MODULO I.

INTRODUCCIÓN.

- **Búsqueda y recuperación de información: perspectivas teóricas.**

    - **El proceso de búsqueda de información.**

El estudio de los sistemas debería comenzar según Robertson por plantear la cuestión de dónde situar los límites de ese sistema, de hecho las diferencias que existen a la hora de emplazar estos límites han dado origen a las dos corrientes teóricas fundamentales en esta área:

- Perspectiva centrada en el sistema.
- Perspectiva centrada en el usuario.

El proceso de búsqueda y recuperación de información ha sido analizado como un proceso de comunicación informal, y como un proceso de resolución de problemas en el que intervienen varios actores:

- Un usuario − persona que busca información porque tiene necesidad informativa.
- Sistema de búsqueda.

    - Documentos
    - Representación de esos documentos
    - Organización de las representaciones de esos documentos
    - Técnicas de recuperación de la información

- Conjunto de resultados que se obtienen del sistema − y que debe satisfacer la necesidad de información que tenía el usuario.

Todos estos elementos influyen en el resultado final del proceso de búsqueda, e interactúan entre ellos, sin embargo, tradicionalmente se ha centrado la investigación casi
exclusivamente en uno de esos factores: en el sistema de búsqueda.

Esta situación ha cambiado a partir de los años 80 con la incorporación de modelos cognitivos al análisis del proceso de recuperación de información. Este cambio ha provocado que la investigación se preocupe por cuestiones:

- ¿Cómo surge la necesidad de información?
- ¿Qué problemas encuentran los usuarios para definir esa necesidad de información?
- ¿Qué estrategias se desarrollan para satisfacerla?

Desde esta perspectiva integradora se han desarrollado varios modelos teóricos del proceso de búsqueda que van a influir directamente en el diseño de los sistemas de recuperación de la información:

- SARACEVIC.

De 1996. Integra siete factores distintos que abarcan el propio sistema de recuperación de información. Recursos informativos que almacena y recursos informáticos que gestiona. Y una interfaz que dialoga con el usuario. El conocimiento, situación, entorno son factores relacionados con el usuario.

- Modelo estratificado de interacción.

Llamado por Robertson porque se desarrolla en tres niveles:

- De superficie.
- Cognitivo.
- De situación.

En el nivel de superficie el usuario interactúa con el sistema mediante un interfaz, usando órdenes o preguntas ! enunciados de búsqueda. Que representan el planteamiento de un problema. En este nivel el sistema responde con información: documentos o metainformación, que en teoría debería satisfacer la necesidad del usuario.

En el nivel cognitivo el usuario interactúa con la respuesta del sistema para evaluar su utilidad en relación con el problema inicial.

Por último en el nivel de situación el usuario interactúa con la situación o problema que ha producido la necesidad de información, intentando aplicar el resultado de la búsqueda a la resolución del problema.

- INGWERSEN.

De 1992. Integra aportaciones de la obra de Belkin y de Vickery de principios de los años 80 (1982 − 1985). El proceso de búsqueda se divide en diez entapas agrupadas en tres fases:

- 1ª fase ! Pre−búsqueda.

Hay 3 etapas:

1. El usuario tiene un problema de información que necesita solucionar.

- El usuario reconoce que su estado de conocimiento es inadecuado para solucionar el problema ! BELKIN (ASK ! estado anómalo de conocimiento).
- El usuario intente resolverlo con la búsqueda de información en un sistema.

- 2ª fase ! Búsqueda propiamente dicha.

Hay 5 etapas. Se inicia con una etapa de interacción con un intermediario: bibliotecario de referencia o documentalista, o con un sistema informático. Esta interacción tiene como objetivo la definición de la necesidad de información y el conocimiento del usuario de las características del sistema.

- Formulación de la estrategia de búsqueda en los términos aceptados por el sistema.
- Actividad de búsqueda.
- Evaluación inicial de los resultados.
- Reformulación del problema, de la estrategia y del enunciado de la búsqueda.
- Retroalimentación ! en función de los primeros resultados obtenidos, el usuario y el sistema retroalimentan o reformulan la búsqueda.

- 3ª fase: Post−búsqueda ! nivel de situación en Saracevic.

- Se evalúa la información recuperada en función de las necesidades.
- Se utiliza la información.

La situación habitual en que un usuario interactúa con un sistema de recuperación de información es que ese usuario se acerque a ese sistema sin tener muy claro lo que quiere, o como mínimo, sin tener muy claro cómo expresarlo. Estas condiciones deberían tenerse en cuenta en el diseño de los SRI (Sistemas de recuperación de información); sin embargo, eso no es lo más frecuente. Lo más habitual es que los sistemas estén pensados para resolver lo que los ingleses denominan **query**, búsquedas analíticas (equiparación de la expresión de búsqueda con el índice). Búsquedas que se han planteado conociendo mínimamente la terminología y la sintaxis que utiliza el sistema. Para comparar la expresión de búsqueda del usuario con los términos del índice de búsqueda.

Para plantear esa búsqueda analítica el usuario debe tener un modelo mental del sistema con el que interactúa: sintaxis Esta situación es muy compleja.

Se han desarrollado sistemas para otro tipo de búsqueda más dinámica, de carácter exploratorio: **browsing**, la información se ojea, se visualiza y se selecciona. El usuario no necesita formular un enunciado de búsqueda, lo que hace es seleccionar información de entre las opciones que el sistema le presenta. Descarga el esfuerzo cognitivo que tiene que realizar el usuario porque sólo tiene que reconocer la información que resolverá su problema y no tiene que reescribirla como en el modelo anterior.

- **Perspectivas centradas en el sistema y en el usuario.**

En función de la importancia que se dé a los diferentes factores que intervienen en el proceso de búsqueda o, en palabras de Robertson: en función de donde se pongan los límitestradicionalmente se han distinguido dos corrientes sobre recuperación de información.

- Enfoque centrado en el sistema.

O enfoque tradicional, algorítmico, fisicalista. El SRI se limita al mecanismo que permite almacenar y recuperar la información. Hay una entrada en el sistema que es la estrategia de búsqueda, **input**, y los documentos recuperados, **output**.

Se centra en los problemas relacionados con la representación de los documentos y con la equiparación de esta representación con las búsquedas de los usuarios.

Su objetivo es mejorar las técnicas de representación y equiparación.

La relevancia es la medida en que un documento o varios documentos recuperados se ajustan al objetivo de la búsqueda. Se utiliza para evaluar los sistemas de búsqueda de información.

Para esta perspectiva es una relevancia objetiva, temática. Lo que mide es el ajuste entre el tema del documento y el de la búsqueda pensando que hay una relación objetiva entre ellos.

Sus autores son Salton, el más importante, Spark Jones, Robertson, Rijsbergen.

Trec Text Retrieval Conference − Se prueba la eficacia de los SRI.

- Perspectiva centrada en el usuario o cognitiva.

Trata de abarcar todos los elementos que influyen en el proceso de búsqueda con la incorporación del usuario y del entorno en el que se produce esa búsqueda.

Si se quieren diseñar sistemas eficaces es necesario saber las características contextuales e individuales que influyen en la formalización de las necesidades de información y en la valoración de los resultados obtenidos.

Su objetivo es adecuar los SRI a las características de los usuarios finales y a las necesidades de un proceso interactivo y que evoluciona.

Se interesa por analizar temas como las condiciones en las que surgen la necesidad de información y como influyen esas condiciones en el proceso de búsqueda, la influencia del entorno.

Investiga la influencia de las características individuales de los usuarios: sexo, edad, nivel profesional, conocimientos de la materia, experiencia con informática en el proceso de búsqueda y en la eficacia que se obtiene de los sistemas. Analiza cómo se desarrolla ese proceso: la interacción

Es frecuente que se la llame centrada en el usuario o cognitiva. Ahora se la reconoce que hay una perspectiva centrada en el usuario en la que hay una orientación cognitiva que se preocupa por el último de los factores: el desarrollo de la búsqueda.

Según Ingwersen: tiene como objetivo entender cómo las categorías de los modelos conceptuales de los usuarios afectan al proceso de búsqueda y se modifican durante el mismo.

La relevancia es contextual o de situación, medida subjetiva que trata de valorar la utilidad del documento o de los documentos recuperados para el usuario que plantea la búsqueda.

Sus autores son Belkin, Ellis, Borgman o Bates.

- **Definición de sistemas de recuperación de información.**

- **Recuperación de información y recuperación de datos.**

Expresión que tiene muchos términos considerados equivalentes:

- recuperación automatizada de información,
- almacenamiento y recuperación de información,
- sistemas de almacenamiento, conservación y recuperación de información (SACRI),
- recuperación de documentos que ya no es tan frecuente como en los 80,
- recuperación de texto − está vinculada a que los SRI, en un principio sólo recuperaban información textual, ahora con imágenes, sonido

- texto libre.
- texto completo.

Las definiciones se establecen a partir de la diferenciación de recuperación de información y recuperación de datos. Se hace en función de cuatro aspectos (de Blair) sintetizando las ocho variables de Rijsbergen:

- Tipo de respuesta que recibe el usuario.

- Recuperación de datos.

La interrogación típica es específica. El sistema responde con una respuesta directa, recuperación directa, que debe ofrecer la información real. Quiero saber .

- Recuperación de información.

Recuperación indirecta que proporciona o dirige a un conjunto de documentos que probablemente contendrán lo que quiere el usuario. La pregunta típica es general, no específica. Quiero saber acerca de

- Relación entre la petición formal y la satisfacción del demandante de información.

- Recuperación de datos.

Relación necesaria entre una petición bien construida y la respuesta correcta. Relación determinista.

- Recuperación de información.

Relación probabilística entre una petición bien construida y una respuesta satisfactoria. Aunque el enunciado de búsqueda esté bien construido la respuesta puede ser o no satisfactoria.

- Criterio del éxito de la recuperación.

- Recuperación de datos.

Corrección de la respuesta. Evaluación objetiva: ¿Contesta el sistema correctamente a la pregunta del demandante?

- Recuperación de información.

Depende de la utilidad de la información recuperada para el usuario. Evaluación subjetiva: ¿Satisface el sistema la necesidad del demandante?

- Velocidad de la recuperación con éxito.

- Recuperación de datos.

Depende principalmente de la velocidad física de acceso del sistema que se está usando.

- Recuperación de información.

Depende principalmente del número de decisiones lógicas que se deben tomar al formular una búsqueda.

Se hace esto porque los SRI surgen por las debilidades de los sistemas de recuperación de datos para la gestión de datos.

- **Caracterización de los SRI.**

Los SRI también llamados sistemas de gestión documental son uno de los modelos de los sistemas de información automatizada, concretamente dentro de esos sistemas, son el modelo que está orientado a la gestión de información textual desestructurada.

Fueron diseñados para superar las limitaciones que presentaban los SGBD relacionales para trabajar con este tipo de información.

Los imperativos del modelo de bases de datos relacional, especialmente, la no admisión de grupos repetitivos o atributos multivaluados o las exigencias de un diseño muy normalizado obligan a la descomposición de un objeto de información en un conjunto de relaciones. El modelo relacional exige que se definan unos campos que van a definirse con su longitud, tipo de información de manera que representar esa información supone crear una estructura compleja de relaciones entre la información que contienen.

La existencia de esos campos repetitivos haría que se repitiese una entrada de registro por cada uno de los

valores de campo repetidos, ya que este sistema para recuperar la información trabaja con cada uno de esos campos. Se descompone el objeto en varias tablas que se relacionan entre sí, se unen luego con la ? en SQL.

Frente a estas limitaciones del modelo relacional los SRI son concebidos para representar documentos o conjuntos de información que no tienen ningún tipo de estructura o algún tipo de estructura formal en la que el valor de cada elemento puede ser variable tanto en el número de veces que aparece como en su longitud. También están diseñados para facilitar el acceso a los documentos por muchas vías, sobre todo por cada una de las palabras que contienen.

Podemos definir estos SRI como sistemas automatizados cuyo objetivo es la recuperación de documentos que contengan información probablemente relevante para satisfacer las necesidades del usuario expresadas con una estrategia de búsqueda. Todo SRI realiza dos tareas básicas por lo menos:

- La representación

Es el proceso por el que el sistema transforma un documento ya guardado o el enunciado de búsqueda del usuario en entradas de índice o puntos de acceso que pretenden representar la información del documento y también la necesidad de información del usuario.

- La búsqueda

La búsqueda en un SRI es el proceso por el que el sistema examina las representaciones de los documentos y las compara con la representación de la consulta. Su finalidad es determinar qué representaciones de los documentos coinciden con la de la búsqueda o son más similares.

- **Características de los Sistemas de Recuperación de Información.**

    - **Organización de la información documental en los SRI.**

Parten de principio de que la información que se tiene que tratar en el sistema informático se procesa a partir de documentos entendidos estos como secuencias más o menos extensas de caracteres que se agrupan formando: palabras ! frases ! párrafos o campos de información ! documentos.

A pesar de que en principio estos sistemas se crearon para trabajar con información desestructurada, hay que entenderlo relativamente ya que los documentos con los que trabajan los SRI tienen siempre una estructura formal interna.

Inicialmente los SRI se diseñaron para trabajar con información bibliográfica. Se pretendía gestionar bases de datos de referencias bibliográficas que sustituían a los documentos originales de forma que si se usan referencias bibliográficas tanto la representación y la búsqueda que hace el SRI se hace sobre documentos secundarios. Con el paso del tiempo se abarató el coste del almacenamiento de la información electrónica y se empezó a considerar normal el almacenamiento y recuperación del texto completo de los documentos. Ahora las dos posibilidades conviven y los SRI trabajan tanto con representaciones, metainformación, como con documentos originales, texto completo.

La metainformación puede ser registros bibliográficos normalizados con el formato MARC, o registros de una base de datos, representados en HTML como el Dublin Core.

Aunque se trabaje con el texto completo la información no está totalmente desestructurada porque cualquier documento electrónico tiene una estructura por un lenguaje de marcado, ésta puede estar orientada a la representación formal de los documentos o también orientada a una descripción formal del contenido de los documentos, en este caso se indicará cual es el autor, título y la información se ordenará jerárquicamente.

Lenguajes como SGML, HTML, y XML lo utilizan.

Esta estructura cuando se trabaja con representaciones de los documentos permite identificar distintas partes del registro que almacena en conjunto la información. Hay varias opciones para almacenar estas estructuras de registro variable, se almacena una etiqueta que lo identifica y, además, hay unas marcas que definen el principio y el final del registro, la longitud de etiqueta de campo es fija, también se puede incluir información sobre la longitud del campo. Una alternativa es guardar toda la información sobre la estructura del registro en un área diferenciada de datos que se suele llamar directorio, esto se hace con los registros en formato MARC. Mientras mayor información se tenga sobre la estructura del registro del documento más precisión habrá en la recuperación.

- **Estructuras de datos para el acceso a la información.**

Los SGBD documentales mantienen unas estructuras determinadas para posibilitar el acceso a la información que contiene el fichero de texto de forma no secuencial. Habitualmente suelen ofrecer diferentes métodos de indización y acceso a la información. Esta flexibilidad da muchas posibilidades a la hora de parametrizar los índices que van a facilitar el acceso al sistema. Ej.: formato MARC campo materias:

650 08 $a Mujeres $x Situación Social $z Andalucía

Se pueden generar índices de materias, alfabéticos, de palabras clave, de subcampos. En el índice de materias se pueden cambiar el orden de campos, o no, y también por palabras clave para acceder a la información. Es el propio centro de información o bibliotecario el que decide los campos a indicar y los índices que se van a realizar.

Todos estos sistemas elaboran índices o estructuras de acceso a los documentos a partir de las palabras que contienen. Este proceso se denomina **indización automática por medio de palabras claves**, lo que hace que se genere un fichero inverso que es un índice que contiene la relación alfabética de todos los términos contenidos en los campos indizables del fichero texto más la información posicional correspondiente. Cuando el usuario introduce un término de búsqueda, el sistema de recuperación accede al fichero inverso para localizar ese término y recoger la información posicional.

Como el tamaño de los registros del fichero de texto es variable, esta información posicional del fichero inverso remite al fichero índice de texto que es donde tenemos la localización correcta del documento en el fichero. La información posicional del término que se específica en el fichero inverso condiciona las posibilidades de recuperación, si en esa información sólo se incluye información sobre el documento sólo podemos indicar al sistema que los términos de búsqueda utilizada estén en un documento. Si queremos que se puedan realizar búsquedas más precisas hay que incluir más información sobre la posición del término.

Ej.: Hacer búsquedas dentro de un párrafo. En la posición indicaremos además del documento el número de párrafo o de campo en el que aparece ese documento. Si queremos realizar búsqueda de adyacencia tenemos que decir el número de palabras que representa esa palabra en el documento.

Respecto a los términos que se incluyen en el índice, en principio cualquier palabra que aparezca en un documento almacenado en el sistema sirve para representar su contenido, puede servir como punto de acceso o entrada de un índice. En la práctica se realiza un proceso de selección para eliminar términos no significativos y reducir el ruido documental en la recuperación. Este proceso de selección puede ser más o menos exhaustivo y puede afectar simplemente a la eliminación de palabras vacías, pero también puede afectar a la normalización de formas flexionadas y derivadas.

Ej.: Normalización de las formas singular/plural, masculino/femenino, diferentes tiempos verbales, normalización de sufijos y prefijos

Cuando realizamos una normalización se pueden adoptar dos soluciones para gestionar las búsquedas:

- Someter al enunciado de búsqueda al mismo proceso de normalización al que se le ha sometido a los términos del índice.
- Mantener un fichero diccionario en el que se conserva información sobre todos los términos que puede utilizar el usuario y desde este fichero diccionario remite al inverso en el que sólo se almacena información sobre los documentos que contienen el término normalizado. En el fichero inverso aparece información sobre el número de veces que aparece el término y el número de documentos en el que aparece.

- **Técnicas de equiparación.**

Los índices de estructura de datos de un SRI permiten realizar operaciones de búsqueda mediante técnicas que comparan o equiparan los enunciados que han utilizado los usuarios con los términos almacenados en los índices del sistema. Estas técnicas permiten recuperar o realizar búsquedas de documentos que contengan uno o varios términos sin importar en qué campo o en qué arte del texto. Búsqueda de documentos que contengan uno o más términos en un campo o subcampo concreto.

La clasificación básica de las técnicas de equiparación empleada se realiza en función del grado de coincidencia entre búsqueda y términos del índice que exigen esas técnicas para que se recuperen documentos. Por ello se diferencian:

- Técnicas de equiparación exacta − los documentos sólo se recuperan cuando cumplen todas las condiciones del enunciado de búsqueda. Si utilizamos el operador 1 se produce mucho silencio y si utilizamos el operador 0 se produce mucho ruido.

- Técnicas de equiparación parcial − los documentos se recuperan en función de su similitud con el planteamiento de búsqueda y no de la coincidencia total de los descriptores con los términos que aparecen en el documento. Eliminan la necesidad de utilizar operadores voléanos. Para facilitar la selección de la información, los documentos recuperados se ordenan en función de su relevancia respecto a la búsqueda, relevancia determinada por el grado de coincidencia con el enunciado de búsqueda.

- **Interacción con el usuario**.

Al mismo tiempo que se mejora la técnica de equiparación, la investigación sobre la experiencia de usuarios reales demostraba que para estos incluso era cuestionable el sistema de búsqueda por equiparación, con frecuencia el problema es que no tienen los recursos necesarios para convertir su carencia de información en una pregunta formal a un sistema de recuperación.

Para mejorar la interacción en cualquier tipo de búsqueda los SRI han ido incorporando diferentes herramientas que mejora tanto la equiparación como la búsqueda exploratoria, una de estas herramientas son los sistemas que permiten la corrección de errores tipográficos y ortográficos del enunciado de búsqueda. Los estudios empíricos demuestran que entre un 7−10% de las búsquedas fallan porque incluyen este tipo de errores, las técnicas que se han desarrollado para corregirlos permiten mejorar la equiparación o coincidencia con los índices. El resto de las mejoras están orientadas a facilitar las búsquedas exploratorias.

- Herramientas utilizadas para facilitar búsquedas exploratorias

    ◊ *Visualización de los índices alfabéticos y sistemáticos,* lo que permite al usuario hacerse una idea del contenido del sistema. También deberían llevar a ese usuario desde el término que ha utilizado hasta el término admitido como punto de acceso y

también deberá permitir encontrar términos de búsqueda en los que previamente no había pensado.

- ◊ *Visualización de los registros o documentos que contiene el sistema*. La mayor parte de los sistemas ofrecen visualizaciones en diferentes formatos con el objetivo de que el usuario tenga información suficiente sobre el documento para facilitar los juicios de relevancia y la potencial selección de nuevos términos de búsqueda.
- ◊ *Enlaces hipertextuales entre registros*. Enlaces que permiten visualizar registros que comparten información de los campos que son puntos de acceso normalmente.
- ◊ *La modificación de las búsquedas mediante técnicas de retroalimentación por relevancia* (relevante feedback) permiten recuperar automáticamente documentaos similares al que un usuario ha considerado relevante.
- ◊ *Presentación de la información a los usuarios* mediante interfaces gráficos y metáforas de la biblioteca.

MODULO 2. REPRESENTACIÓN DE LA INFORMACIÓN TEXTUAL.

- **Conceptos básicos de la representación para la recuperación de información.**

- **Indización por asignación.**

Se emplean palabras o expresiones elegidas por el indicador y asignadas a los documentos para representar su contenido. Estas palabras o expresiones son descriptores libres o, con mayor frecuencia, descriptores de un lenguaje documental (LD). Sirven como puente para enlazar el vocabulario usado por los autores de los documentos y el de los usuarios. La utilización de los LD supone una normalización de los términos a usar, que es formal o morfológica como semántica. Esta normalización tiene como finalidad que cada concepto que aparece en los documentos esté representado por un solo término, además de normalizar la forma hay que controlar la sinonimia y también trata de que cada término represente a un solo concepto, para esto se controla la polisemia.

Los LD establecen relaciones semánticas. Esta normalización que podría ser una ventaja para la recuperación es también una limitación del vocabulario de entrada: de los términos que puede usar el usuario para la recuperación. Esta limitación que se podría contrarrestar fácilmente si se explotaran las relaciones de equivalencia supone un obstáculo para el uso de los LD.

- **Indización por extracción.**

La idea de ampliar el vocabulario de entrada es la que se encuentra en este tipo de indización. Se extrae automáticamente las palabras del texto de los documentos y usarlas como punto de acceso o entradas de índice para representar su contenido. Permite en principio usar todos los términos empleados por los autores de los documentos para caracterizar su contenido.

Gil Leiva lo divide en dos grupos:

- *Métodos no lingüísticos* − usan análisis estadísticos basados en la frecuencia de aparición de los términos en los documentos. En estos análisis se entiende como término: toda cadena de caracteres diferente que aparecen en el documento sin tener en cuenta su significado, diferencias en la representación. Pero tiene unos problemas:

- No tienen capacidad para resolver los problemas de ambigüedad terminológica como la sinonimia o la homonimia, que sí se resolvían en los LD. Estos es importante sobre todo porque son métodos que trabajan con frecuencias de aparición considerando que cada término representa un concepto distinto, penalizan a aquellos autores con mayor riqueza de vocabulario.

- No son capaces de reconocer y trabajar con recursos lingüísticos que permiten dar cohesión a la lengua escrita. Uno de estos mecanismos es la **anáfora**: repetición sistemática de un elemento a lo largo del discurso usando para ello pronombres y elipsis o sinónimos también.

- En su forma más simple, no contabilizan la aparición de términos compuestos, de expresiones, y por el contrario sí contabilizan todas las formas flexionadas de una misma raíz.

- *Métodos lingüísticos* − menos desarrollados. Análisis morfológico, sintáctico o semántico del documento. De estas tres posibilidades la más desarrollada es el análisis morfológico, sólo están en algunos sistemas experimentales. Divididos en dos grupos: lingüísticos y morfológicos, pero no son excluyentes entre ellos.

- **Extracción automatizada de los términos de indización.**

    - **Indización automatizada** − Van Slype define la indización automatizada como una operación que consiste en que el ordenador reconoce los términos o elementos que aparecen en el título, en el resumen o en el texto completo de los documentos y los incorpora a su fichero de búsqueda como características que lo representa.

    - **Análisis léxico** − En recuperación de información, análisis léxico es el proceso por el que se convierte un flujo de caracteres de entrada en un flujo de palabras o elementos. Considerando esos elementos son grupos de caracteres con significado. Este análisis léxico se hace sobre el texto de los documentos como sobre los enunciados de búsqueda de los usuarios. ¿Qué constituye un elemento? A primera vista cada elemento es cada una de las palabras o secuencias de caracteres entre dos separadores (espacios en blanco, puntuación).

    - Caracteres numéricos − no suelen ser buenos términos de indización, por eso no suelen incluirse como elementos. Sin embargo, ciertas secuencias de números pueden ser importantes en determinados entornos (bases de datos legales, bases de datos históricas) Aparecen con frecuencia en expresiones alfanuméricas (bases de datos de documentos técnicos) por lo que muchos sistemas lo consideran como elementos. También son las fechas.
    - Palabras con guiones − ¿hay que unirlas o separarlas? Se piensa que separarlas ayuda a un uso consistente porque así son igual a las que aparecen en los documentos cuando no están unidas. Pero hay que tener en cuenta que a veces los guiones forman parte de una palabra y que también se usan para partir una palabra al final de una línea sin que sean dos palabras.
    - Otros signos de puntuación − puntos, barras. Los puntos pueden formar parte de siglas o acrónimos, las barras para separar elementos de una misma palabra. Diferencia de mayúsculas o minúsculas igual.

No hay dificultad técnica para resolver estos problemas pero hay que tener cuidado con ellos cuando haya que hacer la política de análisis léxico del sistema.

    - **Filtrado de términos** − Esta información en bruto que se obtiene del análisis léxico pasa luego por un filtrado de términos. La función principal es eliminar los elementos no significativos. Desde el principio de los trabajos en recuperación de información se ha reconocido que muchas de las palabras más usadas son desaconsejables como términos de índice ya que cualquier búsqueda que se haga por ellos, recuperará casi todos los documentos de la base de datos porque el valor de discriminación en estos términos que aparecen mucho es muy bajo. El origen de los estudios y propuestas para el filtrado y selección de términos en la indización automática lo forman los trabajos del psicolingüísta Zipf.

Se desarrollaron en los 40. Zipf en su primera ley dice que si ordenamos un conjunto de palabras diferentes y

que son de un mismo corpus documental las ordenamos de manera que decrezca su nivel/frecuencia de aparición y después multiplicamos cada frecuencia por el rango que ocupa esa palabra, el resultado es un valor próximo a una constante.

La conclusión es que existe cierta relación entre la frecuencia y la utilización de las palabras y la importancia que éstas tienen de cara a la representación del contenido de los documentos a los que están asociadas o de los que se han extraído.

Luhn profundizó en los trabajos de Zipf de cara a la recuperación de información, llegó a decir que la significación de cualquier texto está depositada en las palabras que tienen frecuencias intermedias, por tanto las que tienen frecuencia muy baja o muy alta carecen de significado fuera de un corpus documental. Como consecuencia se puede establecer un valor de corte para eliminar aquellas palabras que no serían representativas del contenido.

A pesar de esto, no desarrolla ningún procedimiento para establecer los límites de los valores de corte, esto se debería realizar en función del contenido de cada una de las bases de datos.

Los trabajos de Luhn y Zipf están en la base del procedimiento más habitual para filtrar y eliminar palabras no significativas en cualquier SRI, esto es utilizar un antidiccionario o lista de palabras vacías. En estos hay términos que por su frecuencia de aparición no sirven como características de los documentos.

Los sistemas comerciales suelen ser muy conservadores e incluyen muy pocas palabras vacías: determinantes y proposiciones, palabras que no tienen significado léxico. Convendría añadir a estas palabras aquellas que por la especialización de la base de datos aparecen también con mucha frecuencia: haciendo un estudio de este tipo y establecer un valor de corte.

Este proceso de filtrado de elementos se puede complementar con otros:

- Listado de palabras vacías
- Reducción de términos a la raíz (lematización, lexematización, stemming)
- Ponderación de los términos − valorar y eliminar los de poco valor.

- **Lematización.**

La mayoría de los SRI incluyen algún mecanismo que permite reducir el número de términos de indización utilizando algún control morfológico o de las formas flexionadas de las palabras.

Este tipo de mecanismo se utiliza considerando que aquellos términos con misma raíz tienen también un significado equivalente. La lematización es para reducir considerablemente el fichero de búsqueda sin que esto implique una pérdida importante de información. Eliminar en mayor o menor medida las variantes de un mismo lexema producidas por flexión: singular, plural, masc., fem., los tiempos verbales; y también las formas producidas por derivación: sufijos, prefijos Estos algoritmos se aplican en el proceso de recuperación, de indización y en ambos.

- **Conflación** − Es el proceso que permite reunir todas las variantes morfológicas de una misma raíz. Tiene dos ventajas:

- Limita notablemente el tamaño de los ficheros de búsqueda, aproximadamente un 50 %.
- Simplifica el trabajo del usuario que no tendrá que introducir todas las formas derivadas de una misma raíz cuando haga una consulta, ni con truncamientos.

A pesar de estas ventajas, la lematización no elimina dos problemas: la polisemia ni la sinonimia. Muchos de

ellos no extraen la raíz de las palabras sino que las cortan y unen aquellas que comparten caracteres. Sólo se hace una reducción a la raíz con los análisis morfológicos. Al reducir la palabra a la raíz aumenta la exhaustividad en la recuperación, pero también se reduce la precisión. El mayor problema de estos algoritmos es saber cuál es el nivel óptimo de lematización: de eliminación de caracteres de un término. Hay dos formas en que un lematizador puede ser inexacto:

- **Hiperlematización** − cuando elimina excesivos caracteres de un término. Provocando que en la conflación se unan términos que realmente no estén relacionados semánticamente. En la recuperación habrá documentos no relevantes.

- **Hipolematización** − se eliminan pocos caracteres de lo debido, provocando que en la conflación no se unan términos que sí están relacionados, y en la recuperación habrá silencio documental.

Frakes (1992) clasifica los algoritmos de lematización:

- **Algoritmos de eliminación de afijos**

Eliminan los sufijos o prefijos y dejan un lexema. La mayoría de los estudios y las aplicaciones de reducción de formas variadas y flexionadas por algoritmos de eliminación de afijos se han hecho con el inglés, y el intento de aplicarlos a otras lenguas no ha dado buenos resultados. Los mejores por tanto son los creados para el inglés. Cualquiera de estos algoritmos se basa en un conjunto de terminaciones, un conjunto de condiciones que debe cumplir el lexema que resulte de la reducción y un conjunto de reglas de sustitución.

El más sencillo es el que permite controlar las formas singular y plural, como el de Harman en 1991, es muy simple. Los más importantes para el inglés son el de Porter y el de Lawgli.

- Porter − elimina los plurales y las formas −ed e −ing (students/student). También elimina algunas terminaciones como las formas −ic, −full, −ness (electronic/electron). Asimismo, elimina la −e si la palabra tiene más de dos sílabas (violence/violenc).

Problemas de los algoritmos:

- Difícilmente se pueden aplicar a diferentes lenguas del inglés, se usan poco en Internet, y para el castellano los de más éxito son los que eliminan las formas sing./plural.
- No extraen la forma canónica mediante un análisis morfológico por lo que no funcionan con lenguas que tienen variantes flexionadas más complejas que el inglés.
- Lingüísticamente algunas veces que los afijos aportan más información sobre la palabra que la raíz.

- **Algoritmos de búsqueda en tabla.**

Los términos extraídos de los documentos y sus correspondientes lexemas están almacenados en una tabla por la que se realiza la lematización.

Es el método más simple. Consiste en almacenar en una tabla todos los términos de índice, considerando como términos solamente los lexemas.

Problemas:

◊ No existen estas tablas elaboradas ni para el inglés ni para el castellano.
◊ Aunque existiesen muchos términos de la base de datos no estarían representados

porque son dependientes dela base de datos especializada y no pertenecen al vocabulario estándar y necesitarían uan lematización diferente.

◊ La sobrecarga de almacenamiento. Auqneu en algunas condiciones valdría la pena cambiar sistema de almacenamiento por el de gestión.

♦ **Algoritmos de variedad de sucesores.**

Usan como base para la lematización las frecuencias de las secuencias de letras en un corpus textual o documental.

Se basa en el cálculo de la longitud de los prefijos que mejor admitan la expansión mediante la implantación de sufijos. Se trata de un método análogo empleado en los análisis de lingüística estructural que intenta delimitar los límites de palabras y los morfemas basándose en la distribución de fonemas en un cuerpo de pronunciaciones. También se puede usar con letras y con un corpus de texto en vez de pronunciaciones.

Formal − el método se define:

Sea x una cadena de caracteres de longitud n

Sea xi un prefijo de longitud i de x

Sea D el corpus textual

Sea Dxi el subconjunto de D que comparten aquella secuencia de caracteres que es xi

La variedad de sucesores de xi, denotada por Dxi, es el número de letrsa distintas que ocupan la posición i + 1 en las palabras Dxi.

Ej.:

Corpus doctal: ábaco, ámbar, accidente, ambiente, arco

xi = a xi3 = amb

Dxi = todas Dxi = ambar, ambiente

Sxi = 4 (b, c, m, r) Sxi = 2 (a, i) ! Variedad de sucesores

La variedad de sucesores de una cadena de caracteres es el número de caracteres diferentes que pueden seguir a esa cadena en las palabras de un corpus textual.

A

B C M R

B

A I

R E

N

T

E

A 4 (b, c, m, r)

AMB 2 (a, i)

AMBA 1 (r)

AMBAR 1 (blanco)

Cuando se lleva a cabo esto es un corpus amplio. La variedad de sucesos de un término disminuye a medida que se le añaden caracteres hasta alcanzar el límite de un segmento, en ese momento, la variedad de sucesores vuelve a aumentar y esta información puede usarse para identificar los lexemas que identificarán esas palabras.

Una vez que se han comparado los sucesos de una palabra se puede usar esa información para segmentarla. Hay distintos métodos para segmentarlas:

- Valor de corte − cuando se aplica este método se selecciona un valor de corte para todas las variedades de sucesores y se identifica el límite de un segmento cada vez que se alcanza ese valor.
- Picos y Valles − se hace el corte del segmento en aquellos caracteres cuya variedad de sucesores excede al que lo precede y también excede a la del segmento que lo sigue.
- Palabra completa − el corte se hace después de un segmento si éste es una palabra completa del corpus.

Ej. Según el método de picos y valles: actual, amplio, anunciar, archivo, archivero, arqueología, átomo

A

C M N R T

T C Q

H U

I

V

E O

R

O

A − 5 (c, m, n, t, r)

AR − 2 (c, q)

ARC − 1 (h)

ARCH − 1 (i)

ARCHI − 1 (v)

ARCHIV − 2 (e, o)

ARCHIVE − 1 (r)

ARCHIVER − 1 (o)

ARCHIVERO − blanco

Dxi = ARCHIV (segmento de caracteres).

Ej.2: balda, beneficio biología, bibliografía, biblioteca, bibliotecario, blasón, byte

B − 5 (a, e, i, l, y) BIBLIOTEC − 1 (a)

BI − 2 (o, b) BIBLIOTECA − 1 (r)

BIB − 1 (l) BIBLIOTECAR − 1 (i)

BIBL − 1 (i) BIBLIOTECARI − 1 (o)

BIBLI − 1 (o) BIBLIOTECARIO − 1 (blanco)

BIBLIO − 2 (g, t) Según los picos y valles: Dxi = BIBLIO

BIBLIOT − 1 (e) Según palabra completa: Dxi

BIBLIOTE − 1 (c)

Hay veces que aunque usemos un solo método para establecer los segmentos pueden salir varios segmentos que sirvan para representarla. En este caso se ha decretado que el segmento más corto (12 o menos palabras) se considera a ese segmento como lexema; si aparece en más de doce se consideraría el segundo segmento como lexema. Esta condición parte de que si se basa en más de doce palabras será un prefijo que no tiene que ver con la raíz.

- **N−gramas.**

Unen los términos en función del número de digramas o n−gramas que comparten. Este método que no es exactamente uno de lematización se basa en el número de secuencias de caracteres compartidos o dos palabras o expresiones, se usa para seleccionar una misma forma que represente a varios términos, también para vincular expresiones.

La secuencia de caracteres que se selecciona de una palabra, puede estar formada: por dos caracteres, digrama, tres caracteres, trigrama, o por n−caracteres

Permite asociar términos, calculando medias de asociación de términos basándose en los digramas, trigramas, n−gramas únicos que comparten.

El primer paso para usarlo es establecer el tipo de n−gramas que se van a utilizar. Para cada una de las

palabras del texto se harían tantas divisiones en secuencias de dos caracteres, si usamos digramas, como permita esa expresión.

Ej.:

Actual = ac ct tu ua al (5)

Actualidad = ac ct tu ua al li id da ad (9)

No hay repeticiones en los digramas.

Luego se calcula la **medida de asociación**, para ello se usa el:

- Coeficiente de inicio − DICE. El número de digramas o n−gramas únicos compartidos multiplicado por dos, se divide entre la suma de los digramas únicos de cada expresión. Esto da una cifra que oscila entre 0 y 1. La asociación entre el par de términos será mayor cuanto más se acerque ese resultado a 1.

Con el ejemplo de antes se obtendría:

S = 2 * 5 / 5 + 9 = 0,71.

Esta medida de similitud se usa para todos los pares de términos de la base de datos, luego se asocian los términos usando alguno de los métodos de Cluster: que permitan agrupar aquellos términos más parecidos que se considerarán como un único grupo de cara a la indización.

Las asociaciones son próximas a 0 y si se establece que para que los términos pertenezcan al mismo grupo la similitud tiene que ser superior al 0,6. Las asociaciones que se dan son correctas, es extraño lo contrario.

Ej.:

- Estadística = **es st** ta ad di is st **ti ic ca** (9)
- Estadísticamente = **es st** ta ad di is st **ti ic ca** am me en nt te (14)
- Estratificar = **es st** tr ra at **ti** if fi **ic ca** ar (11)

S (1, 2) = = 0,78

S (1, 3) = = 0,5

- **Ponderación de los términos de indización.**

  - **Algoritmos de ponderación.**

Este método consiste en asignar un valor a los términos que se han extraído de los documentos. Estos términos pueden estar ya lematizados o no. El valor que se da a los términos basándose en el principio de Luhn está relacionado con su frecuencia de aparición en el texto. Podemos considerar la ponderación como una operación de filtrado porque se puede determinar que aquellos términos que no alcancen determinado nivel no se tendrán en cuenta para representar el contenido de ese documento. Se han diseñado muchos algoritmos de ponderación.

  - Algoritmos de frecuencia del término en el documento.

El método más fácil para ponderar un término es en controlar el número de veces que está en el documento y asignarle ese valor. Para neutralizar este problema (no son buenos los términos de indicación que aparecen mucho o poco) se han desarrollado sistemas que permiten normalizar en cierta manera la frecuencia de aparición. El objetivo es moderar el efecto de los términos que aparecen demasiado, y compensar la longitud del documento.

Pij = ! **algoritmo de Harman**

El peso de un término de un documento está directamente relacionado con la frecuencia de aparición en ese documento y es inversamente proporcional a la longitud del documento: número de términos que lo forman.

Pij = Peso del término i en el documento j

Freqij = frecuencia del término i en el documento j

Longj = número de términos en el documento j

	Term 1	Term 2	Term 3
Doc 1	3	1	4
Doc 2	1	0	3
Doc 3	2	6	3

P (3, 2) = = 0,15

- ♦ Algoritmo de frecuencia inversa.

Aunque tiene en cuenta la longitud del documento no es propiamente un algoritmo de frecuencia.

La frecuencia de aparición de un término en un conjunto de documentos debe tenerse en cuenta también para valorar la importancia del término de cara a la indización. Esta idea es la que subyace en todas las funciones de ponderación de frecuencia inversa que expresan el peso de un término relacionándolo de forma inversamente proporcional con su frecuencia de aparición en la base de datos.

Una de las primeras fórmulas de ponderación basada en este principio fue la de Sparck Jones que ha servido de base para todos los algoritmos de frecuencia inversa (IDF) posteriores.

El peso en esta fórmula está relacionad con su frecuencia de aparición en la base de datos y con el número de documentos en la base de datos.

Pij = log 2 N / freqi + 1

Pij = Peso del término i en el documento j.

N = número de documentos en la base de datos.

Freq = número de veces que aparece en el documento.

La mayoría de los algoritmos de IDF relaciona la frecuencia de aparición del término en el documento con la frecuenta de aparición de ese mismo término en la base de datos.

Salton 1:

Pesoij = freqij * (log2 N − log2 dfi + 1).

dfi = número de documentos en los que aparece el término.

Ej. Salton con la tabla anterior (3,2)

Pij = 3 * (log2 3 − log2 3 + 1) = 3 la resta de logaritmos hace que estos se anulen.

Salton 2 − tiene en cuenta la frecuencia de aparición del término en el documento y la relaciona con el número de términos de la base de datos y con la frecuencia total de aparición del término que estamos valorando.

Pesoij = freqij *?+ (log NT − log freqi + 1)

NT = número total de términos.

Freqi = número de veces que aparece el término.

Ej.:

P = 3 + (log 24 − log 10 + 1) = 4,14

- Valor de discriminación

Usar el valor de discriminación representa una alternativa teórica a los modelos anteriores. Los descriptores se valoran por su capacidad para incrementar la media de disimilaridad/diferenciación entre las descripciones de los documentos en una base de datos. En cualquier base de datos se puede calcular la similitud de los documentos en función de las características que estos documentos comparten, los términos de indización. Los documentos serán más parecidos en la medida en que comparten más descriptores.

La consecuencia es que un término que aparezca en muchos documentos tendrá poca capacidad para discriminar entre los documentos y esto hace que tenga poca capacidad para diferenciar documentos relevantes y no relevantes. Se establecerá una medida para discriminar. Basándose en la capacidad para discriminar de algunos descriptores Salton y Yang propusieron un algoritmo de ponderación de los términos. Se basa en el cálculo de un valor de discriminación de los descriptores:

DISCRIMi = MediaSIMi − MediaSIM.

MediaSIMi = Media de similitud entre los documentos si eliminamos el término i.

MediaSIM = Media de similitud.

- Este valor puede ser positivo − indica que eliminar el término i aumenta la diferencia entre los documentos, por tanto su existencia aumenta la discriminación. Tiene valor de discriminación.
- Un valor 0 implica que es indiferente que exista o no este término para la similitud. Valor de discriminación ni positivo ni negativo.
- Un valor negativo implica que la existencia del término aumenta la similitud entre los documentos. Por tanto no tiene valor discriminatorio.

Cuando se ha calculado este valor y sólo siendo positivo se aplicaría la fórmula para la ponderación del término, asignarle un peso:

Pesoij = freqij * DISCRIMi.

Freqij = Frecuencia de término i en el documetno j.

DISCRIMi = Valor de discriminación del término i.

Hay que decidir a partir de qué peso va a ser el límite entre descriptores limitadores y los que no lo son.

- **Otras técnicas de análisis morfológico o textual.**

Se aplican a la recuperación de información como consecuencia de los malos resultados que producen los algoritmos de lematización: eliminación de afijos para normalizar los términos en aquellas lenguas que tienen flexiones más complejas que el inglés. Sólo son experimentales, y se importan de investigaciones sobre el proceso del lenguaje natural.

♦ **Análisis morfológico (Tagging).**

Se basa en un diccionario léxico que contempla todas las posibles variantes de una raíz junto con un conjunto de reglas para las flexiones y con la información de la categoría gramatical a la que corresponde.

Se obtiene un listado en el que aparece el conjunto de posibles formas canónicas del término, junto a la información asociada a cada una de las flexiones.

♦ **Análisis morfosintáctico (POST: Part Of Speech Tagging).**

Complementa al análisis morfológico, para ello se necesitan unas reglas que permitan al sistema diferenciar lo que es una oración, sintagma El sistema decide cuál de las posibles formas de un término es la que corresponde a cada caso. Esto se hace por probabilidad, o por un análisis morfosintáctico completo, normalmente por probabilidad. Sirve para desambiguar términos.

MODULO III.

MODELOS PARA LA RECUPERACIÓN DE INFORMACIÓN.

- **Modelos de recuperación de información.**

Las primeras investigaciones sobre recuperación de información que se desarrollaron en los 60 tenían un carácter totalmente práctico y que no se empezó a complementar con modelos teóricos hasta los 70.

El desarrollo de estos modelos teóricos en los que se basan las técnicas de recuperación no sólo permite comprobar la eficacia de las técnicas, sino que también facilita el análisis y racionalización que pueden ser necesarias para comprobar la eficacia de los sistemas con un conjunto real de documentos, búsquedas y juicios de relevancia.

Sparck Jones y Willet han sintetizado los modelos en cuatro:

El grado de concreción de estos modelos es muy difícil porque mientras algunos sólo establecen fundamentos teóricos, otros proporcionan información concreta sobre las tareas que tiene que desarrollar un sistema de búsqueda.

♦ **Modelo booleano.**

La lógica de Boole ha sido la base usada para desarrollar la mayoría de los sistemas de recuperación de datos y también la mayoría de los sistemas tradicionales de recuperación de información.

Se basa en una combinación de los términos de búsqueda en el momento de la recuperación usando una serie de operadores, los operadores voléanos: Y ! intersección (AND), O ! unión (OR), NO ! exclusión (NOT). También permiten usar operadores de proximidad: CERCA ! NEAR, CON ! WITH (en un mismo párrafo).

Su uso en las bases de dados con la opción de restringir la búsqueda a campos concretos: autor, título permite plantear preguntas muy sofisticadas que pueden obtener buenos resultados en la recuperación. Pero el uso adecuado de este modelo tiene unas exigencias:

- Es necesario que los usuarios formalicen su necesidad de información en un planteamiento de búsqueda concreto y preciso (**query**) para poder realizar una búsqueda analítica o por equiparación, no exploratoria.
- Es necesario entender el significado de los operadores y usarlos correctamente, algo que no está al alcance de usuarios no expertos.
- Todos los documentos recuperados tienen que cumplir todas las condiciones de la estrategia de búsqueda. **Técnicas de equiparación exactas**.

El modelo booleano aunque es el más extendido tiene numerosas limitaciones que han provocado diversas críticas:

- La lógica booleana no se adquiere por intuición, sino que requiere formación por parte del usuario. El Y lógico en la lengua es unión de términos, aquí es una intersección. El significado de los operadores es distinto al que tienen en sentido lógico.

Ej.: Información sobre medicina en Francia y España s. XVII: el usuario quiere los documentos en Francia, España y en los dos.

- El operador AND es demasiado restrictivo. Produce mucho silencio. Al revés que el operador OR que es altamente expansivo y por tanto produce ruido documental.
- Dado un par de términos sobre los que se quiere consultar la base de datos, al usuario no le es posible establecer si el término x es más importante que el término y ni en qué proporción.
- En respuesta a una pregunta, el álgebra de Boole no es capaz de mostrar los documentos en función de algún grado de relevancia. Los documentos se recuperarán si cumplen las condiciones.
-

Ej.: Residuos **OR** Contaminación **OR** Reciclaje.

AND ! sólo recupero los que tengan los tres términos.

Tiene la posibilidad de recuperar los documentos por orden de relevancia utilizando el álgebra de Boole, con una estrategia de búsqueda compleja, que combine los tres términos:

Residuos AND Contaminación AND Reciclaje

Residuos AND Contaminación NOT Reciclaje

Residuos AND Reciclaje NOT Contaminación

Reciclaje AND Contaminación NOT Residuos

Residuos NOT Contaminación NOT Reciclaje

Reciclaje NOT Residuos NOT Contaminación

Contaminación NOT Reciclaje NOT Residuos

Esta es una alternativa a los modelos de equiparación parcial. Se desarrollan modelos que permitan la equiparación y la ordenación por relevancia de los documentos.

- **Modelo espacio−vectorial.**

Es el modelo más reconocido, incluso más que el booleano. Se empieza a desarrollar en los 70, el investigador más importante en este modelo es Salton, estaba a la cabeza de un equipo de investigadores que desarrolló el SMART.

En este modelo los términos se consideran coordenados de un espacio informativo multidimensional, los términos de indización sitúan un documento y una búsqueda en un espacio que tiene n direcciones. Los términos de indización son considerados como coordenadas de un espacio informativo multidimensional. Documentos y búsquedas que se representan en un espacio, ya que documentos y búsquedas se representan como vectores en ese espacio. Cada componente del vector está representado por el peso asignado a ese término utilizando algún algoritmo de ponderación, o por el término de indización correspondiente.

Para recuperar los documentos se calcula su similitud, su proximidad en este espacio. Utilizando alguna función de similaridad. Los documentos se ordenan en función de esa similitud con la búsqueda; es decir, aparecerán primero los documentos más próximos al vector de la búsqueda.

- **Modelo probabilística.**

Se desarrolla a mediados de los 70, como alternativa al modelo espacio−vectorial. Con este modelo han trabajado distintos equipos y los dos sistemas experimentales que más lo han usado: OKAPI (para catálogos en línea) e INQUERY (para bases de datos).

Toma como punto de partida la premisa de que la función primordial de cualquier SRI es ordenar una colección de documentos en orden decreciente de probabilidad, de relevancia para una necesidad informativa. En este modelo los documentos y búsquedas se representan mediante características que son términos de ponderación. La recuperación se realiza mediante función matemática que valora la probabilidad de que un documento sea relevante para una búsqueda. La salida de los documentos es ordenada en función de esa probabilidad de relevancia.

- **Modelo cognitivo.**

Modelo totalmente teórico. El principio del desarrollo de este modelo lo estableció DeMey para quien cualquier proceso de información está mediado por un sistema de categorías de conceptos que constituyen un modelo del mundo. En la interacción del usuario con el sistema existen dos modelos del mundo:

- El del usuario.
- El que ha creado el sistema.

Para una recuperación eficaz es necesario que el usuario entienda las categorías conceptuales que componen el sistema y desarrollar sistemas que se adapten a modelos mentales que se adapten a cualquier usuario.

Se dan fundamentos teóricos para desarrollar sistemas que faciliten comprensión por parte del usuario de las categorías conceptuales que fundamentan el SRI.

- ♦ Sistemas expertos.
- ♦ Interfaces amigables.

- **Modelo espacio−vectorial.**

Cualquiera de los puntos situados en el espacio se definen por dos coordenadas x e y. Si fueran tres coordenadas A = (x, y, z). Es fundamental el orden en que aparezcan los elementos de esos vectores.

Tiene como punto de partida un conjunto de documentos al que se le ha asignado un número no específico de términos de indización que son los que van a definir el espacio en el que nos moveremos y que servirán para determinar el número de dimensiones que va a tener cada uno de los vectores que se incluyan en ese espacio, va a servir para establecer el número de elementos que hay que tener en cuenta para representar cada elemento que hay en ese espacio vectorial.

Ej.: Términos de indización:

Interacción

Automatización

Indización D1 = (1010100) T = (t1, t2, t3, t4, t5, t6, t7)

Browsing D2 = (0110010) T = (t1, t2, t3 tn)

Usuarios

Estadística

Internet

En este mismo espacio vectorial también se representan las búsquedas. Representando los documentos y la búsqueda buscar la similitud entre ellos.

Ej.: ! B1 = (0010101) Se representa presencia o ausencia de términos

Se multiplica y luego se suma

D1 = (1010100)

0 +0 + 1 + 0 + 1 + 0 + 0 = 2

Podemos usar valores binomios o valores ponderados, e incluso ponderar los resultados de la búsqueda.

Ejercicios: Recuperar como vectores la indización ponderada de los siguientes documentos de una base de datos:

Vectores:

Metadatos: 1000000

Docs. Electrónicos: 1010001

Catalogación: 1110000

Rec. Inf.: 1001000

Bibliotecas: 0100000

Automatización: 0100100

Marc 21: 0010000

Documentos: 0001001

Formación: 0001000

Indización: 0000111

Estadística: 0000110

Usuarios: 0000110

Lingüística: 0000010

Internet: 0000001

Bibliotecarios: 0001000

Búsqueda 1 (01100000000000)

Sim B1 D1 = 10 1

Sim B1 D2 = 6 3 Orden de la recuperación

Sim B1 D3 = 8 2

Sim B1 D4 = 4 4

Equiparación parcial − los resultados que produce son muchos documentos y se intenta compensar con la ordenación por relevancia y con algún sistema para limitar la búsqueda.

Búsqueda 2 (00010001110000)

Sim B2 D1 = 3

Sim B2 D4 = 22

Sim B2 D7 = 6

Búsqueda 3 (00001100010000) Automatización e indización en las bibliotecas

Sim B3 D2 = 9 2

Sim B3 D5 = 14 1 Orden de la recuperación

Sim B3 D6 = 6 4

Sim B3 D7 = 7 3

- ♦ Para calcular la similitud se usan diferentes **funciones de similaridad.**
- ♦ Producto escalar.

Sumatorio del producto de cada uno de los elementos del factor de búsqueda:

Se calcula la intersección

- ♦ Coseno.

Se calcula el coseno del ángulo que se forma entre el vector de la búsqueda y el vector del documento. Esta medida varía entre 0 y 1, y los documentos serán más similares cuanto más se acerquen a 1.

- ♦ Coeficiente de Dice.

x = vector búsqueda 2 ·

y = vector documento

- ♦ Coeficiente de Jaccard.

- **Técnicas automatizadas de clasificación: el análisis de cluster.**
- **Definición y características.**

Se pueden clasificar los documentos usando las técnicas de clustering. Se ocupan de la forma en que se agrupan los términos de indización asignados a los documentos o los documentos mismos para poner de manifiesto la relación que hay entre documentos de materias similares creando grupos con características comunes.

El punto de partida es la llamada **hipótesis cluster** que establece qué documentos estrechamente relacionados mediante estas técnicas tienden a ser relevantes para las mismas búsquedas. El objetivo de la agrupación de documentos en clusters es simplificar el acceso y la manipulación de los ficheros de las bases de datos.

En un fichero sobre el que se han realizado técnicas de clustering los documentos que pertenecen a un mismo cluster se almacenan en localizaciones adyacentes y de esta manera un único acceso permite recuperar todos los documentos de un mismo cluster. Por tanto en la recuperación los términos de la búsqueda no se comparan con todos los documentos de la base de datos para calcular la similitud sino con un representante de cada una de las clases. Si ese representante es relevante todos los documentos de ese grupo lo serán. Los representantes se denominan **centroides**.

Las técnicas de Cluster y los sistemas bibliotecarios tienen un mismo objetivo: organizar temáticamente la información almacenada.

Mientras que usando un sistema de clasificación bibliográfica los documentos se agrupan en función de una sola característica dominante, en el caso de la clasificación automatizada el objetivo se define por múltiples características que se tienen en cuenta para agruparlo a un grupo. Es **politética**: un objeto se define por múltiples características.

- ♦ Algorítmica − algoritmo formado por:
    - ◊ Función de distancia − que dice cuánto tienen que parecerse los documentos para incorporarse al grupo.
    - ◊ Regla de aglomeración que define los criterios que se deben seguir para incorporar nuevos documentos a los grupos.
        - · Genera clases sin denominar cada una de las características de los documentos, sí tiene denominación pero no el grupo que se genera.

- **Métodos para la generación de cluster.**

Basados todos ellos en el cálculo de la similitud entre pares de objetos y los métodos básicamente pueden ser de dos tipos:

- Métodos heurísticos.

Dividen un conjunto de documentos en series de subconjuntos entre los que no hay relaciones jerárquicas. Usan los parámetros que permiten controlar el proceso de creación de los grupos. El parámetro fundamental es el umbral de similitud que tiene que existir entre un documento y un grupo para que ese documento entre a formar parte.

- Número de clusters que se van a formar.
- El tamaño máximo y mínimo de esos grupos.

Se basan en el cálculo de la similitud entre pares de objetos y en la definición de representantes de cada cluster que es el que va a servir de base para las operaciones de comparación con el conjunto de documentos que se están clasificando. Normalmente ese centroide no es un ítem concreto sino que se define de manera artificial como un promedio de las características de los documentos que forman ese cluster.

- One pass − los elementos que se van a clasificar se toman en orden arbitrario. Un documento que va a pasar a ser el centroide. Cada uno de los documentos que se incorporan se compara con los grupos que ya existen y si de acuerdo a un umbral de similitud, previamente establecido, se puede incorporar a alguno de los grupos ya existentes, se añade y sino ese nuevo documento pasa a formar un nuevo grupo.

Ej.: D1 forma un nuevo grupo y es centroide

D2 se compara con D1 su similitud, si es por ejemplo 6 pasa a ese grupo y se forma un nuevo centroide (Ca).

D3 se compara a Ca y si la similitud es 8 entra y se calcula un nuevo centroide.

D4 se compara con Ca y si es 3, al no superar el umbral de similitud pasa a formar un nuevo grupo y ser centroide.

D5 se compara con ambos grupos con el que más similitud tenga, entra.

- Documento semilla − los representantes de cada cluster están previamente determinados, antes de la clasificación de los documentos ya se han elegido los centroides.

Los métodos heurísticos producen clasificaciones de manera rápida y con coste informático relativamente bajo. El inconveniente es que con grandes cantidades de documentos tienen un comportamiento bastante arbitrario y muchas veces los grupos que se generan dependen del orden en que se vayan incorporando los

documentos.

- Métodos jerárquicos.

Exigen como punto de partida el cálculo de la similitud entre todos los pares de documentos de la base de datos. La construcción de la jerarquía a partir de estas similitudes

- Por una técnica divisiva − se crearán los cluster de arriba abajo, grupo con características comunes y luego grupos más específicos.
- Por una técnica acumulativa − grupos de abajo a arriba. A partir de grupos pequeños se irán construyendo grupos más grandes.

Para crear clusters jerárquicos se usa más la técnica acumulativa que la divisiva.

- Métodos para la generación de clusters jerárquicos
  - ♦ Método de enlace sencillo o sigle link − se agrupan los documentos más similares y los nuevos documentos se incorporan al cluster tomando como medida la similitud entre el par más similar de documentos. Genera pocos grupos de gran tamaño y con límites poco definidos.
  - ♦ Método de enlace completo o complete link − se agrupan también los documentos más similares pero la medida para incorporar nuevos documentos a un grupo es la similitud entre el par menos similar de documentos.
  - ♦ Método de enlace promedio o group average − se toma como medida un promedio de la media de similitud de los documentos del grupo respecto a los documentos que se van a incorporar.

Además de determinar un umbral de similitud para que los documentos se unan al cluster, es necesario definir una regla de acumulación. Esta regla es la que permite diferenciar los métodos jerárquicos para la generación de clusters.

Aunque en diferente medida todos estos métodos jerárquicos dan lugar a grupos bien formados y estables sin importar el orden de incorporar los documentos. El problema de los métodos jerárquicos es que al exigir un cálculo previo de la similitud de todos los documentos su implementación puede resultar más costosa de los recursos informáticos que requieren.

MODULO IV.

PRESENTACIÓN DE LA INFORMACIÓN AL USUARIO Y MEJORAS PARA LA INTERACCIÓN.

- **Introducción.**

El desarrollo de la tecnología de las telecomunicaciones en las últimas décadas ha contribuido a incrementar el volumen de los recursos de información disponibles en línea y a popularizar:

- **La utilización de los SRI.**

Por una parte se ha incrementado la información en línea con la incorporación de recursos electrónicos que constituyen colecciones dinámicas enlazadas y distribuidas cuyo crecimiento es exponencial. La tecnología de las telecomunicaciones ha facilitado la interconexión de sistemas. Las redes telemáticas dan el soporte necesario para conectar sistemas que usan software diferente, que tienen bases de datos con contenido distinto y que aplican técnicas de representación y recuperación distintas.

Protocolos como Z39.50 permiten consultar desde un mismo interfaz catálogos, bases de datos y otros recursos de información que tienen características técnicas distintas, así como bases de datos distintas. Para que esta interconexión sea realmente eficaz sería necesario homogeneizar las técnicas de representación del contenido de los documentos que se usan en los distintos sistemas que se conectan.

- **El acceso a la información en línea.**

Esto significa que la recuperación de información se ha desplazado desde hace ya unos años desde el terreno de los especialistas al de los usurario finales. Este nuevo grupo de usuarios es más numeroso y heterogéneo cada vez, tanto en las habilidades con que se enfrenta a los SRI como en lo que demanda de estos SRI. En muchos aspectos la tecnología parece haberse desarrollado más deprisa que nuestro conocimiento a las tareas a las que se aplica y que la evolución de nuestras capacidades para adaptarnos a esos avances tecnológicos.

La tecnología no facilita en la misma medida herramientas para realizar búsquedas más eficaces, o para manipular los resultados obtenidos. Se han mejorado mucho las técnicas para equiparar y representar las búsquedas, pero se nos facilita poca información en los SRI sobre qué se puede encontrar, cómo realizar una búsqueda o es muy complejo Este desajuste se debe a que el modelo conceptual de los SRI apenas ha cambiado en estos años y para interactuar con los SRI los usuarios siguen necesitando manejar las mismas habilidades y los mismos conocimientos que poco después de la aparición de los sistemas en línea, los OPAC.

En 1986 Borgman estableció que el usuario necesitaba tres tipos de conocimiento cuando interactuaba con un sistema de recuperación de información. En el año 2000 sigue pensando lo mismo en lo referente a los sistemas avanzados de recuperación de información.

- Conocimiento conceptual − permite transformar su necesidad de información en una estrategia de búsqueda adecuada gracias a un correcto modelo mental del sistema.
- Conocimiento semántico − necesario para poder implementar ese planteamiento de búsqueda en un sistema concreto. En qué campos se puede buscar, modificar búsqueda, tipo de búsqueda, operadores voléanos
- Habilidades técnicas − para manejar un ordenador, funcionamiento teclado, ratón, pantalla.

Actualmente los usuarios en su mayor parte tienen las habilidades técnicas necesarias para usar con mayor o menor pericia un ordenador. Pero este número de usuarios se reduce considerablemente cuando se trata de los conocimientos conceptuales, sintácticos y semánticos que son necesarios para realizar una búsqueda satisfactoria. Esta situación se complica más cuando se trata de búsquedas por materias en las que el usuario tiene que expresar aquello que desconoce.

El problema radica en que se sigue usando como base para el diseño de estos sistemas un paradigma anticuado basado en una sola respuesta, el conjunto de documentos, a preguntas o búsquedas que se supone que van a ser precisas y específicas olvidando aquellas necesidades de información que son más ambiguas o más difíciles de definir.

En este tipo de sistemas no es posible en principio ver y seleccionar físicamente una fuente de información. El usuario tiene que usar alguna representación de esa fuente de información para recuperarla. Para representarla es necesario formular un enunciado de búsqueda que es el que se va a usar para compararlo con la representación de los documentos que hay en el sistema ! **búsqueda analítica**.

El sistema sólo responde cuando se ha planteado la búsqueda. Los resultados sobre el comportamiento de búsqueda de los usuarios finales parecen cuestionar la eficacia de las estrategias analíticas, especialmente en aquellos casos en que se trata de búsqueda de materias, que siempre son peor definidas. Los usuarios no tienen un correcto modelo mental del sistema, no saben usar la lógica booleana, problema para encontrar términos adecuados o alternativos, y necesidades de información que pueden cambiar durante el proceso de búsqueda.

En la práctica es casi imposible que un usuario formule una estrategia que satisfaga con un solo resultado su necesidad de información. Esta dificultad ha estimulado la idea de que es necesario fomentar el planteamiento de **búsquedas exploratorias** en las que el enunciado inicial, que es muy genérico, se va modificando durante la interacción con el sistema; mientras, el usuario va conociendo las características y estructura del sistema con el que trabaja, y también va perfilando su necesidad de información.

La mejora de la eficacia de los SRI pasa por el desarrollo de modelos alternativos que suponen una reducción del esfuerzo cognitivo que realiza el usuario durante la búsqueda trasladando el esfuerzo al sistema. Son modelos alternativos que facilitan las búsquedas exploratorias, éstas suponen menos esfuerzo cognitivo que las analíticas porque se reconoce la información en la pantalla requiriendo menos esfuerzo que recuperación de la memoria.

Dentro de estos métodos que implican una búsqueda dinámica y exploratoria, implican alguna forma de browsing:

- Berrypicking − de los 80. Lo compara con la recogida de arándanos: seleccionando alguno de diferentes búsquedas.
- Pearl browsing − usan uno o varios documentos relevantes para a partir de ellos recuperar el material relacionado con ese conjunto inicial: mismo autor, mismos descriptores, citados entre ellos

· **Browsing.**

Cove y Walsh lo definen de forma muy pragmática: como el arte de no saber lo que se quiere hasta que se encuentra. Forma de búsqueda en la que el usuario interactúa con la información almacenada en el sistema y su representación para seleccionar los documentos relevantes y concretar el objetivo de búsqueda.

Hay tres tipos de browsing en función de la claridad con la que está definido el objetivo de la búsqueda y en función de la sistematización de las tácticas que se emplean:

- Browsing dirigido o específico − forma sistemática que se dirige a un objetivo claramente especificado, pero los criterios de búsqueda son imprecisos. Ej.: ojear una lista de documentos para buscar un documento conocido.
- Browsing semidirigido y predictivo o de propósito general − tiene el objetivo menos claro y definido, procede de una manera menos sistemática. Ej.: consulta a una base de datos cuyo tema interesa, usando un solo término de búsqueda para ver lo que hay sobre ese tema.
- Browsing no dirigido o de hallazgo fortuito − actividad puramente aleatoria, desestructurada y no dirigida, cuyo objetivo no está claramente definido. Ej.: búsqueda en un catálogo a ver si hay algo que interesa, o zapping en la televisión.

- **Técnicas para la modificación de la búsqueda**.

Como consecuencia de los problemas que tienen los usuarios para expresar su necesidad de información el planteamiento inicial de búsqueda que hacen puede ser una representación inadecuada o incompleta de esa necesidad, bien porque el usuario no tienen claro lo que quiere, búsqueda no bien definida, o porque el usuario no entienda la sintaxis ni la terminología del sistema de recuperación.

En este contexto podríamos definir la ampliación de la búsqueda (modificación) como el proceso de complementar el planteamiento de búsqueda inicial con términos adicionales como método para mejorar los resultados de la recuperación. El método en sí mismo se puede aplicar a cualquier situación independientemente de las técnicas de recuperación que use el sistema. Además entendida la modificación en un sentido amplio, no sólo significa incluir nuevos términos sino también eliminar términos del planteamiento inicial.

En cualquier forma de modificación de la búsqueda hay que considerar dos factores, que han servido a Efthimiadis para establecer una clasificación de los tipos de modificación:

- **Método usado para seleccionar los métodos.**
  - Manual − cuando el usuario busca los términos que va a emplear o los toma de algún documento previamente recuperado, pero es el usuario quien teclea los nuevos.
  - Automática − es el propio sistema quien selecciona los términos y los introduce en una nueva búsqueda.
  - Interactiva − cuando el sistema ofrece al usuario una relación de términos que éste puede seleccionar o no para replantear la búsqueda.

- **Fuente utilizada para seleccionar los términos** − en función de este factor la ampliación puede estar basada en:
  - Resultados de una búsqueda previa
  - Estructuras de conocimiento

Estas cinco posibilidades se pueden combinar entre ellas.

- **Modificación automática basada en los resultados.**

Se conoce como retroalimentación por relevancia (relevance feedback). En este proceso los documentos que han sido relevantes en una interacción previa se convierten en fuente de términos con los que el sistema va a reformular la búsqueda. Esta definición corresponde a la retroalimentación por relevancia positiva.

También ha sistemas que trabajan con la relevancia negativa: eliminan de la búsqueda los términos que aparecen en los documentos que el usuario ha considerado que no le interesan.

El diseño de los SRI debe tener en cuenta dos factores básicos:

- Método utilizado para obtener juicios de relevancia sobre los documentos − Criterios a tener en cuenta:

- Tamaño del conjunto de documentos relevantes en el que el sistema debe basar su estimación − Un solo documento relevante. Los estudios experimentales han demostrado que los mejores resultados se obtienen entre 10−20 documentos visualizados y 5 documentos relevantes.
- Formato de presentación en el que deben basarse los juicios de relevancia que hace el usuario − qué cantidad de información sobre el documento necesita ver el usuario para hacer un juicio de relevancia. Los estudios experimentales confirman que los usuarios deben enjuiciar después de ver la presentación más completa que exista en el sistema: un resumen, los basados sólo en el título no dan buenos resultados.
- Naturaleza de los juicios de relevancia − la mayor parte de los SRI usarán juicios de relevancia binarios, sólo consideran documentos relevante o no relevante, no consideran documento poco relevante, bastante empobreciendo las posibilidades de los términos, los usuarios no valoran sólo la relevancia temática de los documentos sino su relevancia contextual, su utilidad.

- Algoritmos para la selección de los términos de los documentos relevantes y para su ponderación − intentan determinar qué términos se van a usar de los documentos recuperados. Si los juicios de relevancia se basan en un solo documento no se necesitarán los algoritmos, pero sí se necesitarán si los juicios de relevancia se basan en más de un documento.

Asignar peso a la representatividad de los términos y ordenarlos en función de ese peso para usar sólo los más

representativos.

- Algoritmo de Porter − se tiene en cuenta la ocurrencia del término en los documentos relevantes y la frecuencia de aparición de ese término dentro de la colección.

Porter = −

R − conjunto de documentos relevantes recuperados

r − conjunto de documentos relevantes que tienen el término t

n − número de documentos de la colección que tienen el término t

N − número de documentos en la colección

Para seleccionar los términos además de usar un algoritmo hay que establecer un valor de corte, que suele ser un número máximo de términos de ese listado ordenado que serán incorporados a la búsqueda. Los experimentos realizados varían en cuanto al número de términos, pero entre unos 20−30 términos.

Cuando los sistemas son interactivos, ofrecen al usuario un listado de los términos para que éste seleccione los que le parecen más importantes que se incluyen luego en la búsqueda.

- **Estructuras de conocimiento.**

Es otra herramienta que se usa actualmente para modificar las búsquedas. Durante años las que se usaban tradicionalmente en los sistemas de información: clasificaciones, listas de encabezamientos de materias, tesauros; se consideraron un obstáculo para la recuperación de información, los argumentos que se usaban para defender esto eran que estas estructuras usaban vocabulario controlado y sintaxis propia y que era muy difícil que el usuario usara en sus búsquedas los mismos términos y la misma sintaxis que se habían usado en la indización.

Esta valoración se realizaba desde un punto de vista en el que se considera sólo el modelo de las búsquedas analíticas y que trata de defender desde un optimismo tecnológico poco crítico la eficacia de los sistemas automatizados de indización y recuperación basados sólo en el LN (lenguaje natural).

En la última década y ante la evidencia de que los usuarios no expertos necesitan ayuda, muchos sistemas de recuperación comerciales incluyen algún tipo de estructura de conocimiento o explotan más las posibilidades de las estructuras que se empleaban tradicionalmente.

Las razones para que se revaloricen son que estas estructuras permiten mejorar el conocimiento conceptual del usuario sobre qué hay en el espacio de información del sistema (qué contiene el sistema), cómo se organiza ese espacio de información, y sobre los términos que se han usado para representar los documentos. También pueden usarse para hacer coincidir el vocabulario del usuario con el del sistema o para crear equivalencias entre las estructuras usadas en diferentes sistemas de recuperación. Por tanto pueden usarse tanto para realizar búsquedas exploratorias como para ayudar a formular búsquedas analíticas y también para realizar operaciones de retroalimentación por relevancia.

Efthimiadis clasifica estas estructuras de conocimiento en dos grupos:

- **Dependientes de la colección** − se generan a partir de los documentos que ya están almacenados en el sistema, se generan empleando algún método automático basándose en características de los documentos como los términos de indización creando grupos de

documentos temáticamente afines.

- Citas − generar matrices con citas en un período, autores citan a otros, se calculan similaridades, se consigue visualizar tendencias, colegios invisibles, escuelas
- Tesauro por asociación − generado automáticamente en el que la relación entre los términos se establece en función del número de veces que esos términos aparecen juntos. La forma es igual a la anterior: matrices, período, similitud

- **Independientes de la colección** − los tradicionales LD usados en bibliotecas y centros de documentación: tesauros, listas de materias

- Redes semánticas − para solucionar problemas de acceso a la información multilingüe. El proyecto más importante en la UE es EurowordNet con equivalencias en LN en inglés, francés, castellano, catalán En esta red se ven reflejadas las relaciones de los términos similares a los tesauros.

- **Utilidades.**

- Reconocer espacio informativo − mostrar al usuario qué información tiene el sistema y cómo está organizada. Se usan los índices alfabéticos de materias, autores, títulos Es más práctico usar los índices sistemáticos basados en estructuras del conocimiento. Por ejemplo en Murcia se empieza a usar la CDU para mostrar al usuario cómo están organizados los libros.
- Proporcionar términos equivalentes − las relaciones de equivalencia que hay en los LD o las clases de sinónimos que se pueden incorporar a los sistemas de búsqueda permiten llevar desde sus propios términos a los que han utilizado en el

índice del sistema al usuario. Esto se puede hacer el reenvío informando al usuario mediante algún tipo de referencia. No es frecuente que los SRI usen este tipo de herramienta, lo que es inexplicable, sobre todo en los catálogos. Se usa en CISNE, el catálogo de la Universidad Complutense de Madrid.

- Proporcionar términos alternativos − las relaciones jerárquicas y asociativas de los LD pueden facilitar al usuario términos de búsqueda en los que no había pensado y pueden ser de su interés. Las redes semánticas generadas automáticamente se basan en la misma idea y facilitan términos relacionados porque habitualmente aparecen juntos en los documentos pero no dicen de qué tipo es esa relación.

- Facilitar la interconexión de sistemas − la tecnología ha proporcionado el acceso a múltiples bases de datos, catálogos, pero para que esto sea funcional se necesita que el contenido de los recursos de información a los que se acceden sea homogéneo, sobre todo respecto a los puntos de acceso Uno de los mayores problemas de ahora es el multilingüismo. Para solucionarlo se recurre a estructuras de conocimiento. Como EurowordNet. Otro proyecto sobre ficheros de autoridades, en el marco del programa COBRA+ de la Asociación Concertada de la Comisión Europea (DGXIII), sus finalidades:

- Facilitar acceso a los documentos electrónicos.
- Facilitar la interconexión de sistemas.

Dentro de esta Comisión hay un grupo de trabajo de acceso temático multilingüe que impulsó MACS con un fichero de autoridades de materias con los principales repertorios de materias: inglesa, Library of Congress Subject Headings; francesa, RAMEAU: Répertoire d'autorité matiére encyclopedique et alphabetique unifié; y alemán: RWD.

En España se intenta solucionar el multilingüismo de interior, como el Catálogo de la Azkue Biblioteca, Proyecto ABK, que es en eusquera y en castellano. En Cataluña también se trabaja en este sentido.

- **Interfaces**

El interés por la investigación sobre las interfaces de los sistemas se explica porque es el medio a través del cual el usuario dialoga con el sistema. En el desarrollo de éstas a lo largo de los últimos treinta años se han establecido tres generaciones:

♦ **1ª generación. Visualización de índices y términos de búsquedas.**

En los años 70. Esta orientada a un entorno de trabajo en el que va a ser usada por expertos en recuperación de información, en centros especializados, con ficheros estructurados en campos, acceso secuencial a esos ficheros, y para búsquedas analíticas usando lógica booleana. Interfaz de órdenes y al incluir menús, estos son de líneas.

♦ **2ª generación. Enlaces hipertextuales.**

Década de los 80. El entorno cambia. Orientada a usuarios finales que conocen en alguna medida el sistema. Se usa en centros especializados y centros de acceso público. Combinan los SGBD con los SGD. Se trabaja con documentos en texto completo y con estructuras hipertextuales. Se usa para búsquedas analíticas, aunque también empiezan para sistemas avanzados de recuperación de información, no ya sólo con el álgebra de Boole.

Las interfaces empiezan con menús desplegables, incorporando cuadros de diálogo, e interfaces gráficas con un lenguaje basado en iconos, el GUI.

♦ **3ª generación. Presentaciones gráficas.**

Y los front−end−interfaces que se crean independientemente de la base de datos, incluyen facilidades para el manejo de la información documental y se pueden implementar a distintos sistemas de recuperación. La interfaz se desarrolla para el cliente y puede ser igual para todos los servidores.

- Entorno
  - ◊ Acceso universal.
  - ◊ Sistemas distribuidos / WWW.
  - ◊ Multimedia.
  - ◊ Navegación y browsing frente a búsquedas analíticas.
  - ◊ Interactividad − usuario y sistema interactúan.
- Interfaces.
  - ◊ Interfaces gráficas − basadas en metáforas que representan espacio de información en 2D, 3D o realidad virtual.
  - ◊ Manipulación directa − es lo que se busca, no órdenes a través de lenguaje.
  - ◊ Perfiles de usuario − diferenciados.

El avance de estas tres generaciones ha hecho pasar del LIBERTAS: basado en un menú de selección de órdenes, con interfaz poco amigable y presentación de los índices sin enlaces hipertextuales. Al WebCat: con enlaces hipertextuales.

Hildreth dice que las mejoras son superficiales, sin mejoras fundamentales.

La investigación en mejoras en las interfaces para recuperación de información en INNOPAC y en bases de datos documentales se centra en tres aspectos:

♦ **Formas de visualizar índices y registros**

- Tipo de respuesta que debe dar el sistema a una búsqueda de un usuario

Responder mostrando:

◊ Directamente los registros − como ABSYS y cualquier buscador de Internet, muestra los registros que hay, presentación abreviada, no lleva a los índices. El problema es ordenar los registros y sobrecarga de información.
◊ Mostrando índices del sistema que hay sobre el campo buscado − como INNOPAC.

- Formato más adecuado para mostrar al usuario

Abreviados, completos, en bases de datos los registros etiquetados o en formato MARC o ISBD. En ABSYS es abreviado y en INNOPAC es texto completo con etiquetas.

- Orden en que se deben visualizar los términos de un índice.

Afecta a los encabezamientos de materia y subencabezamientos. Orden alfabético seguido de secuencia completa del encabezamiento produciendo modificaciones en el orden lógico de presentación por los calificadores o modificadores, o subencabezamientos.

En 1997 se formó un grupo de trabajo en la IFLA, el Task Force on Guidelines para redactar unas recomendaciones para la visualización de la información bibliográfica en catálogos en línea, en sitios web, en cualquier tipo de interfaz gráfico, y en recursos multimedia. Redactó un borrador provisional con recomendaciones publicado en Internet para ser base de un debate aún abierto, fue muy polémico. Recomendaciones a los problemas planteados:

- Tipo de respuesta que debe dar el sistema a una búsqueda de un usuario − mostrar al usuario los índices de los campos de búsqueda siempre que el resultado sea más de un registro, sino mostrar ese registro completo.
- Formato más adecuado para mostrar al usuario − mostrar formatos abreviados como respuesta posterior al índice alfabético dando al usuario la posibilidad de visualizar formas más completas en formato etiquetado o en ISBD.
- Orden en que se deben visualizar los términos de un índice − ordenar los índices teniendo en cuenta la secuencia alfabética de cada uno de los subcampos sin incluir los modificadores.

♦ **Tipo de enlaces hipertextuales son más útiles y manejables**.

Por influencia del entorno Web se incorporan a los catálogos en línea y a las bases de datos que no tienen una estructura hipertextual. Facilitan la navegación y búsquedas exploratorias de forma muy intuitiva y eliminan el esfuerzo de que el usuario replantee la búsqueda. Se emplean para unir términos del índice con los registros a que está asignado y también para unir campos de un registro con registros relacionados porque comparten ese punto de acceso.

Ahora los enlaces hipertextuales son falsos, el sistema al seleccionar nosotros un término del índice plantea una búsqueda por equiparación por esos términos, no hay una estructura de nodos, el formato MARC no lo soporta.

El usuario se pierde con facilidad, hay que hacer que visualice un mapa de su recorrido, dónde está y cómo ha llegado ahí. Esto en los buscadores de Internet se pone por ejemplo como: World > Español > Medios de comunicación Esto no se hace en los catálogos en línea.

♦ **Presentaciones gráficas.**

Tratan de situar al usuario de manera metafórica en un espacio de información, favorece el uso de las capacidades cognitivas de reconocimiento espacial frente a los sistemas tradicionales que emplean las capacidades cognitivas de tipo conceptual porque trabajan con lenguaje y con la lógica.

Las representaciones gráficas se han desarrollado de manera experimental y la mayoría usa como base para crear la presentación técnicas automatizadas de clasificación.

- En 2D − uno de los métodos más usados son los mapas organizativos usan redes neuronales para crear espacios de información como el de Kohonen (las zonas más amarillas tienen más documentos, sobre un fondo rojo oscuro).
- En 3D.
    - ◊ Las que se basan en nodos espaciales − muy usadas para navegar en entornos hipertextuales en los que hay unos racimos de nodos enlazados con unas líneas de conexión. Ejemplo: Narcissus (Universidad de Birmingham).
    - ◊ Las basadas en mapas estructurales − en las que se crea una metáfora de la estructura de la información presentada al usuario. Ejemplo: LyberWorld para INQUERY con dos presentaciones gráficas: una esfera de relevancia después de una búsqueda analítica, o los conos de navegación con una estructura jerárquica.
- En realidad virtual − uno de los proyectos es VIRGILIO, de dos institutos alemanes y la Universidad de la Sapienza (Roma). Trata de facilitar el acceso a una base de datos musical.

20

TÉCNICAS AVANZADAS DE INDIZACIÓN Y RECUPERACIÓN DE LA INFORMACIÓN.

PRIMER CURSO. PRIMER CUATRIMESTRE.