles. Ejemplos de éstos pueden ser: soportes CD-ROM, DVD, discos magneto-ópticos...

Tal y como hemos indicado, una de las posibilidades ofrecidas para almacenar documentos se encuentra en los discos ópticos. Brevemente indicamos algunas de ellas:

- Discos ópticos configurables por el usuario:
  - WORM (Write Once, Read Many)-(CD-W): se pueden grabar una única vez, tienen alta potencia de escritura.
  - WMRA (Write Many, Read Always) = EDOD (Erasable Digital Optical Discs) = magnetoópticos: permiten escribir, borrar y leer datos de la misma forma que un disco duro magnético.
- Discos ópticos configurables por el fabricante:
  - CD-ROM: son dispositivos que se han grabado una única vez, generalmente en un proceso industrial donde se hacen múltiples copias de tal forma que el usuario únicamente puede leer.
  - DVD: similar al CD-ROM pero con la diferencia en la longitud de onda del láser (650 nm).

Para el caso de los sistemas documentales está claro que se necesita una gestión electrónica de todos los documentos que se almacenan. Estos sistemas discos ópticos.

Entre las muchas ventajas que supone una gestión electrónica de los documentos se encuentran:

- Reducen el volumen de almacenamiento.
- Fortalecen la seguridad del acceso a la información.
- Gran variedad de posibles búsquedas, éstas se adaptan a la necesidad de conocimiento.
- Misma validez que el papel.
- Seguridad.

Un ejemplo de sistemas que utilizan almacenamiento óptico son los COLD (Computer Output to Laser Disk).

## 2.3. SALIDAS DE LA INFORMACIÓN.

Dependiendo del uso que se le quiera dar a esa información podemos encontrarnos:

- Si queremos utilizar la información almacenada a analógica:
  - RIP (pasan del mapa de bits a formato analógico).
  - Impresoras → dpi (puntos/pulgada) para impresión.
  - Litografía → lpi (líneas/pulgada).
  - Filmadoras.
  - spi (muestras/pulgada) para imágenes digitales.
  - ppi (pixels/pulgada) para presentación en pantalla.
- Explotación directa de la información digital:
  - Grabación en soportes CD-ROM, DVD.
  - Formatos PDF.
  - Lenguaje XML.

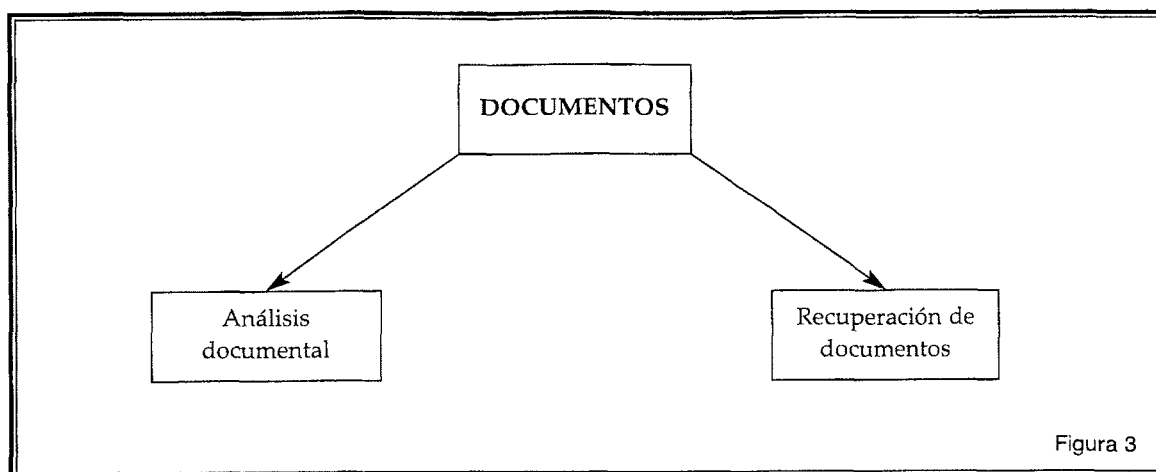
### 3. ORGANIZACIÓN FUNCIONAL DE UN SISTEMA DOCUMÁTICO.

La RI (recuperación de la información) consiste en recuperar la información que un usuario requiere mediante un conjunto de consultas a los documentos contenidos en la base de datos. Estas consultas suelen estar realizadas en un lenguaje de interrogación. Su resultado debe ser exacto, exhaustivo, preciso, oportuno, íntegro y significativo.

Un documento es un objeto de datos, que puede ser texto, imágenes multimedia, fotografías, vídeo animado, etc. Respecto a éste, el sistema de recuperación de la información ha de ser capaz de:

- Introducir nuevos documentos en la base de datos.
- Modificar los documentos de la base de datos.
- Eliminar los documentos de la base de datos.
- Búsqueda de documentos en la base de datos. Búsqueda que se realiza a través de metainformación (información imperfecta que tiene un factor de distorsión inherente al proceso) del documento. Esto supone la existencia de algún método para localizar el documento.
- Recuperación y presentación del documento al usuario.

En el apartado anterior ya habíamos realizado un procesamiento electrónico del documento. Una vez realizado éste, existen dos procesos que hay que realizar para obtener un Sistema de Recuperación de Información.



### 3.1. ANÁLISIS DOCUMENTAL.

El análisis documental consiste en seleccionar las ideas más relevantes de un documento para que luego sea posible su recuperación en el siguiente proceso. Este análisis puede tomar la forma de un resumen, un índice alfabético de materias o códigos.

#### • FASES DEL ANÁLISIS DOCUMENTAL.

##### 1. Análisis formal (nivel de asiento).

Intenta extraer los elementos característicos de un documento que lo distinguen típicamente de los demás: tipo, autor, título, editorial, fecha, número de páginas, idioma original, etc. Se realiza en dos pasos:

- Catalogación. Diseñar los puntos de acceso que los documentos han de tener en el catálogo para que puedan ser recuperados. A cada documento se le asigna un identificador único. Los documentos se estructuran en campos.
- Descripción documental. Se describe el documento.

##### 2. Análisis de contenido.

Se describe aquello sobre lo que trata el documento. Consta de tres partes:

1. Clasificación. Permite ordenar el conocimiento indicando la rama del conocimiento a la que pertenece.
2. Indización (nivel de indización). Analiza los conceptos que puede representar el contenido: patrones o palabras de igual frecuencia, lugar..., para extraer los conceptos y los traduce al lenguaje documental. A este nivel es cuando se realiza:
  - Eliminación de términos no indizables a través del uso de la lista de palabras vacías.
  - Puede asignar una ponderación a los términos basada en la frecuencia en la que aparecen en documentos.

- Reducción de las palabras a su forma raíz (corte de palabras).
- Búsqueda de la relación entre los términos introduciéndolos en un tesauro.

La extracción de información relevante se puede llevar a cabo a partir de OCR (Optical Character Recognition), una vez que los documentos se escanearon en la fase de captura de la información. Existen dos métodos para el reconocimiento de caracteres:

- Comparación matricial: compara los caracteres con plantillas estándar, que tienen un conjunto de patrones o fonts. Primero se determina el tipo de letra, se construye una matriz de la marca de tinta y la comparan con la máscara de este tipo de letra.
- Métodos basados en análisis de patrones o extracción de características: cada algoritmo analiza las características de cada letra. Contiene reglas gramaticales. Es más completo, sofisticado, flexible, preciso y lento que el anterior. También consume más recursos. Por ejemplo: OMNIPAGE v.11 de Calera y LEXIFIER de Xerox.

Los términos con los que se ha indizado el documento han de ser los mismos que se usen para recuperar la información, por lo que se suele trabajar en lenguaje documental. Esta información se almacena en una base de datos de extractos.

3. Condensación (nivel de resumen o abstract). Realiza un resumen del texto en lenguaje natural para que cuando el usuario esté realizando una consulta vea a través del resumen si es el documento que quiere recuperar.

Tanto la fase de análisis formal como la de análisis de contenidos son realizadas por el módulo de extracción. Éste almacenará los extractos y el documento donde sea preceptivo.

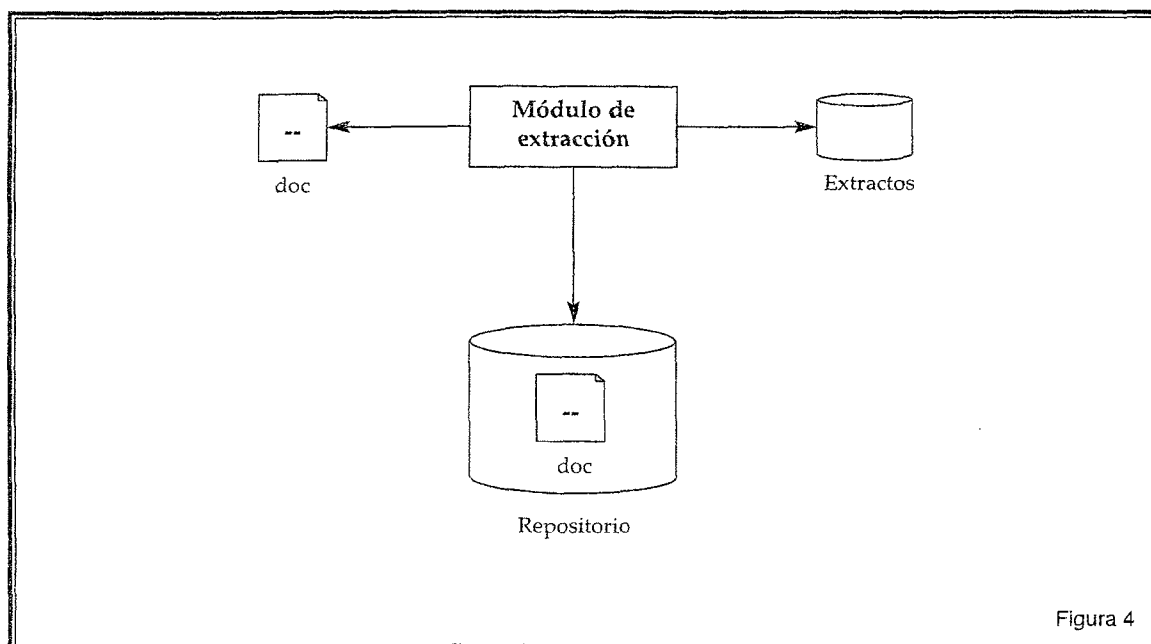


Figura 4

El documento íntegro se almacena en el repositorio de información. Puede ser en forma de BLOBS (binary large objects) en una BD o en forma de ficheros normales como en el sistema TAMINO.

- ELEMENTOS UTILIZADOS POR EL MÓDULO DE EXTRACCIÓN.

Existe una serie de elementos que van a servir de ayuda en el proceso de extracción:

- Lista de palabras vacías: son las palabras que no aportan ningún significado tales como artículos, preposiciones... y que además no nos sirven para diferenciar un documento de otro.
- Diccionarios: sirven para encontrar palabras relacionadas. Existen distintos tipos de diccionarios de términos, tanto de sinónimos, antónimos, etc.
- Tesauro: cada término puede estar relacionado con otros términos siguiendo distintos criterios tales como la jerarquía, semántica, etc. Esto es tesauro, una lista de los términos que se relacionan con uno dado. Esta relación nos servirá de ayuda para la localización.

### 3.2. ALMACENAMIENTO EN LA BASE DE DATOS DOCUMENTAL.

En el apartado anterior hemos visto que la última operación será almacenar los resultados en la base de datos. Como lo que estamos tratando son documentos entonces utilizaremos una base de datos documental.

- CARACTERÍSTICAS DE UNA BASE DE DATOS DOCUMENTAL

- Gestionan información no estructurada (suelen ser documentos).
- Gestionan un gran volumen de información, con su referencia o resumen para poderla localizar (título, autor...).
- La administración es similar a la de las bases de datos relacionales.

- DIFERENCIAS ENTRE UNA BASE DE DATOS DOCUMENTAL (BDD) Y UNA BASE DE DATOS RELACIONAL (BDR).

- Objetos que almacena: para el caso de una BDD almacenaría cualquier información clasificable, por ejemplo texto, vídeo, etc.; mientras que si estamos en una BDR lo que almacena son tablas.
- Tipo de recuperación: para el caso de una BDD existe una cierta incertidumbre respecto a la respuesta que se va a producir (probabilística) mientras que en el caso de las BDR es exacta, cada pregunta tiene su respuesta asociada (determinística).
- Criterio de éxito: para el caso de las BDD el criterio se rige por el grado de satisfacción del cliente respecto a la consulta que ha realizado mientras que en la BDR se basa en la corrección y exactitud con la que se haya producido la respuesta.
- Tiempo de respuesta: el tiempo que el sistema tarda en dar una respuesta en el caso de las BDD va a depender de las decisiones y acciones del usuario respecto de la búsqueda. Imaginar que se quiere buscar hoteles en Andalucía, se obtiene como respuesta un grupo de enlaces, la res-

puesta será más o menos rápida si el enlace que se escoge es el que nos conviene y da fin a la consulta o si por el contrario el primero no nos conviene y tenemos que ir mirando todos los enlaces hasta llegar al que se desea. Para el caso de las BDR el tiempo de respuesta depende únicamente del soporte físico y la perfección en la búsqueda.

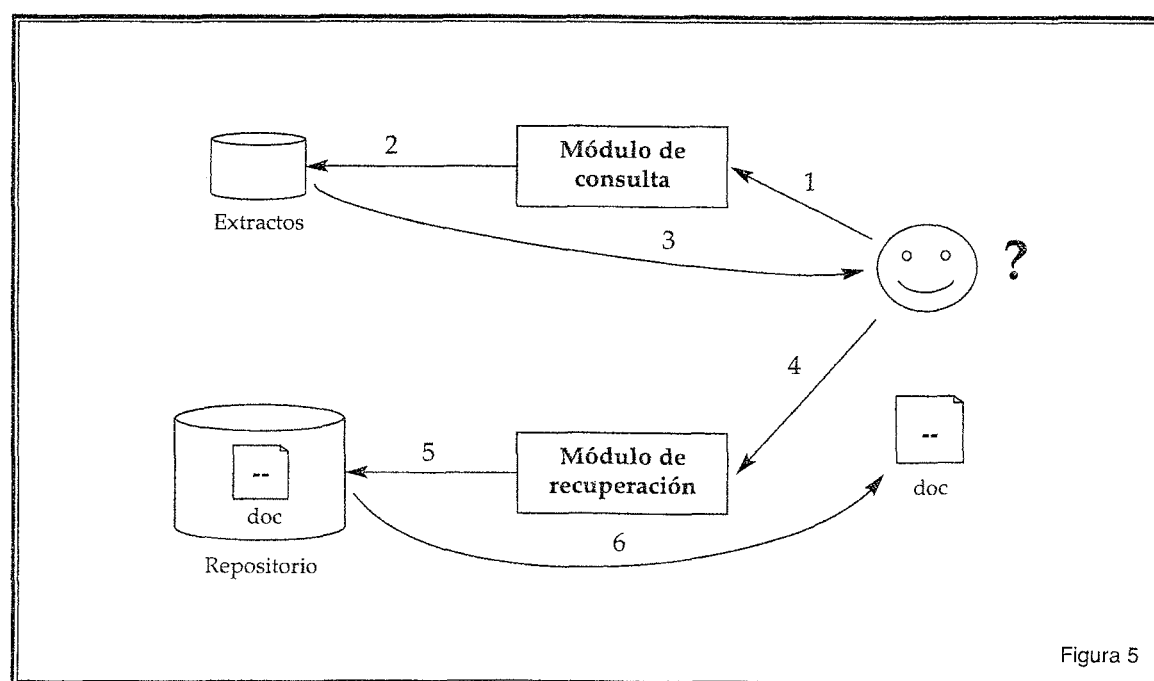
- Registros: la información se almacena en registros variables y flexibles para el caso de las BDD ya que no todos los objetos tienen el mismo tamaño. En el caso de las BDR el tamaño de los registros es fijo (debido a las estructuras de tablas) y han de ser pequeños para que sean manejables.
- Tamaño de las bases de datos: para el caso de las BDD el tamaño varía desde grande a muy grande. El hecho de tener información en distintos formatos hace que ésta sea muy grande. En muchos casos se requiere disponer de los programas a partir de los cuales esa información es visible, véase alguna imagen que requiera un visualizador... Para el caso de las BDR el tamaño va a ser intermedio.

### 3.3. CONSULTA Y RECUPERACIÓN DE LA INFORMACIÓN.

Una vez que se tiene la base de datos cargada con la información que necesita y dispuesta para su funcionamiento el usuario puede comenzar a realizar sus peticiones. El primer paso para esto es expresar su petición en un lenguaje que el sistema entienda, es decir, expresar la petición en un lenguaje de consulta. En el dibujo y en las líneas siguientes se describe el proceso completo desde que el usuario realiza la petición del documento hasta que el sistema le suministra el documento deseado.

Tal y como se muestra en la FIGURA 5, podremos dividir el proceso de consulta y recuperación en los siguientes pasos:

1. Necesidad del usuario. En este apartado, el usuario muestra qué es lo que desea conseguir del sistema. Lo expresa en un lenguaje que el módulo de consulta puede interpretar. Este módulo de consulta no es otra cosa que un interfaz entre el usuario y los metadatos (información extra que hemos almacenado en la BD de extractos).



2. Petición de extractos. El módulo de consulta acude a una base de datos, que contiene los resúmenes de los documentos que están relacionados con la consulta.
3. Exposición de extractos relacionados. El sistema le muestra un breve resumen de los documentos que están relacionados con su consulta para que el usuario escoja el que crea que más se parece a su petición.
4. Selección del documento. El usuario selecciona uno de los resúmenes listados con la pretensión de recuperar el documento que lo contiene. Esta información va hacia el módulo de recuperación.
5. Solicitud del documento. El módulo de recuperación solicita a la base de datos el documento que ha seleccionado finalmente el usuario.
6. Recuperación del documento. El sistema le muestra el documento al usuario.

Una vez que el usuario ha terminado con la consulta ésta puede almacenarse para usos posteriores, crearse perfiles de búsqueda para el supuesto de que en un futuro volviera a realizar peticiones similares.

#### • LENGUAJES UTILIZADOS EN UNA BDD.

Desde el lenguaje que utiliza el usuario para describir su petición hasta el lenguaje que utiliza el sistema para preguntarle a la base de datos se produce una pérdida de información, debida a la característica intrínseca de cada lenguaje.

- Lenguaje natural. Consiste en expresar la necesidad de conocimiento en palabras: «Quiero ir de vacaciones a SS». El usuario debe saber y expresar la necesidad de información que requiere. El problema de expresar las peticiones de esta forma reside en la ambigüedad del propio lenguaje.
- Lenguaje documental. Es el lenguaje de consulta del sistema. Esto tiene su dificultad.
- Lenguaje de interrogación de la BD. Es el lenguaje que se utiliza para interactuar con la BD. Por ejemplo SQL, esto tiene la dificultad de traducir a ese lenguaje la necesidad de información que se requiere.

#### • ORGANIZACIÓN EN FUNCIÓN DE LA BÚSQUEDA.

La información extraída puede ser:

- Índice o asiento: son los identificadores básicos de la información: autor, título, fechas, número de páginas...
- Descriptores o signatura (nivel de indización): palabras clave que identifican el documento de forma habitual. Este resultado es la base del diálogo hombre-sistema para recuperar y almacenar los documentos.
- Resúmenes o abstract: incluye un resumen del texto íntegro antes de acceder al documento.

### 3.4. TÉCNICAS Y LENGUAJES ESPECÍFICOS DE INTERROGACIÓN.

Los principales modelos de recuperación de información de la base de datos serán:

- Sistemas booleanos:

- El usuario ha de realizar la consulta a través de sentencias que van unidas mediante operadores lógicos tales como AND, OR y NOT. Las búsquedas van a ser en sus correspondientes términos de indización o palabras clave y los operadores asociados a ella.
- Las consultas que sean imprecisas van a generar mucho ruido, es decir, recuperación de textos no deseados, mientras que una consulta demasiado precisa puede ignorar textos que son relevantes.
- El índice de retorno IR está en torno al 0.5 y el de precisión también.
- Es el tipo de prueba con reintento. Se va a repetir hasta llegar al documento deseado. Una vez construida la consulta se podrá almacenar el perfil de la búsqueda para un posible uso posterior.
- Se usa sobre todo cuando hay que gestionar grandes volúmenes.
- Este modelo requiere un buen conocimiento del álgebra booleana.

Sobre este modelo han surgido algunas extensiones que se recogen bajo el modelo booleano extendido.

- Técnica de índices invertidos:

Surgió orientada al perfeccionamiento de las búsquedas mediante un acceso no booleano.

- Se crea un fichero auxiliar o índice durante el proceso de catalogación que contiene los términos clave (fichero de términos) o apuntadores a los documentos (fichero de apuntadores) a los documentos donde aparecen dichos términos.
- A esta técnica se le llama indización.
- El proceso de recuperación booleano se acelera, ya que basta con aplicar los operadores OR, AND y NOT a las referencias contenidas en los índices.
- Es ampliamente utilizada.

- Lenguaje natural:

- Se expresa la consulta en un lenguaje corriente, sin necesidad de recurrir al álgebra de boole.
- Se alcanzan resultados similares a los de acceso booleano.
- Hace uso de índices invertidos y ficheros clave. Se comparan las cadenas introducidas contra los índices invertidos para seleccionar las referencias comunes.

- Indexación y recuperación automáticas vectoriales:

Los ficheros inversos o índices se crean a partir de la indexación. Estos ficheros se crean durante el momento de la catalogación. Hay dos tipos:

- Indexación por palabras: son ficheros que guardan todas las palabras diferentes que se hallan en los documentos con una indicación sobre el documento en el que se encuentran.

Así los ficheros índices son una lista ordenada de palabras con la posición de éstas en cada documento. De esta forma se sabe para una palabra determinada cuántas veces aparece y en qué documentos.

- Indexación por string de caracteres: se dirige la búsqueda a cadenas de caracteres completas. Almacena junto a cada término el número de ocasiones que aparece en un documento. Así los textos quedan representados por vectores cuyos elementos son las frecuencias de todas las claves de dicho texto. El vector tiene de longitud el número de claves distintas que haya en el índice.

- Recupera los textos por comparación entre el vector de búsqueda (el de la consulta) y el vector de referencia (el de cada documento).

- IR (índice de recuperación) = 0.5, IP (índice de precisión) = 0.5

- Lógica borrosa:

- La lógica borrosa establece diversos grados de certeza relativos al grado de relevancia de las claves y combina los valores obtenidos para clasificar los textos con arreglo a su relevancia estimada. Los grados de certeza están en un rango (0.1). Tiene como ventaja que es muy eficaz.

- Mejora la recuperación booleana.

- Métodos vectoriales y probabilísticos:

Basados en el estudio de la frecuencia de aparición de los términos del texto.

- Modelos vectoriales: almacenan el número de veces que aparecen las claves en el documento.

Cuantas más veces aparezca la clave más relevante es el documento.

- Modelos probabilísticos: reflejan la aparición del término en el documento respecto de otros documentos de la colección.

En ambos casos se construye una matriz M donde las filas son documentos y las columnas son las claves y el valor  $M_{i,j}$  es la frecuencia de aparición de una clave en el documento.

El procedimiento de recuperación consiste en localizar documentos próximos a la consulta con arreglo a una definición de proximidad adecuada.

Gracias a estos métodos se le asigna un peso a cada documento dentro del conjunto de documentos recuperados para que sean ordenados después siguiendo un cierto orden de relevancia.

- Retroalimentación:

Este mecanismo se incorpora a las consultas anteriores para mejorar el resultado final, incorporándoles un mecanismo de retroalimentación en el vector de recuperación.

El usuario afina su búsqueda: una vez realizada la primera búsqueda el usuario indica si entre los textos recuperados hay algunos relevantes o no. Así el sistema se adapta acercándose a los documentos más parecidos y alejándose de los que no lo son:  $IR = IP = 0.7$  puede llegar a 0.8.

- Normas gamma (Salton, Fox, Wu):

Una forma de obtener documentos ordenados por su relevancia es interpretar las consultas booleanas de acuerdo a esta norma.

Estudios posteriores han indicado que es más eficaz que el sistema booleano.

- Sistemas expertos:

Los diccionarios representan conocimientos adquiridos y estructurados antes de producirse las consultas. Se pueden aplicar técnicas de sistemas expertos basados en reglas para recuperar la información. Se asignan grados de confianza a las relaciones supuestas y se crean reglas o normas de aplicación en función de los tópicos de las consultas. Aplicables en BD pequeñas consiguiéndose un alto índice de precisión y retorno.

#### 4. PARÁMETROS BÁSICOS DE EVALUACIÓN DE RESULTADOS OBTENIDOS.

Evidentemente nos gustaría saber si las búsquedas en un sistema son rápidas y eficientes o si, por el contrario, son tan lentas que el usuario no está satisfecho, con lo que pondría en peligro todo el sistema. Por ejemplo: si se intenta acceder a una página web de un supermercado para hacer la compra del día, y el cliente se ha tirado dos horas para acceder a los productos deseados eso puede provocar que la próxima vez el cliente se lo piense dos veces antes de utilizar el sistema. Para eso existen ciertos parámetros que se describen a continuación:

- Eficacia en la ejecución.

Es el tiempo que tarda el sistema en realizar una operación. Es un parámetro importante porque un largo tiempo en recuperar un documento interfiere en la utilidad del sistema. Por ejemplo: si el usuario espera en la web, por un documento recuperado, durante mucho tiempo se suele impacientar y desiste de su empeño respecto de esa recuperación.

- Eficiencia en el almacenamiento.

Se corresponde con el número de bytes que se necesita para almacenar los datos.

- Efectividad de la recuperación.

Se basa en la relevancia de los documentos. Éste es un parámetro un tanto subjetivo, ya que una misma consulta con una misma respuesta puede ser más o menos efectiva en distintas personas. Por ejemplo: un novato en java y un experto si realizan la misma consulta, «Quiero ejemplos de java» como su objetivo no es el mismo y el sistema recupera la misma información no estarán igual de satisfechos con el resultado.

Respecto a la efectividad de la recuperación, según Lancaster podríamos encontrar otros parámetros, algunos de ellos todavía vigentes: cobertura, exhaustividad, precisión, tiempo de respuesta, esfuerzo del usuario y formato.

Nosotros vamos a considerar dos parámetros importantes:

- Índice de retorno (IR) o tasa de exhaustividad: mide la cantidad de información relevante obtenida en una búsqueda dada, sobre el número total de documentos relevantes. Excepto para tests realizados sobre pequeñas colecciones, este denominador es generalmente desconocido y debe estimarse por muestreo o por otros métodos.

$$IR = \frac{A}{A + C} \approx [0.6, 0.8]$$

- Tasa o índice de precisión (IP): mide la calidad de la información devuelta. Esto es el número de documentos relevantes recuperados respecto del número total de documentos extraídos.

$$IP = \frac{A}{A + B} \approx [0.2, 0.8]$$

En esta tabla se muestra el código utilizado para designar la información extraída o no extraída, relevante y no relevante a nuestra consulta para utilizarlo posteriormente.

	DOCUMENTOS PERTINENTES	DOCUMENTOS NO PERTINENTES
Extraídos .....	A	B
No extraídos .....	C	D



