UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

UN SISTEMA DE RECUPERACION DE INFORMACION DE PROPOSITOS GENERALES

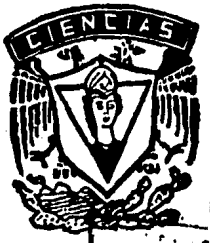
T E S I S

QUE PARA OBTENER EL TITULO DE

M A T E M A T I C O

P R E S E N T A :

MARCO ANTONIO RAMIREZ GONZALEZ



MEXICO, D F. 1994

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

VNIVERSIDAD NACIONAL
AVFNºMA DE
MEXICO

FACULTAD DE CIENCIAS
División de Estudios
Profesionales
Exp. Núm. 55

M. EN C. VIRGINIA ABRIN BATULE
Jefe de la División de Estudios Profesionales
Universidad Nacional Autónoma de México.
P r e s e n t e .

Por medio de la presente, nos permitimos informar a Usted, que habiendo revisado el trabajo de tesis que realizó el pasante RAMIREZ GONZALEZ MARCO ANTONIO con número de cuenta 7716522-0 con el título: "UN SISTEMA DE RECUPERACION DE INFORMACION DE PROPOSITOS GENERALES

Consideramos que reune los méritos necesarios para que pueda continuar el trámite de su Examen Profesional para obtener el título de MATEMATICO.

GRADO NOMBRE Y APELLIDOS COMPLETOS	FIRMA
ING. HECTOR CAMPOS ESTRADA Director de Tesis	
MAT. ENRIQUE DUEÑAS BLANQUEL	
M.EN C. GUSTAVO ARTURO MARQUEZ FLORES	
ACT. NOE MOACYR VALLEJO GONZALEZ Suplente	
MAT. HUGO VILLASEÑOR HERNANDEZ Suplente	

Ciudad Universitaria, D.F., a 17 de MAYO de 1994

DEDICATORIA

A MIS PADRES:

POR HABERME DADO ESTA OPORTUNIDAD...

A BETO Y A ROSY:

POR QUE SE CUMPLAN TODOS SUS SUEÑOS.

A MIS SOBRINOS:

ALEJANDRO, VIVIANA, ADRIANA Y JOSE ANGEL.
PUES ELLOS SON NUESTRA ESPERANZA.

A ALEJANDRO:

POR TU EJEMPLO.

A JOSE ANGEL:

QUE REALICES TUS ILUSIONES.

A MIS CUÑADAS CECILIA Y RAQUEL.

A MAMA RUFINA...Y A PAPA LUIS † QUE NOS ENSEÑASTE A LUCHAR
HASTA EL ULTIMO MOMENTO.

A LA FAMILIA QUE PROCREARON.

AGRADECIMIENTOS

A el Ing Héctor Campos E. por su apoyo en todo momento para la realización de éste trabajo y por su amistad.

A la Lic. Sonia Marquez V, por sus valiosas aportaciones, particularmente en la elaboración del capítulo 2 y del anexo.

A mis compañeros de trabajo en la U.N.A.M. y posteriormente en la P.G.R. por su interés en la conclusión de este trabajo y por su amistad

Al mis compañeros de la P.G.R.: Liliana, Rosario, Alfonso, Ricardo y Armando por su ayuda para que el trabajo pudiera salir a tiempo

A todos ellos, de verdad muchas gracias.

Sinceramente...

Marco.

INDICE

OBJETIVO.

El objetivo del presente trabajo es, por una parte, proporcionar un panorama bastante general de los conceptos e ideas desarrolladas en relación al tema, ya que las investigaciones realizadas hasta la fecha en nuestro país no son abundantes. Así, se espera que el material expuesto sirva como introducción a quien desee adentrarse en la gran variedad de perspectivas de estudio que ofrece la recuperación de información. Por otra parte se desarrollará un sistema de recuperación de información que intente en la medida de lo posible, ser aplicado en cualquier área del conocimiento, es decir, que cualquier persona que posea un volumen considerablemente grande de textos o documentos, encuentre en el sistema que habremos de desarrollar, una herramienta para almacenar tales textos y posteriormente tener acceso rápido y con eficacia suficiente para satisfacer sus requerimientos de información, todo ello claro está, por medio de una computadora personal.

INTRODUCCION.

Sea cual sea el ámbito en el que llevemos a cabo nuestras actividades de estudio o investigación, hemos necesitado en cierto momento información sobre algún tema específico. El problema de almacenar y recuperar información consiste en lo siguiente: tenemos una gran cantidad de información, a la cual, el accesar con rapidez y precisión es cada vez más difícil. Esto ocasiona que al menos una parte de la información que es de interés sea ignorada, ya que no es posible cubrirla durante la búsqueda, debido a que ésto lleva implícito una enorme cantidad de trabajo y esfuerzo. Este problema se manifiesta en una situación paradójica: existe demasiada información y al mismo tiempo demasiada poca información. Es decir, el volumen de textos es demasiado grande, pero no podemos tener un buen acceso a la información deseada, precisamente a causa del gran tamaño de la colección de textos.

Veamos el problema un poco más de cerca: Supongamos que se tiene alamacenada una gran cantidad de documentos y una persona (USUARIO) se encuentra con un requerimiento de información (PREGUNTA), cuya respuesta es un conjunto de documentos que satisfacen la petición de información solicitada. Una manera en que el usuario puede obtener ese conjunto, es leyendo todos los documentos que se tienen almacenados, reteniendo los que sean relevantes y descartando los que no lo sean. En cierta forma, esto constituye una recuperación "perfecta", pero obviamente es una solución impracticable.

Con el desarrollo de la informática, llegó a creerse que las computadoras serían capaces de "leer" una colección de documentos y extraer aquellos que fueran relevantes a una pregunta. Pero la caracterización automática de un texto, es decir, el proceso mediante el cual la máquina intente

realizar el proceso humano de leer, es un problema bastante serio. Leer implica tratar de extraer información, sintáctica y semántica del texto y usarla para decidir si cada documento es relevante o no, a una pregunta dada. La dificultad radica no solamente en saber cómo extraer la información, sino también en cómo usarla para decidir la relevancia.

Podemos decir que esta noción (**RELEVANCIA**), es el centro de la recuperación de información. La estrategia de la recuperación automática consiste en recuperar todos los documentos relevantes y al mismo tiempo, tratar de no recuperar aquellos que no lo sean. La caracterización de un documento, debe permitir la recuperación del mismo, cuando sea relevante a una pregunta. Para un ser humano es intelectualmente posible decidir la relevancia de un documento en respuesta a una pregunta, para que una computadora lo haga, es necesario construir un modelo dentro del cual la decisiones de relevancia puedan ser cuantificadas.

A continuacion se presenta un resumen de lo que habrá de tratar cada uno de los capítulos del presente trabajo, con el fin de proporcionar una visión más detallada del objetivo del mismo.

CAPITULO 1.- INTRODUCCION A LOS SITEMAS DE RECUPERACION DE INFORMACION.

Se realizará una investigación acerca del concepto de recuperación de información y sus diferencias con respecto a la recuperación de datos, de igual manera se abordarán temas inherentes a la clasificación de documentos. También se habrán de mencionar las condiciones necesarias para que una recuperación pueda considerarse óptima. Se tratarán las

teorías en que se basa la evaluación de la eficacia en la recuperación de información.

CAPITULO 2.- REPRESENTACION DE LA INFORMACION.

Analizar de qué manera algunos elementos lingüísticos, tales como la sinonimia y la polisemia, pueden influir en la respuesta que ofrece un sistema de recuperación de información a un requerimiento específico. Además, se expondrán algunas herramientas que disminuyen en cierto grado los problemas causados por dichos elementos.

CAPITULO 3 .- ALGUNOS MODELOS FORMALES PARA LA RECUPERACION DE INFORMACION.

Revisión de modelos en los que se basan algunos sistemas de recuperación de información (estrategias de almacenamiento y búsqueda). Se mencionará en que consisten, cuales son sus ventajas y desventajas.

CAPITULO 4 .- PROPUESTA DE NUESTRO SISTEMA.

Se llevará a cabo un análisis para la elección de la metodología a usar en el desarrollo del sistema, así como de las estrategias de clasificación y búsqueda de documentos, para tratar de alcanzar una recuperación exitosa.

CONCLUSIONES.

En este capítulo se evaluarán los resultados obtenidos a lo largo del desarrollo de nuestro proyecto.

ANEXO.- SISTEMAS DE RECUPERACION DE INFORMACION EXISTENTES.

Aquí haremos una descripción muy general de los sistemas de recuperación de información, qué técnicas usan y los modelos en que se basan. Identificar qué sistemas existen en México y cuales de ellos han sido desarrollados por mexicanos.

CAPITULO 1

INTRODUCCION A LOS SISTEMAS DE RECUPERACION DE INFORMACION

1.1 DESCRIPCION GENERAL DE UN SISTEMA DE RECUPERACION DE INFORMACION

El siguiente diagrama ilustra un sistema de recuperación de información. Se muestran tres componentes: entrada, proceso y salida.

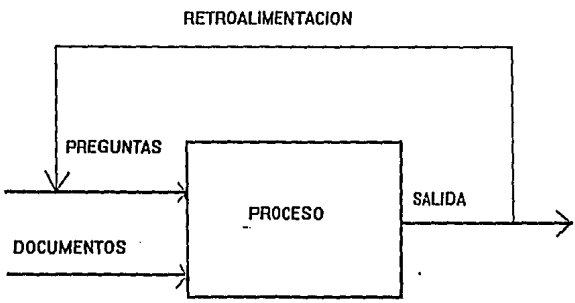


FIG 1.1

Comencemos con la entrada, el problema principal es obtener una representación de cada documento (y cada pregunta), de tal forma que puedan ser usados en una computadora. Algunos sistemas generan una representación del documento la cual es almacenada, esto significa que el texto íntegro del documento se pierde una vez que es procesado para generar tal representación, la cual, en la mayoría de los casos se constituye de una lista de palabras extraídas del documento, la característica de estas palabras es que se consideran significativas, es decir describen lo mejor posible al documento. Cabe aquí aclarar que actualmente y gracias a los adelantos en los medios de almacenamiento electrónico de datos, es posible desarrollar sistemas que guarden el texto completo del documento, trataremos más adelante y con más detalle tales sistemas.

Ante la problemática de tener que analizar el contenido de los documentos y "leer" las preguntas de búsqueda, sería ideal poder utilizar el lenguaje natural, lo cual es un problema sin solución a la fecha. Una alternativa es crear un lenguaje artificial mediante el cual todas las preguntas y documentos puedan ser expresados, esto presupone que el usuario deberá pensar cómo expresar la información que necesita mediante ese lenguaje.

Es conveniente que el usuario tenga la facilidad de poder cambiar la forma de su pregunta después de obtener un conjunto de documentos recuperados y de esta manera tratar de obtener una mayor eficacia mediante subsecuentes recuperaciones, en tal caso se dice que se trata de un sistema con retroalimentación, la idea es que el sistema funcione de manera dinámica e interactiva con el usuario.

Volviendo a nuestro esquema, la segunda parte es el proceso, en el cual se realizará la recuperación, involucra la estructuración de la información en forma adecuada, es decir, su clasificación. Además, debe ejecutar las estrategias de búsqueda de los documentos que habrán de dar respuesta a una pregunta.

Finalmente, la salida, la cual es normalmente un conjunto de citas o números para hacer referencia a los documentos recuperados y que deben ser relevantes a la pregunta. En los casos en que el sistema sea capaz de almacenar el texto completo del documento, también deberá de proporcionar acceso al documento mismo, para que pueda ser consultado.

Aunque la recuperación de información puede ser subdividida de muchas maneras, podemos decir que hay tres áreas con las cuales podríamos cubrir gran parte de la materia: análisis de contenido, clasificación de la información, y evaluación de la eficacia de la recuperación. La primera se encarga de la descripción de los contenidos de los documentos en forma apropiada para su procesamiento, la segunda de explotar las relaciones entre los documentos para mejorar la eficiencia y la eficacia de las estrategias de búsqueda y la tercera de medir la eficacia de la recuperación. Cada una de estas áreas serán tratadas con detalle más adelante.

1.2 DIFERENCIAS ENTRE RECUPERACION DE DATOS Y RECUPERACION DE INFORMACION

Podemos describir cuatro áreas principales en las que el acceso a datos y el acceso a documentos son diferentes.

a) Método de pregunta

Un sistema recuperador de datos opera mediante pregunta directa. La forma de preguntar por la información deseada es específica, por ejemplo: "¿ Cual es el salario del Sr. Iglesias ?" o "¿Cual es el monto de las comisiones pagadas a los vendedores el mes pasado en la ciudad de Querétaro?", etc. Esta operación no es tan directa en un sistema de recuperación de información, el cual no recupera la información específica deseada, sino referencias a un conjunto de documentos que probablemente contendrán tal información. A este nivel, algunos sistemas son llamados sistemas recuperadores de referencias. Aquí, las preguntas tienen un sentido más general que en la recuperación de datos, por ejemplo: "¿ Qué reportes se tienen que hablen sobre la postura de nuestros competidores en el mercado?", "Qué artículos de la Constitución de los Estados Unidos Mexicanos están relacionados con las garantías individuales?", "¿En cuáles de sus discursos el Presidente de la República Mexicana habló sobre el Tratado de Libre Comercio?", etc. Tales preguntas son indirectas en el sentido de que normalmente no pueden ser traducidas en preguntas formales y como resultado de ello, información no relevante será recuperada junto con información que sí lo es.

b) Relación entre la pregunta y la respuesta que satisface al usuario.

En recuperación de datos existe una relación necesaria entre una pregunta formal bien construída y la respuesta correcta a esa pregunta. Por ejemplo: una pregunta formal para conocer el domicilio del Sr. Alatriste debe ser contestada con el domicilio correcto, ningún otro lo será. Esta relación entre la pregunta y la respuesta correcta significa que los sistemas de recuperación de datos son, desde un punto de vista lógico, determínisticos. Para la recuperación

de documentos, existe usualmente una relación probabilística entre la pregunta formal y la posibilidad de que la pregunta sea satisfecha. Esta posibilidad puede ser muy alta, pero no es más que eso, una posibilidad. Esta relación probabilística entre la pregunta formal y la posibilidad de satisfacción, significa que desde un punto de vista lógico, un sistema de recuperación de información es no-determinístico. Dicho de otra forma, la misma pregunta formal sometida a dos diferentes sistemas de recuperación de datos (con idénticas bases de datos), recuperarán idénticos conjuntos de datos; pero si es sometida a dos diferentes sistemas de recuperación de documentos (con idénticas colecciones de documentos), no necesariamente se recuperarán los mismos conjuntos de documentos.

c) Criterio de recuperación exitosa.

Para recuperación de datos el criterio de éxito es relativamente sencillo, sólo es necesario preguntar si el sistema responde la pregunta correctamente, se puede decir que el criterio de éxito en la recuperación de datos es la exactitud, este criterio tan objetivo da lugar a que el problema de evaluar el sistema, en términos de la eficacia, sea bastante simple. No es tan fácil determinar la eficacia de un sistema de recuperación de información, en este caso se debe preguntar si el sistema satisface las necesidades del usuario, o si se encontraron todos los documentos útiles o relevantes, de aquí que el criterio de éxito en recuperación de documentos, es la relevancia, un criterio mucho más subjetivo que la exáctitud y, en consecuencia mucho más difícil de medir. Como resultado de esto, los sistemas de recuperación de información pueden ser evaluados en base a su grado de eficacia.

d) Rapidez en la recuperación exitosa.

Debido a que los sistemas de recuperación de datos son esencialmente determinísticos, la rapidez con la que un usuario recibe una respuesta satisfactoria, depende en gran parte de la rapidez de búsqueda física. La naturaleza no-determinística de la recuperación de documentos, hace que su rapidez dependa no tanto de la búsqueda física, sino de

el número de decisiones lógicas que el usuario debe hacer durante su búsqueda, estas decisiones lógicas consisten en actividades tales como: construcción formal de las preguntas, evaluación de la utilidad de los documentos recuperados, revisión de la pregunta formal en caso de que el conjunto recuperado no sea satisfactorio, etc. Si un usuario debe buscar en grandes conjuntos de documentos recuperados o debe revisar continuamente su pregunta, entonces estas actividades, y no la rapidez de acceso físico, determinarán qué tan rápida será la búsqueda.

La distinción entre acceso a datos y acceso a documentos tiene un gran efecto sobre el diseño y la operación de los sistemas de recuperación de información. Desde el punto de vista del diseño, un sistema recuperador de datos es relativamente sencillo en el sentido de que un dato tiene normalmente un solo punto de acceso. Por ejemplo: para recuperar el valor del salario del Sr. Iglesias de una base de datos, la descripción "salario del señor Iglesias" es el único punto de acceso a ese dato específico. Los documentos, por otro lado, tienen diversos puntos de acceso, un documento puede ser recuperado, mediante descripciones o combinaciones de éstas, tales como autor, título, fecha, número de documento, tipo de documento, materia, destinatario, etc. Esta lista puede extenderse dependiendo del tipo de documentos y de acceso requeridos. Por ejemplo: al referirnos a un texto que hable de la Revolución Mexicana, podríamos tener acceso al mismo si preguntamos sobre "Revolución", pero también lo accesaríamos tal vez si preguntamos por "México", o "Villa", o "Zapata", o "Francisco". Es importante hacer notar aquí que, a causa del gran número de puntos de acceso que tienen los documentos, el diseño lógico de un sistema de recuperación de información, no es tan sencillo como el diseño de un sistema recuperador de datos. Para que la recuperación de información sea óptima, debe permitir al usuario encontrar los documentos con el menor número de decisiones lógicas posible, esto significa que el sistema no solo debe proporcionar un gran número de puntos de acceso a un documento, también debe proporcionar acceso a ese documento

a través de combinaciones apropiadas de esos puntos de acceso.

1.3 LA NATURALEZA INTENTO-ERROR DE LA RECUPERACION DE INFORMACION.

La distinción entre datos y documentos tiene importantes consecuencias para los usuarios de sistemas de recuperación de información. La diferencia más clara radica en lo que se espera de la eficacia de la recuperación. En la recuperación de datos el usuario espera recuperar exáctamente el dato que busca, o bien, saber si el dato existe o no en la base de datos. En recuperación de información existen varios puntos de acceso a documentos individuales, y estos puntos de acceso normalmente no son específicos, es decir, existen varios documentos con una misma fecha, tipo de documento, materia, autor, descripción, etc. Esto significa que, a diferencia de la recuperación de datos, en recuperación de información rara vez se recupera toda y únicamente la información deseada. Es muy posible que junto con los documentos requeridos, sea también recuperado un conjunto de documentos irrelevantes (ruido informático), de aquí que el usuario tenga que hacer una revisión del total de documentos recuperados, descartando los menos útiles y quedándose con aquellos que desea.

En este sentido un buen sistema de recuperación de información puede ser tan específico como se desee pero nunca lo será tanto como un sistema de recuperación de datos.

Existe otra consecuencia importante de la falta de especificidad en la descripción de documentos. La consecuencia es que el usuario no puede saber con certeza, si existen documentos importantes en la base de datos, que no fueron recuperados después de haber realizado la búsqueda (silencio informático). Supongamos por ejemplo que un usuario busca exhaustivamente documentos haciendo uso del sistema, pero no recupera ningún documento útil, en este caso no se puede concluir que no existen en la base documentos útiles. Por otro lado cuando el usuario recupera un conjunto de documentos útiles, es demasiado dificil

inferir si tiene todos los documentos que existen en la base y que desea recuperar. Podemos entonces distinguir dos fases para la búsqueda de documentos:

1) Encontrar los documentos necesarios.

2) Asegurarse de que estos documentos son todos los que se encuentran disponibles en la base.

Además, si no se obtienen buenos resultados en una búsqueda, no quiere decir necesariamente que el sistema sea ineficaz, quizás proporcionó los mejores documentos en una pobre colección; de manera semejante, si aparentemente se obtienen buenos resultados, no significa necesariamente que el sistema esté funcionando en forma adecuada, puede ser que no esté regresando los mejores documentos, y que siempre regrese algunos de ellos, debido al gran tamaño de la colección.

1.4 ANALISIS AUTOMATICO DE TEXTOS (INDEXACION)

Un problema central, que influye en todos los aspectos de la recuperación de información es el siguiente: ¿Cómo podrían ser representados los documentos de tal forma que pudieran ser recuperados adecuadamente?. El análisis de textos tiene como objetivo resolver este problema.

El punto de partida del proceso del análisis de textos puede ser el texto del documento completo, un resumen, el título, etc.; el resultado de este proceso debe ser una representación del documento en forma tal que pueda ser manejado mediante una computadora.

Durante el proceso, una de las primeras tareas que se llevan a cabo es la eliminación de palabras no significativas: existen ciertos tipos de palabras que por sí solas no dicen nada sobre el tema que trata el documento, tales como artículos, preposiciones, conjunciones, pronombres, adverbios, etc., éstas deben eliminarse y la manera más común de hacerlo es mediante la comparación de las palabras

contenidas en el documento contra una lista de estos términos no significativos. Si una palabra del documento se encuentra en esa lista, entonces no deberá ser tomada en cuenta para describir al documento. De esta manera centramos nuestra atención en aquellas palabras del documento que "nos dicen algo" sobre él, es decir, que son significativas.

Inicialmente, supongamos que tenemos una lista con todas las palabras significativas que forman parte del lenguaje natural, llamémosmeles términos índice o palabras clave, entonces afirmamos que un documento está indexado por todos aquellos términos índice que figuran como palabras significativas dentro del mismo. Es decir, el documento queda representado por el conjunto de términos índice que contiene.

Tradicionalmente, los sistemas de recuperación de información realizan un proceso de truncación de palabras con el fin de agrupar en una sola noción o idea, las palabras que presentan una misma raíz (sufijo) y que por lo tanto tienen significados iguales o semejantes. Este proceso de truncación trae consigo una serie de dificultades las cuales serán analizadas más adelante, en el Capítulo 2, donde además se tratarán algunos métodos de detección de palabras equivalentes en cuanto a su significado, pero que morfológicamente son distintas.

Un enfoque un poco más preciso de lo que es la indexación es el siguiente: el objetivo del proceso de truncación es generar una lista de nombres de clase (términos índice o palabras clave), cada uno de los cuales representa una clase de palabras, las cuales contienen una misma noción o idea y que aparecen en el texto de entrada. Aquí un término índice ya no es una palabra aislada, sino una clase de palabras representando un mismo significado. Un documento será indexado por un nombre de clase, si una de sus palabras significativas aparece como miembro de esa clase.

Al principio de esta sección se dijo que el punto de partida para el proceso de análisis de textos podría ser el texto completo del documento o un resumen; en el primer caso es obvio que se habrá de dar entrada al texto íntegro del documento, los sistemas que así lo permiten son denominados sistemas "FULL TEXT", los primeros sistemas de recuperación de documentos no permitían este tipo de almacenamiento, ya que resultaba costoso y no existían medios de almacenamiento como los que han surgido a finales de la década de los 80's y principio de los 90's, esto ha provocado que los sistemas de Full Text tengan cada vez un mayor auge. En el segundo caso lo que se hace es introducir al sistema un extracto del documento, también denominado "ABSTRACT", que contiene los aspectos más relevantes del documento es sobre éste que se realizarán las búsquedas. El extracto debe contener información adicional para poder localizar el documento propiamente dicho, en caso de que se requiera. En México, el sistema UNAM JURE, aplicado en el Instituto de Investigaciones Jurídicas de la UNAM, es la referencia más próxima que tenemos sobre un sistema que utiliza la técnica del abstract.

Resumiendo: en primera instancia tenemos el documento, el cual puede ser visto como una secuencia de palabras, el primer paso es la eliminación de las palabras que no son significativas, entonces nos quedamos con las palabras que sí lo son, posteriormente, éstas se agrupan en clases, así que los documentos quedan descritos por conjuntos de clases, denominadas comunmente palabras clave o términos índice.

Por otra parte debemos considerar la pregunta, la cual puede consistir de la combinación booleana de uno ó más términos índice, pero que tambien puede consistir de una cadena de palabras, en el segundo caso sería sometida al mismo proceso de análisis de textos para obtener una lista de términos índices, los cuales serán buscados en la colección de documentos. La figura 1.2 muestra de manera gráfica el proceso de análisis de textos. La entrada es el documento, puede ser Full Text, Abstract, etc. y la salida es un conjunto de términos índice que representan al documento.

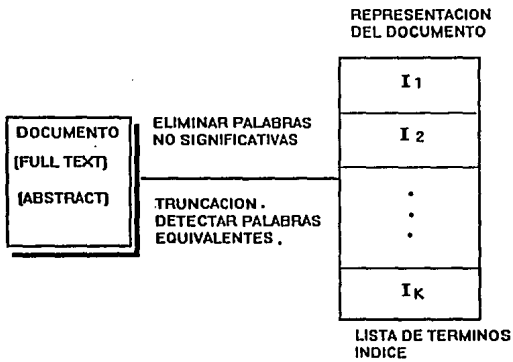


FIGURA 1.2

En la figura 1.3 podemos ver la forma en que se encuentran relacionados la pregunta, el universo de términos índice, y los documentos en la colección.

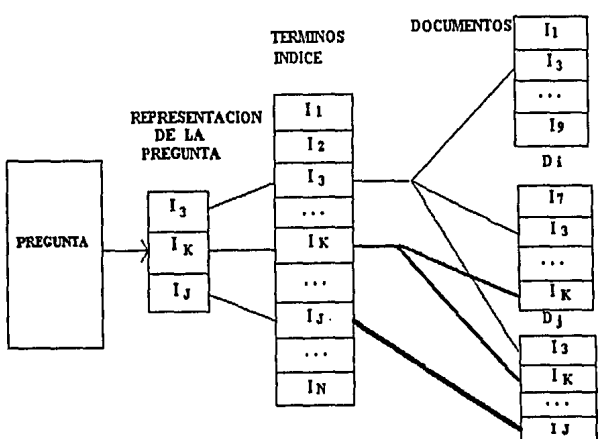


FIGURA 1.3

En la figura anterior, podemos observar que el análisis de la pregunta da como resultado su representación con términos índice ($I_3, I_k, .., I_j$), los cuales forman parte del conjunto de términos índice de la colección, es decir uno o más documentos contienen al menos uno de esos tres, las flechas indican cuales documentos los contienen, así el documento D_k contiene a los tres términos de la pregunta, por lo tanto sería un buen candidato a satisfacerla.

Hasta este punto hemos descrito de una manera bastante general el proceso de análisis de textos, desafortunadamente, el proceso de análisis de textos es a la fecha una tarea inconclusa; debido a la naturaleza intento-error de la recuperación de información, los métodos de representación de documentos desarrollados, son necesariamente una descripción incompleta del contenido intelectual del documento.

Otro punto que valdría la pena considerar en todo sistema de recuperación de información es el número de veces que aparece un término índice en cada documento (Ponderación de términos índice), esta frecuencia podría servir como indicador de qué tan importante es un término índice dentro de un documento, de tal manera que se puedan detectar niveles de importancia y recuperar sólo aquellos documentos donde el término rebase una frecuencia dada.

Es oportuno agregar el porqué es importante la representación de documentos: la recuperación se basa en cómo los documentos han sido representados en el sistema y la eficacia dependerá directamente, más que otra cosa, en la calidad de esas representaciones.

1.5 CRITERIO DE PREDICCION Y PUNTO INUTIL.

Debido a la naturaleza intento-error de la recuperación de información y al rápido crecimiento de las colecciones de documentos, es de gran importancia la forma en que el usuario formule su pregunta, mediante ella sería conveniente balancear dos aspectos complejos para una mejor búsqueda:

a) Debe predecir cómo están representados (o indexados) en el sistema los documentos que necesita .

b) Debe recuperar conjuntos de documentos lo suficientemente pequeños para poder revisarlos y encontrar los que le sean útiles.

La primera dificultad que afronta un usuario de un sistema de recuperación de información es la PREDICCION (por medio de una pregunta formal), de las palabras o frases que han sido usadas para representar (indexar) los documentos que satisfarían la pregunta.

Cuando la colección de documentos en un sistema de recuperación de información se vuelve grande, este proceso de predicción se vuelve complicado debido al tamaño resultante de los conjuntos de documentos recuperados. En otras palabras, no es suficiente, predecir uno o varios de los términos usados para indexar los documentos deseados, se debe además, recuperar esos documentos en un conjunto suficientemente pequeño como para poderlos revisar. El número de documentos en un conjunto recuperado que contiene los documentos deseados, debe ser más pequeño que el "PUNTO INUTIL" del usuario. Este punto inútil es el máximo número de documentos recuperados que el usuario está dispuesto a revisar. Dicho número variará de un usuario a otro, y aún de una búsqueda a otra para un mismo usuario, esto también es influenciado por la disponibilidad físca de los documentos al ser revisados, si es fácil hacerlo, entonces el usuario tendrá una mayor tolerancia con respecto a los conjuntos grandes. Consideremos un ejemplo: Supongamos que en un sistema de recuperación de información existe un documento D_i, el cual contiene la información que se requiere. D_i tiene seis índices asignados: I_a, I_b, I_c, I_d, I_e, I_f. Supongamos que el punto inútil del usuario es m, supongamos además que el primer término índice seleccionado para formular la pregunta de búsqueda es I_b, hasta aquí se ha satisfecho el primer criterio para lograr una búsqueda exitosa, prediciendo mediante una pregunta formal, un término usado para indexar el documento D_i que se está buscando. Este primer requisito es llamado el criterio de predicción (CP), pero no es suficiente para que la selección de I_b como término de búsqueda, tenga como resultado la recuperación exitosa de D_i; el número de documentos recuperados (llamémosle la

AMPLITUD de I_b), debe ser menor que m, este es el criterio del punto inútil (CPI).

Más que por su tamaño, el punto inútil es importante por la forma en que afecta la estrategia de búsqueda. Por ejemplo, supongamos como antes, que utilizamos a Ib como pregunta formal y que D_i es el único documento que desea el usuario, esta pregunta, está claro, recuperará todos los documentos que tengan asignado a I_b como índice, y el tamaño del conjunto recuperado será igual a la amplitud de I_b, si esta amplitud es más grande que m (el punto inútil), entonces el usuario no estará dispuesto a revisar dicho conjunto, ya que sería demasiado grande, teniendo como consecuencia que el documento no sea localizado. Así, la estrategia debe dirigirse ahora a reducir el tamaño del conjunto de documentos recuperados a un número menor que m, hablando en sentido práctico, el usuario se interesa primero en satisfacer el CPI y sólo posteriormente con cumplir el CP. Esto podría hacerse de dos formas: primero, el usuario debería seleccionar otro término I_j, el cual satisfaría probablemente el CP, entonces podrá elegir una de las siguientes alternativas:

1) Descartar I_b y usar I_j como pregunta de búsqueda.

2) Construir una pregunta de búsqueda formal que consista de la conjunción (intersección booleana) de I_b con I_j.

Como se ha dicho, el determinar si un término seleccionado para una pregunta satisface el CP depende de que el usuario revise los documentos que están indexados por el término, esto es relativamente sencillo si la amplitud del término en cuestion es menor que m, es decir se cumple el CPI y lo único que el usuario tiene que hacer es revisar completamente el conjunto de documentos recuperados, si uno de esos documentos satisface el requerimiento del usuario, entonces podemos determinar que el CP se cumple y se concluye la búsqueda en forma exitosa. Pero en sistemas de recuperación de documentos grandes, esta determinación no es tan sencilla, en este caso, la mayoría de los términos índices tendrá una amplitud mayor que m y si el usuario escoge uno de estos términos para realizar su búsqueda,

entonces se verá forzado a escoger una de las dos estrategias descritas anteriormente. Supongamos, como antes, que el usuario únicamente será satisfecho con D_i y que inicialmente selecciona el término Ik para su búsqueda, los dos criterios se cumplirían como sigue:

El CPI se satisface si y sólo si $|I_k| < m$

El CP se satisface si y solo si $I_k \in \{I_a, I_b, I_c, I_d, I_e, I_f\}$

donde $|I_k|$ es la cardinalidad del conjunto de documentos indexados por I_k.

Supongamos además que el CPI no se cumple, es decir, que $|I_k| > m$, entonces el CP para I_k no puede ser determinado. Si el usuario escoge la estrategia 1) y sustituye I_k con un nuevo término I_j, entonces el CP para I_k no podrá ser determinado. Si adopta la estrategia 2) y formula una búsqueda conjuntiva con I_k e I_j $(I_k \cap I_j)$, entonces el CPI y el CP, deben ser probados para esta intersección de términos. Si ambos criterios se satisfacen, D_i será recuperado y la búsqueda termina.

Consideremos la posibilidad de que el CPI se cumpla, es decir $|I_k| < m$, pero el CP no, esto es, sucede una (o ambas) de las siguientes cosas:

I_k no está en el conjunto $\{I_a, I_b, I_c, I_d, I_e, I_f\}$

I_j no está en el conjunto {Ia, Ib, Ic, Id, Ie, If}

ahora, aunque el usuario puede establecer con certeza que el CPI se cumple y que el CP no se cumple con la conjunción de I_k e I_j, no puede asegurar que I_k (o I_j) individualmente no satisfacen el CP, tal vez sólo pueda hacer una estimación de ello.

1.6 SISTEMAS GRANDES Y EL USO DE TERMINOS DE BUSQUEDA INEFICIENTES.

Cuando la colección de documentos en un sistema de recuperación de información se vuelve grande y la amplitud de los términos índice es excesiva, cada vez es más dificil

recuperar conjuntos de documentos cuyas respectivas cardinalidades sean menores que el punto inútil. En este tipo de sistemas, la segunda estrategia para satisfacer el CPI es más conveniente, existen tres razones para esto.

1) Existen muchos términos índices con amplitudes demasiado grandes, y el valor típico de m normalmente es bastante pequeño, como resultado, puede ser difícil encontrar términos cuya amplitud sea menor que m . Cuando se usa la estrategia (1), el usuario utilizará los términos que primero lleguen a su mente, puede ser que ninguno de esos términos cumpla el CPI, ya que muy probablemente tendrán una amplitud bastante grande y entonces, cuando al fin encuentre un término que lo satisface, lo más seguro es que dicho término no sea tan bueno como para satifacer el CP, la razón de esto es la siguiente: lo que el usuario tiene que hacer es predecir los términos usados para indexar los documentos que necesita, entonces esperaríamos que los primeros términos que el usuario pensaría usar para la búsqueda serían los mismos que se usaron para indexar los documentos que satisfarían la pregunta. Si ninguno de estos primeros términos satisfacen el CPI, entonces el usuario debe encontrar términos adicionales para la pregunta. Para que una búsqueda tenga posibilidades de éxito, el usuario (que quiere el documento D_i), debe realizar esencialmente las mismas tareas que el indexador realizó para caracterizar a D_i, en otras palabras, el usuario está actuando como un indexador, consecuentemente, si suponemos que el indexador asigna seis términos a D_i, el usuario quizás sería capaz de predecir tres o cuatro de esos términos. Ahora, si el usuario usa esos tres o cuatro términos individualmente como pregunta y ninguno de ellos satisface el CPI, entonces la mejor forma en que D_i puede ser recuperado en un conjunto de tamaño menor que m es usar esos tres o cuatro términos en algún tipo de combinación conjunta. Esta es la estrategia (2). Claro que el problema es más complicado por el hecho de que el usuario no sabe cuales tres o cuatro términos seleccionar para que se cumpla el CP.

2) El CPI puede ser satisfecho más rápidamente intersectando términos sucesivamente, en lugar de tratar con nuevos términos individuales (estrategia (1)). En otras

palabras, mientras que los primeros cuatro o cinco términos que el usuario considera para formular sus preguntas, pueden tener cada uno amplitudes mayores que m, es menos posible, aún en sistemas grandes, que la intersección de esos cuatro o cinco términos produzca un conjunto recuperado mayor que m (a menos que m sea demasiado pequeña).

3) La razón final para usar la estrategia (2), es la tendencia de los individuos a apreciar una posibilidad particular para eventos inciertos, en estos casos, la gente estima un valor como punto de partida, el cual se va ajustando hasta llegar a la respuesta final, tal valor puede ser sugerido por la formulación del problema o por el resultado de un cálculo parcial. Esta idea está relacionada de la siguiente manera con el problema de formular preguntas: cuando se construye la pregunta, el usuario (como se ha dicho), debe predecir los términos índices que se han asignado a los documentos que está buscando, esta es una toma de decisión bajo incertidumbre, los valores iniciales con los que se comienza, son los primeros términos que se proponen como términos de búsqueda. Si esos términos no satisfacen el CPI, el usuario se verá obligado a usar nuevos términos o combinaciones de términos para búsquedas alternativas. Sin embargo, a causa de la tendencia de tomar decisiones enfocadas a satisfacer primero el CPI, tratará de mantener sus términos originales y modificar la pregunta agregando nuevos términos. Por ejemplo, si el usuario comienza seleccionando I_j como pregunta, e I_j no cumple el CPI, entonces lo más conveniente es que intersecte algún otro elemento con I_j (estrategia (2)), si la intersección de estos términos no satisface el CPI, entonces se aumentarían nuevos términos a la intersección hasta que el CPI se cumpla, En un sistema de recuperación de información grande esto sería hasta llegar a cuatro o cinco términos, supongamos que la intersección de cinco términos produce un conjunto recuperado cuyo tamaño es menor que m, pero cuando el usuario revisa dicho conjunto no encuentra lo que quiere (el CP no se cumple), en este punto el usuario debe descartar uno de los cinco términos y sustituírlo con uno nuevo, la pregunta es: ¿Cuál de los términos es más conveniente descartar? es razonable pensar que descartará el

menos importante de los cinco. Podemos suponer que el orden de selección de los términos es decreciente en cuanto a la importancia. Como resultado el último término seleccionado sería el menos importante (o uno de los menos importantes). De aquí que, en nuestro ejemplo, el quinto término será probablemente el menos importante y consecuentemente, el primero en ser remplazado.

La estrategia (2) puede ser dividida en dos etapas:

i) Verificar que el conjunto recuperado satisface el CPI, y

ii) Checar si el conjunto recuperado satisface el CP.

Gráficamente lo podemos ver a continuación:

1-Se selecciona el 1er. término	I_1	CPI no se cumple
2-Se selecciona el 2o. término	$I_1 \cap I_2$	CPI no se cumple
3-Se selecciona el 3er. término	$I_1 \cap I_2 \cap I_3$	CPI no se cumple
4-Se selecciona el 4o. término	$I_1 \cap I_2 \cap I_3 \cap I_4$	CPI se cumple se prueba el CP no se cumple.
5.- se sustituye el 4o. término	$I_1 \cap I_2 \cap I_3 \cap I_5$	CPI no se cumple
6.- se sustituye el 4o. término	$I_1 \cap I_2 \cap I_3 \cap I_6$	CPI no se cumple

.
.
.

Aunque el usuario puede ser capaz de formular nuevas preguntas de esta forma, es poco probable que estas nuevas preguntas recuperen documentos que satisfagan el CP. La razón es la siguiente: como hemos dicho, la tarea básica del usuario es predecir los términos usados para indexar los documentos que está buscando y se ha comprobado que en promedio, el número de términos para predecir con éxito, generalmente está entre dos y cuatro, si se alcanza ese número de términos al tratar de formar la intersección que habrá de producir un conjunto recuperado que satisfaga el

CPI, entonces los términos adicionales generarán conjuntos recuperados de menor tamaño y es muy probable que algunos de los documentos deseados puedan haber sido excluídos durante este proceso. Por supuesto, el número de términos índice que el usuario predice puede variar de documento a documento y de usuario a usuario, pero en promedio, tan pronto como el número de términos en la intersección es más grande que el número de términos con los cuales se podría esperar que el usuario hiciera una predicción exitosa, entonces es menos posible que la búsqueda sea afortunada. Es decir, cuando se intersecta un cierto número de términos con el fin de satisfacer el CPI, y el CP no se ha cumplido, entonces podemos empezar a sospechar que la búsqueda no tendrá éxito.

1.7 BUSQUEDA ASOCIATIVA Y EL CRITERIO DE PREDICCION

El principal impulso al trabajo teórico en el diseño de sistemas de recuperación de documentos, ha sido dirigido a ayudar a expander el número de términos de búsqueda útiles que un usuario tiene disponibles en el curso de una búsqueda, podríamos decir que la eficacia de un sistema de este tipo depende en gran medida del número de alternativas de búsqueda que se brinden al usuario. Este subcampo del diseño de sistemas de recuperación de documentos se conoce como BUSQUEDA ASOCIATIVA.

La búsqueda asociativa utiliza una gran variedad de métodos para establecer relaciones entre términos índice y, por inferencia, entre documentos en una base de datos. Tales relaciones pueden ser establecidas normativamente por medio de la construcción manual de un tesaurus, o pueden ser establecidas descriptivamente monitoreando las relaciones estadísticas entre términos índice, por ejemplo la co-ocurrrencia de los mismos. La co-ocurencia es una medida de qué tan frecuentemente dos o más términos son usados para indexar al mismo documento.

Los tipos de relaciones que se pueden establecer entre términos índices son bastante numerosos, pero todos están diseñados para lo mismo: permitir al usuario que empieza su búsqueda con uno o dos términos índice, agregar otros que de

alguna manera estén relacionados con los iniciales, para brindarle más posibilidades de construir preguntas.

Otro tipo de relación que puede manejarse es la relación entre documentos, la cual consiste en establecer una medida de asociación, o de similiaridad entre ellos, de tal manera que se genera una agrupación de documentos mutuamente relacionados. Además se elige un representante para cada uno de estos grupos de documentos, de tal forma que las búsquedas se realizarán sobre los representantes de cada grupo para posteriormente, si se desea, accesar a los demás documentos del grupo seleccionado.

Una medida de asociación está diseñada para cuantificar la semejanza entre documentos, de tal forma que es posible construir grupos de ellos, con la siguiente característica: dado un elemento en un grupo, su medida de asociación respecto a los demás elementos de ese grupo, es menor que su medida con respecto a los documentos que no pertenecen al mismo. Además, dados dos documentos, la medida de asociación entre ellos tiende a ser mayor cuando el número o la proporción de atributos compartidos por ellos aumenta.

A continuación se muestran algunas medidas de asociación conocidas, en ellas se asume que un documento está representado por un conjunto de términos índice y que la medida de contar $|\bullet|$ (número de términos que indexan al documento) proporciona el tamaño del conjunto mencionado. Por ejemplo: $|X|=12$ significa que el documento X tiene doce términos índice asignados.

$|X \cap Y|$ Coeficiente de comparación simple

$$\frac{2|X \cap Y|}{|X| + |Y|}$$ Coeficiente de Dice

$$\frac{|X \cap Y|}{|X|^{1/2} \times |Y|^{1/2}}$$ Coeficiente del coseno

El coeficiente de comparación simple es el número de términos índice que tienen en común los documentos X e Y, a diferencia de los otros dos coeficientes, éste no toma en cuenta los tamaños de los mismos. Estos últimos pueden ser considerados versiones normalizadas del primero y reflejan de una manera más real el grado de asociación entre documentos. Veamos un ejemplo:

Sean $$S_1(X,Y) = |X \cap Y|$$

$$S_2(x,y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

Supongamos que $|X_1| = 1$, $|Y_1| = 1$, $|X \cap Y| = 1$, lo cual significa que X_1 y Y_1 son idénticos, supongamos además que $|X_2| = 8$, $|Y_2| = 8$, $|X \cap Y| = 1$, es decir X_2 y Y_2 tienen cada uno ocho elementos, pero sólo uno de ellos es común a ambos documentos, lo cual indica que no son muy parecidos que digamos. Evaluando S_1 y S_2 para estas dos parejas de documentos tenemos que:

$$S_1(X_1,Y_1) = |X_1 \cap Y_1| = 1$$
$$S_1(X_2,Y_2) = |X_2 \cap Y_2| = 1$$

$$S_2(X_1,Y_1) = \frac{2|X_1 \cap Y_1|}{|X_1| + |Y_1|} = 1$$

$$S2(X_2,Y_2) = \frac{2|X_2 \cap Y_2|}{|X_2| + |Y_2|} = 1/8$$

Tenemos que $S_1(X_1,Y_1) = S_1(X_2,Y_2)$ lo cual no resulta lógico ya que X_1 y Y_1 son idénticos mientras que X_2 y Y_2 no lo son. Por otro lado observemos que con S_2, la medida de asociación

varía entre 0 (documentos totalmente distintos) y 1 (documentos idénticos), así S_2 mide de una forma más real el parecido entre documentos. Gráficamente:

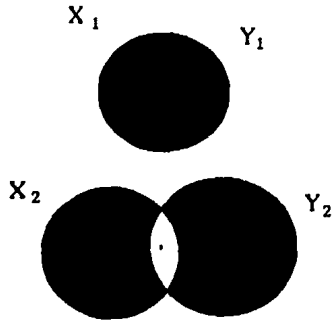


FIG 1.4

Una manera alternativa para construir grupos de documentos es utilizar una medida de disimilaridad, con ella podemos establecer qué tan diferentes son dos documentos, para decidir si pueden pertenecer a un mismo grupo o no. En seguida damos una definición matemática de disimilaridad, la cual es, en cierto sentido, semejante al concepto de distancia que se utiliza en matemáticas:

Si P es el conjunto de documentos a ser agrupados, un coeficiente de disimilaridad es una función de P X P (el producto Cartesiano) en los números reales no negativos que satisface las siguientes propiedades:

1) $D(X,Y) \geq 0$ para todo $X, Y \in P$

2) $D(X,X) = 0$ para todo $X \in P$

3) $D(X,Y) = D(Y,X)$ para todo $X, Y \in P$

Veamos un ejemplo de coeficiente de disimilaridad:

$$D(X,Y) = \frac{|(X \cup Y) - (X \cap Y)|}{|X| + |Y|}$$

Si ignoramos el caso en que ambos conjuntos de términos son vacíos tenemos que $(X \cup Y) - (X \cap Y)$ es diferente del conjunto vacío, así que su cardinalidad es mayor que cero y por lo tanto $D(X,Y) > 0$.

Por otro lado:

$$D(X,X) = \frac{|(X \cup X) - (X \cap X)|}{|X| + |X|} = \frac{|X - X|}{2|X|}$$

$$= \frac{|\emptyset|}{2|X|} = 0$$

Finalmente, es claro que $D(X,Y) = D(Y,X)$.

Así pues, vemos que existen varias formas de cuantificar la semejanza entre documentos para conformar grupos. Es muy importante hacer notar que estos métodos para medir el parecido entre documentos, se basan en una suposición: los documentos que estan fuertemente asociados entre sí, tienden a ser relevantes a las mismas pregunta. Si esta hipótesis es válida, entonces podríamos suponer que estructurando la colección de tal manera que los documentos que están fuertemente asociados aparezcan en el mismo grupo, se obtendría una mayor velocidad y eficacia en la recuperación. Para terminar con esta sección ilustramos a continuación con un ejemplo, una manera sencilla con la cual podrían ser agrupados un conjunto de documentos. Consideremos el conjunto $\{D_1, D_2, D_3, D_4, D_5, D_6\}$; haciendo uso de una medida de asociación, cuantificamos la semejanza para cada par de documentos dentro de la colección, para este ejemplo utilizamos una matriz de similaridad con los valores obtenidos:

D_1						
D_2	.4					
D_3	.5	.8				
D_4	.8	.7	.6			
D_5	.8	.5	.7	.8		
D_6	.6	.6	.7	.9	.3	
	D_1	D_2	D_3	D_4	D_5	D_6

Podemos observar que mientras D_6 y D_5 no están muy asociados, entre D_6 y D_4 la asociación es fuerte. El siguiente paso es establecer un valor límite, dados dos documentos se considerarán en el mismo grupo si su medida de asociación está por debajo de ese valor límite. Por ejemplo, si el valor límite es .75, entonces todos los pares de documentos cuya medida de asociación sea mayor que .75, pertenecerán aun mismo grupo. Gráficamente:

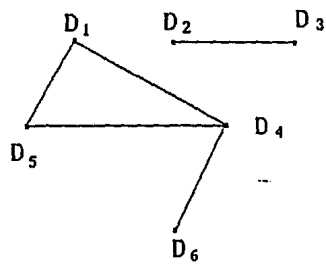


FIG 1.5

A partir de esta idea se han realizado trabajos en los cuales se aplican los conceptos de la Teoría de Gráficas para implementar modelos de agrupación y recuperación de documentos.

1.8 EVALUACION DE LA EFICACIA DE LOS SISTEMAS DE RECUPERACION DE INFORMACION.

El acceso físico a la información es la forma en la que el sistema encuentra la información deseada una vez que se conoce su ubicación. El acceso lógico (o intelectual) por otro lado, consiste en encontrar la ubicación de la

información deseada; por ejemplo, si vamos a una biblioteca y queremos saber dónde está el libro cuya colocación es Z699.V35.1979 se trata de un problema de acceso físico, pero el querer encontrar un libro que contenga la información sobre un tema en particular, es un problema de acceso lógico. Es por eso que el problema de acceso lógico debe ser resuelto antes que el problema de acceso físico.

Una computadora puede acelerar enormemente el acceso físico a los registros deseados en un sistema, pero esto no quiere decir que mejorará el acceso lógico a la información deseada. La rapidez de un sistema de recuperación de información depende directamente del número de decisiones lógicas que deba hacer el usuario en el curso de una búsqueda y no del número de registros que puedan ser revisados en un intervalo de tiempo dado. El problema de recuperar información no es un problema de cálculo, verlo así es confundir el acceso lógico con el acceso físico.

Tomando en cuenta que el propósito de la recuperación de información es proporcionar acceso intelectual a los documentos, cabría preguntarse qué deberíamos medir para determinar si un sistema de este tipo está realmente brindando un adecuado acceso intelectual hacia la colección.

Usualmente se dice que el objetivo de todo sistema de recuperación de información es dar al usuario acceso a los documentos que son RELEVANTES con respecto a una necesidad de información dada, así que lo primero que se nos ocurriría medir sería la relevancia. Desafortunadamente el concepto de relevancia no es fácil de definir ya que es una noción subjetiva; dada una pregunta, un documento puede ser relevante para un usuario, mientras que para otro puede no serlo.

Generalmente cualquier persona puede saber perfectamente cuales documentos contienen información relevante para él, pero no puede explicar tan fácilmente qué es la relevancia, es decir, no puede describir los criterios mediante los cuales ha decidido la relevancia de un documento y la no relevancia de otro. Podemos decir que el decidir si algo es relevante ó no, es de ese tipo de actividades que uno puede

practicar y a través de la práctica mejorar su ejecución, pero el ejercicio de esa práctica en poco o nada contribuye a definir o describir con detalle qué se está haciendo. Podemos saber cómo hacer juicios de relevancia, pero no podemos describir con presición la manera en que lo hacemos, es decir, no podemos describir un procedimiento general para tomar decisiones sobre la relevancia o no, de un documento.

Sería natural pensar que la relevancia es un buen indicador de la eficacia de cualquier sistema de recuperación de información, pero desafortunadamente no existen todavía métodos para cuantificar este parámetro tan subjetivo, únicamente se han hecho intentos, pero todos ellos distan mucho de reflejar con precisión qué tan eficaz es la recuperación en sistemas de este tipo.

Debido a que la recuperación de información es un proceso de intento-error podemos esperar que el resultado de cualquier búsqueda traiga consigo además de documentos útiles algunos que no lo son, no sólo eso sino que es probable que algunos documentos útiles en la base de datos queden sin recuperar. Así que los resultados de cualquier búsqueda los podemos dividir en cuatro casos diferentes:

a) Documentos recuperados que son relevantes.

b) Documentos recuperados que no son relevantes.

c) Documentos no recuperados que son relevantes.

d) Documentos no recuperados que no son relevantes.

El resultado ideal sería cuando el sistema recupere todos y únicamente documentos relevantes, es decir todos los elementos del conjunto de documentos recuperados son relevantes y no existe en el resto de la colección alguno que no lo sea. Claro que en la práctica esto rara vez ocurrirá, por no decir nunca.

Para la mayoría de los usuarios no es indispensable recuperar la totalidad de los documentos relevantes que hay en la colección, les basta recuperar una cantidad suficiente

de documentos útiles y alcanzar un nivel aceptable de satisfacción. Pero existen situaciones en las cuales es necesaria una búsqueda exhaustiva como podrían ser: búsqueda de patentes, investigaciones originales y asuntos legales; este tipo de actividades requieren criterios claros, mediante los cuales el usuario pueda decidir si todos los documentos potencialmente útiles han sido recuperados.

	RELEVANTES	NO RELEVANTES	
RECUPERADOS	X	U	NUM TOT DE RECUPERADOS
NO RECUPERADOS	V	Y	
	NUM TOT DE RELEVANTES		

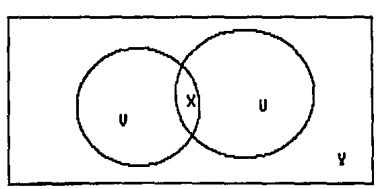


FIG 1.6

El porcentaje recuperado del total de documentos potencialmente relevantes o útiles es una de las más claras formas mediante las cuales comparar la eficacia de la recuperación. Llamemos a este porcentaje el ALCANCE (RECALL) del sistema.

El ALCANCE puede ser definido en términos de las clases de documentos representados en la figura de la página anterior como:

$$\frac{X}{X + V}$$

En términos probabilísticos esto es la probabilidad condicional de que un documento relevante sea recuperado,

$$P(DRec|DRel)$$

claro que para poderlo determinar, nos faltaría conocer el número de documentos relevantes a la pregunta dada, que existen en la colección y que no fueron recuperados (V).

Otro indicador utilizado es la proporción de documentos que se consideran útiles con respecto al total recuperado, este porcentaje es llamado la PRECISION y se define como la probabilidad de que un documento sea relevante cuando ha sido recuperado

$$P(DRel|DRec)$$

en este caso siempre podemos calcular ese porcentaje o esa probabilidad. En términos de las variables presentadas en la figura tendríamos:

$$\frac{X}{X + U}$$

La PRECISION y el ALCANCE juntos son estimaciones de qué tan eficaz es un sistema para recuperar todos y únicamente los documentos de utilidad para un usuario.

La precisión puede ser calculada si dado un conjunto de documentos recuperados, el usuario (quien hizo la pregunta) determina cuántos de ellos son relevantes y ese número se divide entre el total de documentos recuperados. Por otra

parte el ALCANCE solamente puede ser vagamente estimado, he aquí unos intentos de hacerlo:

1. Limitar la recuperación a bases de documentos relativamente pequeñas, de tal modo que sea factible detectar los documentos útiles que no fueron recuperados durante la búsqueda, la justificación de este método viene de experimentos físicos, cuyos resultados al ser realizados en pequeña escala, pueden ser generalizados. Pero existe el inconveniente de que en recuperación de información no podemos asegurar que dichos resultados sean semejantes cuando se usa una base de documentos grande, ya que en este caso el conjunto de documentos recuperados será tan grande, como para no poderlos revisar todos, provocando que el usuario tenga que volver a plantear su pregunta y como consecuencia, el experimento ya no es exáctamente el mismo que cuando se utilizó una base pequeña.

2. Antes de que el usuario comience a hacer su búsqueda mediante el sistema, identifica algunos documentos útiles que sabe existen en la base de datos, el ALCANCE sería la proporción de esos documentos que se han recuperado durante un proceso normal. Pero es raro que un usuario pueda anticipar una cantidad significativa de documentos útiles, es más fácil para el ser humano reconocer documentos que recordarlos

3. Tomar una muestra aleatoria de la colección y evaluar la utilidad de los documentos en la muestra, para, posteriormente, utilizar técnicas estadísticas para estimar el número de documentos útiles existentes en la colección, basándose en aquellos encontrados en la muestra. El problema aquí es que las técnicas estadísticas no son adecuadas para el ambiente de la recuperación de información cuando las colecciones alcanzan tamaños muy grandes y el conjunto de documentos útiles puede ser menor que el uno porciento de la colección y para estimarlo el tamaño de la muestra debería ser demasiado grande, como para que ésta sea revisada.

4. Seleccionar muestras de documentos no recuperados mediante el uso de marcos de muestreo. Mediante esta técnica se extraen muestras de documentos no recuperados de subconjuntos de la base de datos que se sabe tienen alta proporción de documentos relevantes. Es más probable encontrar documentos relevantes en estos subconjuntos que en la colección entera y podrían tomarse pequeñas muestras que conservarían altos niveles de confianza estadística. Si estos subconjuntos son cuidadosamente seleccionados, y abarcan la parte de la colección que posiblemente contenga documentos útiles, entonces la estimación del ALCANCE se estaría maximizando, obteniendo con esto una buena aproximación a su valor real, además bajo tales circunstancias el valor real del ALCANCE dificilmente rebasará al valor estimado mediante este método. El problema aquí es que no es tan fácil encontrar los marcos de muestreo que contengan la mayoría de documentos útiles no recuperados. veamos un ejemplo utilizando el método denominado de modificación lógica:

Consideremos una pregunta formal del tipo "A y B y C y D" donde A, B, C y D son términos índice de búsqueda usados para recuperar documentos, una forma en que podríamos obtener subconjuntos "ricos" en documentos sería modificando la estructura lógica de la pregunta, de la siguiente manera: creando cuatro preguntas distintas basadas en los cuatro términos originales:

A y B y C y no D

A y B y D y no C

A y C y D y no B

B y C y D y no A

Los marcos de muestreo y el conjunto recuperado serían ajenos como se muestra en la figura 1.7.

Debido a la diversidad de maneras en que se pueden expresar las ideas mediante el lenguaje natural, entonces es muy probable que aparezcan documentos relevantes en estos nuevos conjuntos, si la cantidad de ellos es considerable, pueden tomarse muestras aleatorias de ellos, para que sean evaluadas por el usuario que generó la pregunta y de esta forma poder estimar la cantidad de documentos relevantes que no fueron recuperados.

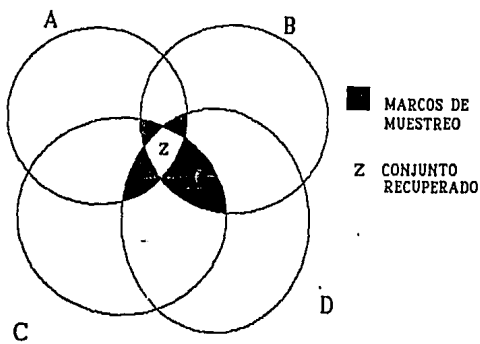


FIG 1.7

Una manera alternativa de generar marcos de muestreo es modificar el contenido semántico de la pregunta original, así se generan conjuntos que se revisan, y aquellos con mayor cantidad de documentos relevantes no recuperados originalmente, sirven como marcos de muestreo y las muestras son sometidas a juicio del usuario para evaluar la proporción de documentos relevantes no recuperados.

Como se dijo al principio de esta sección, el concepto de relevancia es subjetivo y esa es la razón por la que a pesar de los intentos mostrados, no ha sido posible aún cuantificar qué tan efectivo es un sistema de este tipo, como se ha visto, tales intentos no proporcionan una solución adecuada al problema de evaluar la eficacia.

CAPITULO 2

OTROS ASPECTOS RELEVANTES SOBRE INDEXACION

2.1 FACTORES DE LENGUAJE QUE AFECTAN EL PROCESO DE INDEXACION.

Un lenguaje de indexamiento es la colección total de términos índice usados en el conjunto de documentos por indexar.

Un articulo (item) en una colección denota una unidad en la colección. Por ejemplo, un artículo, en una colección puede ser un libro en una biblioteca, un artículo en un diario, un reporte técnico en una base de datos. Indexamiento es un proceso analítico que consiste en la identificación y selección de los conceptos representados, el propósito y contenido del documento, y de la representación de estos conceptos por medio de términos aceptables para el sistema de recuperación. Como hemos visto el indexamiento se caracteriza por dos tipos distintos de actividades; análisis de contenido y selección de términos.

En el primer capítulo se habló sobre la forma en que mediante una computadora, se podría realizar el análisis automático de textos, una de las primeras aproximaciones para realizar esta tarea fue la siguiente:

Una vez que se se ha realizado la eliminación de palabras no significativas, se procede a lo que se conoce como truncación de palabras, lo cual consiste en la eliminación de sufijos (terminaciones), con el fin de obtener la "raíz" de la palabra, por ejemplo, biblia, bibliografía y biblioteca, pueden tener como prefijo la cadena de letras "bibli(o)", así podemos afirmar que muchas palabras que son equivalentes en ese sentido, adquieren la misma forma al remover sus sufijos. De esta manera se pueden agrupar

palabras equivalentes con el fin de que el usuario pregunte por la raíz de la palabra en lugar de tener que hacer una pregunta para cada una de las derivaciones de esa raiz (modificadores de número, género, tamaño, etc.). La hipótesis fundamental (en el contexto de recuperación de información), es que si dos palabras tienen la misma raíz, entonces se refieren al mismo concepto.

Desafortunadamente no con todas las palabras sucede así: radiología y radiografía tienen como raíz, radio; que a su vez es un prefijo de orígen latino cuyo significado es rayo, pero radio también es una palabra que denota un elemento químico, o un hueso del antebrazo, o el segmento de recta que une un punto de una circunferencia con su centro, o un aparato radiorreceptor. De tal manera radiologia, radiodifusora, radioactivo y radio, tienen la misma raíz pero no son palabras equivalentes. De hecho existen palabras que siendo esencialmente equivalentes, pueden significar diferentes cosas en diferentes contextos. Sin embargo la truncación de palabras no representa una mala alternativa para el análisis de contenido de documentos, así que existen sistemas de recuperación de información que llevan a cabo este proceso, tanto para la indexación, como para la formulación de preguntas formales. El resultado final de este proceso es un conjunto de clases para cada raíz detectada. Un nombre de clase es asignado a un documento si y solo si uno de sus miembros ocurre como palabra significativa en el texto del documento.

El desempeño de los sistemas de recuperación de información, es expresado frecuentemente en términos del alcance (recall) y la precisión (v. capítulo 1). Esto es importante al seleccionar el metodo de indexación, que sea más apropiado para producir el grado de alcance y precisión deseados. La hipótesis predominante es que un buen método de

indexamiento, proporciona la clave para el desempeño adecuado de los sistemas de recuperación de información.

Tradicionalmente, la representación de documentos es un proceso de análisis de contenido y selección de términos para proporcionar acceso a ese documento, cada documento es representado por términos índice apropiados, los cuales son asignados en base a sus similitudes semánticas a la clase identificada. De aquí que en lugar de buscar en un amplio documento, se utiliza un archivo de términos índice que puede ser ordenado y manipulado. Como cada término índice está asociado con números únicos de documentos, los documentos pueden ser identificados. Los archivos índice de autores, títulos, números de reportes, fórmulas químicas y números de seguridad social, pueden proporcionar otros puntos de acceso y son útiles a los usuarios al momento de recuperación.

El método tradicional de indexamiento, opera bajo el supuesto de que la relevancia existe entre los términos índice elegidos, con respecto al documento que representan. Hoy en día, las actividades para la representación de documentos, están dominadas por este acceso semántico a la relevancia. Similarmente, para recuperar el documento, funciona el proceso dual: (a) el usuario identifica los conceptos encontrados en la pregunta y (b) procede a seleccionar los términos índice que representan estos conceptos. Se aplica el mismo principio de relevancia basado en la relación semántica entre los términos elegidos y los conceptos que la pregunta involuca.

En el análisis de contenido de un documento, no todas las ideas, tópicos y conceptos son seleccionados para ser indexados. Algunos conceptos son centrales al documento, algunos son mencionados de paso; algunos son nombrados por

interés histórico. La selección de las ideas es un hecho subjetivo.

En los últimos años, se han desarrollado muchos sistemas de recuperación de texto completo (Full-text), en los cuales no es necesaria la indexación humana. Cada documento es indexado automáticamente por cada palabra no trivial en el texto.

Un sistema de indexamiento automático está basado en el supuesto de que los términos usados en el texto representan óptimamente los conceptos en el documento. El interés es eliminar el proceso de indexamiento manual con su correspondiente alto costo. Desafortunadamente aunque alcanza consistencia, el indexamiento de palabras en texto-libre (free-text) crea problemas de sinónimos y homónimos. Por ejemplo, un medicamento específico puede ser conocido por varios nombres. Para asegurar cada posibilidad de búsqueda, debe designarse cada sinónimo a una sola entidad. Por otro lado, palabras para las cuales existen homónimos, tales como banco, papel y peso, son asociados con diferentes significados dependiendo del contexto. Usar este tipo de términos de búsqueda en lenguaje natural podría recuperar documentos no relevantes.

2.2 ESPECIFICIDAD, EXHAUSTIVIDAD Y DENSIDAD.

Tres conceptos importantes son especificidad, exhaustividad y densidad de indexamiento. Cada concepto es ligado con el alcance y la precisión del lenguaje de indexamiento usado.

Especificidad. La especificidad es una medida del grado de precisión con el cual los temas o tópicos de los documentos pueden representarse por medio del lenguaje de indexamiento. Un alto nivel de especificidad denota un alto grado de precisión, con el cual los términos índice son usados en la

descripción de los conceptos contenidos en un documento. Por ejemplo, si el tema de los documentos está relacionado con "sistemas expertos", y si el término índice "sistemas expertos" es un descriptor admisible, hay una correspondencia precisa del concepto y del término. Por otro lado, falta especificidad si el término índice más cercano al concepto "sistemas expertos" es "inteligencia artificial", este término no es específico ya que sistemas expertos son una forma de aplicación de la inteligencia artificial.

En términos del desempeño de la recuperación, la precisión es el porcentaje de los documentos relevantes contenidos en el conjunto obtenido o recuperado. Con términos índice altamente especificos, cada término índice puede cubrir un amplio dominio de tópicos, de los cuales no todos están relacionados al área específica deseada. El conjunto de documentos obtenidos podría ser muy amplio, y más documentos no relevantes podrían ser incluídos. La Precisión del Sistema se pierde. Al mismo tiempo, como se obtienen más documentos con menor relevancia, algunos de esos documentos pueden contener información pertinente al tópico buscado. Consideremos el ejemplo anterior en el que requeríamos el tópico sistemas expertos. Bajo inteligencia artificial, puede encontrarse información relevante sobre una ciencia conocida que tenga importantes implicaciones para la construcción de sistemas expertos, la especificidad del lenguaje de indexamiento es el factor más importante que afecta la precisión en la búsqueda.

Exahustividad. La exhaustividad de indexamiento se refiere al grado de indexamiento en cuanto al alcance de los tópicos encontrados en el documento. Es una medida de cuales de todos los distintos temas ó tópicos discutidos en un documento dado son analizados, reconocidos e indexados. En

cualquier publicación, el documento frecuentemente trata más de un concepto ó tema. Naturalmente algunos están dando más énfasis y son más centrales al documento. La tarea del indexador es descubrir los conceptos y decidir cuales deberán ser tomados en cuenta. Un proceso exhaustivo de indexamiento representa un intento por indexar todos los conceptos contenidos en el documento.

La exhaustividad en el indexamiento es también una consideración en términos de alcance y precisión. Cada término índice sirve como una etiqueta para un tema o concepto en el documento. Si un término es asignado a un documento, es también una representación de información para el documento. Si cada faceta de un documento fuese indexada y se usaran más de 30 términos índice para representarlo, entonces una búsqueda con cualquiera de esos términos deberá ser capaz de recuperar el documento. Con el indexamiento exhaustivo, existe una alta probabilidad de que más de los documentos relevantes, representados por los términos índice puedan ser recuperados. El indexamiento exhaustivo asegura alto alcance. Es decir unalto nivel de exhaustividad lleva a muchas "llamadas" pero poca precisión.

Un documento puede tratar dos ó tres conceptos, que en el indexamiento exhaustivo podría solamente producir unos cuantos términos índice. Supongamos un documento que representa cinco diferentes conceptos; es concebible indexar exahustivamente diferentes aspectos de tres de los cinco temas con muchos términos e ignorar los otros dos. En casos de este tipo, el número de términos asignados no es un indicador preciso del grado de exhaustividad. Por lo que puede resultar engañoso medir la exhaustividad por el número promedio de términos asignados por documento.

Densidad de indexamiento. Frecuentemente la frase densidad de indexamiento es usada intercambiablemente con exhaustividad de indexamiento, es el número promedio de términos índice seleccionados para representar cada documento. Si hay sólo dos conceptos que se discuten en un documento dado, asignar veinte términos no será más exhaustivo que usar dos términos apropiados para cubrir el contenido del documento. Obviamente, es difícil checar el grado de exhaustividad entre un grupo de indexadores ó monitorear el mismo índice durante un período de tiempo. Es igualmente difícil producir una medida cuantitativa para obtener el grado de exhaustividad. La Densidad de indexamiento es puramente un estimado de exhaustividad. Los indexadores experimentados .pueden llevar a cabo un grado deseado de exahustividad, dado un límite superior de número de términos índice permitidos.

2.3 INDICES POR PALABRA CLAVE EN CONTEXTO Y FUERA DE CONTEXTO.

Para propósitos de recuperación de información, en ocasiones es conveniente tener listas textuales producidas por computadora con la descripción de los documentos (títulos, resúmenes o frases), arrreglados de tal forma que todos los descriptores que contengan alguna palabra en particular puedan ser determinados rápidamente por una inspección manual sobre la lista. Un arreglo conveniente es conocido como índice de tipo PALABRA CLAVE EN CONTEXTO (KWIC por sus siglas en inglés *keyword in context*). A continuación explicamos con un ejemplo la forma de índice KWIC.

Consideremos el siguiente conjunto de títulos (los números son para propósitos de identificación).

1876 Las computadoras electrónicas y la legislación.
3048 Algunos aspectos legales de las computadoras.
1498 Procesamiento de estatutos legales por computadora.

6577 Leyes sobre patentes relativas al uso de las computadoras.
4885 Programas de cómputo y la ley sobre patentes.

En la siguiente lista, la cual está en forma de índice KWIC, cada título se repite varias veces con desplazamientos y truncaciones apropiadas, de tal modo que el comienzo de cada palabra significante en el título puede ser encontrada buscando sobre cierta columna localizada aproximadamente sobre la posición central de la página. La lista se ordena alfabéticamente por la palabra colocada en el centro.

```
1876                    Las   computadoras electrónicas y la legislación.
3048  ctos legales de las     computadoras.
1498 tatutos legales por      computadora.
6577 tivas al uso de las      computadoras.
4885         Programas de     cómputo y la ley sobre patentes.
1876      Las computadoras    electrónicas y la legislación.
1498     Procesamiento de     estatutos legales por computadora.
3048      Algunos aspectos    legales de las computadoras.
1498  iento de estatutos      legales por computadora.
1876  s electrónicas y la     legislación.
4885  as de cómputo y la      ley sobre patentes.
6577                          Leyes sobre patentes relativas al uso de las
4885  uto y la ley sobre      patentes.
6577         Leyes sobre      patentes relativas al uso de las
computadoras
1498                          Procesamiento de estatutos legales por compu
4885                          Programas de cómputo y la ley sobre
patentes.
```

La parte de índice KWIC que contiene las palabras colocadas centralmente se conoce como ranura. Las palabras en la ranura son precedidas usualmente por espacios en blanco, con el fin de hacerlas fácil de localizar.

En el índice KWIC anterior las palabras no significativas no se toman en cuenta para hacer la separación y el ordenamiento.

Un índice KWIC puede ser un medio costoso de representación de los títulos, ya que si cada uno de ellos contiene en promedio N términos significativos, el almacenamiento en forma de índice KWIC implica un factor de expansión igual a

N. Sin embargo, un índice KWIC impreso puede ser explorado manualmente de manera rápida y es, por lo tanto, una ayuda útil para localizar títulos dentro de una colección relativamente pequeña. Una ventaja adicional es que una vez que un índice KWIC ha sido impreso, se encuentra disponible sin depender de la computadora.

El índice por palbra clave fuera de contexto (KWOC por sus siglas en inglés *keyword out of context*), es similar al índice KWIC, con la diferencia que las palabras que forman la ranura son colocadas a un lado del texto, así, no se requiere ningún desplazamiento. El siguiente índice KWOC corresponde al índice KWIC mostrado anteriormente.

```
1876 computadoras    Las computadoras electrónicas y la legislación.
3048 computadoras    Aspectos legales de las computadoras.
1498 computadora     Procesamiento de estatutos legales por computadora.
6577 computadoras    Leyes sobre patentes relativas al uso de las
computad
4885 cómputo         Programas de cómputo y la ley sobre patentes.
1876 electrónicas    Las computadoras electrónicas y la legislación.
1498 estatutos       Procesamiento de estatutos legales por computadora.
3048 legales         Algunos aspectos legales de las computadoras.
1498 legales         Procesamiento de estatutos legales por computadora.
1876 legislación     Las computadoras electrónicas y la legislación.
4885 ley             Programas de cómputo y la ley sobre patentes.
6577 Leyes           Leyes sobre patentes relativas al uso de las
4885 Patentes        Programas de cómputo y la ley sobre patentes.
6577 patentes        Leyes sobre patentes relativas al uso de las
computad
1498 Procesamiento   Procesamiento de estatutos legales por computadora
4885 Programas       Programas de cómputo y la ley sobre patentes.
```

2.4 VOCABULARIO CONTROLADO.

Un **vocabulario controlado** en sistemas de recuperación de información es el conjunto de términos índice que pueden usarse para indexar documentos, así como para formular preguntas formales para la recuperación de los mismos.

Un vocabulario controlado tiene como propósito cumplir dos objetivos:

1.- Realizar lo mejor posible la representación del contenido de los documentos al tiempo de indexamiento y satisfacer el criterio de predicción al tiempo de búsqueda.

2.- Facilitar las búsquedas en el sistema, al mantener relacionados, de algún modo, los términos más cercanos en cuanto a sus significados.

Para llevar a cabo estos objetivos, el vocabulario se controla a través de la elección de un término elegido de entre un grupo de términos con significados similares. El indexador y el usuario son guiados al término elegido por su uso o referencias. Un término es escogido porque es el más frecuentemente usado por los usuarios del sistema, ó por ser considerado el término más apropiado para el concepto. Además, el vocabulario controlado debe clasificar las palabras polisémicas. Hay palabras con diferentes significados, escritas exactamente de la misma forma. Por ejemplo, "puro" puede significar un tipo de cigarro, ó denotar algo sin mezcla, ó castidad.

Un buen vocabulario controlado muestra las relaciones semánticas entre los términos. El vocabulario puede dar indicios para hacer mejores formulaciones de búsqueda para la recuperación de un número máximo de documentos relevantes. Por ejemplo, si se necesita un alto porcentaje de documentos relevantes para un tópico dado, el vocabulario puede llevar al usuario a todos los términos relacionados así como a todos los términos generales y específicos asociados con el tópico para una búsqueda genérica.

El tipo de vocabulario controlado más antiguo que existe es la clasificación. La clasificación es el agrupamiento de entidades de acuerdo a sus atributos y reglas. Las clasificaciones son diseñadas para organizar cosas que sirven a algún propósito específico. Por ejemplo, los libros

son clasificados por su contenido. Los esquemas de clasificación para libros son frecuentemente construídos de un arreglo lógico del universo de conocimiento. Intentan ser comprensibles, consistentes, flexibles y expandibles lo suficientemente para adaptar y para anticipar cambios.

Se han desarrollado procedimientos similares a los conceptos usados en las clasificaciones. Aunque las funciones básicas se mantienen, estos nuevos mecanismos de Control de Vocabulario más estructurados, son conocidos como *Thesaurus* ó *Thesauri* (Plural en Inglés de *Thesaurus*).

Un ***thesaurus*** es un vocabulario controlado de términos relacionados semántica y genéricamente que cubre una área específica del conocimiento.

Para construir un *thesaurus* es necesario saber cuál será el dominio en que funcionará. Una de las más importantes consideraciones de un *thesaurus* es su constante necesidad de actualización y revisión, y el soporte de manejo en su creación y su mantenimiento deben asegurarse antes de iniciar su construcción. A continuación describimos una forma de construirlo.

Generación de términos.- Los términos son recolectados de tantas fuentes relevantes como sea posible. Fuentes potenciales son los diccionarios especializados, glosarios, términos índice encontrados en los índices por materia de los libros, y bibliografías temáticas.

Categorización de términos por temas.- Después que el grupo de descriptores ha sido seleccionado para representar el dominio del tema, se categorizan lógicamente en subgrupos de acuerdo a su contenido. El objetivo es ligar los términos relacionados semánticamente dentro de cada grupo para una

fácil recuperación. Similarmente, ciertos términos de diferentes grupos deben ser también conectados.

Ordenamiento jerárquico de descriptores.- El mayor problema en desarrollo de un thesaurus como auxilar en el proceso de recuperación de información, es cómo derivar las relaciones semánticas sobre las cuales el *thesaurus* es construído. Básicamente estas relaciones semánticas pueden ser: Normativas ó descriptivas. Las relaciones normativas son regulativas o ideales y se establecen por convención o acuerdo, podemos tomar un ejemplo de la Biología, en un thesaurus hipotético colocaríamos la descripción "ballena" en relación subordinada (más concreta que) a la descripción "mamífero", de acuerdo a las clasificaciones biológicas establecidas, pero "ballena" también podría ser colocada en relación subordinada a "criaturas marinas", "animales grandes", etc. Así, las clasificaciones normativas imponen un punto de vista particular sobre un sistema de relaciones entre términos o clases de objetos a los cuales se refieren. Tal sistema regulativo puede ocasionar problemas cuando se han establecido relaciones entre términos de manera diferente a la que establecerían los usuarios, esta problemática lleva a la conclusión de que un thesaurus útil a la recuperación de información debe ser descriptivo, en este caso las relaciones entre términos se derivan de las representaciones de los documentos en la base de datos, la manera más obvia de derivar relaciones semánticas entre documentos es tabular la frecuencia de co-ocurrencia entre términos, es decir, tomar en cuenta el número de veces que dos términos cualesquiera son utilizados juntos para representar el contenido de documentos similares. La hipótesis fundamental es que si dos descriptores son frecuentemente utilizados juntos para representar documentos, entonces estos dos descriptores están relacionados semánticamente.

Una de las propiedades interesantes de algunas relaciones semánticas es que son transitivas. Esto significa que no solamente tenemos las relaciones que se dan explicitamente con la co-ocurrencia, ya que se puede inferir la existencia de otras relaciones implícitas al suponer la transitividad. Por ejemplo si tenemos que el término A co-ocurre con el término B y el término B co-ocurre con el término C, podemos inferir que probablemente existe una relación semántica entre A y C, sin tomar en cuenta el hecho de que no co-ocurren.

Enriquecimiento de los términos de entrada. Para facilitar al usuario el acceso a descriptores localizados en el *thesaurus*, es conveniente mostrar un amplio número de términos por los cuales el usuario puede preguntar. Estos son conocidos como **entradas**, su papel es dirigir rápidamente al usuario a los términos usado por el sistema, para indexar documentos.

Factores que influyen en el mantenimiento de un *Thesaurus*. Para ser un mecanismo efectivo de control de vocabulario, el thesaurus debe ser constantemente actualizado. Se debe agregar nuevos tópicos y descriptores. Por ejemplo, "reproductores de sonido digital" y "discos compactos" son términos nuevos introducidos en los años recientes, de igual manera, el desuso de términos antiguos implica la eliminación de descriptores obsoletos.

Finalmente enunciamos algunas características adicionales que debe tener un *thesaurus*:

- Un *thesaurus* debería incluir solamente aquellos términos que probablemente sean de interés para la identificación del contenido en un tema determinado (por ejemplo, la palabra "mano" puede usarse en un *thesaurus* que trate sobre biología, pero no debería ser incluído si su frecuencia de

ocurrencia es demasiado alta en expresiones como "información de primera mano".

- Los términos ambiguos sólo deberían ser incluídos para el sentido que pudiera ser importante en la colección de documentos (al menos dos categorías de *thesaurus* podrían ser usadas para términos tales como "campo", una correspondiente al área de conocimiento y la otra a su sentido técnico en álgebra, en un thesaurus que trata sobre matemáticas, no es necesaria una categoría para el sentido de la palabra "campo" cuando se refiere al agro o a algún país).

2.5 CONSTRUCCION AUTOMATICA DE *THESAURUS*.

La construcción automática de *thesaurus* se basa en el uso de vectores de documentos de la forma:

$D_i = \langle d_{i1}, d_{i2}, \ldots, d_{it} \rangle$

donde cada dij es el peso asignado a el j-ésimo identificador del documento. Por ejemplo, si existen tres términos ALFA, BETA Y GAMMA, entonces

$D_1 = \langle 2,4,0 \rangle$

significa que el documento número 1 está identificado por el término ALFA con un peso de 2, BETA con un peso igual a 4 y GAMMA con unpeso de 0. La longitud t del vector corresponde a el número de términos distintos asignados a la colección total y los pesos con valor cero significan que esos términos no están asignados al vector de documento dado. Entonces, una colección de documentos es representable por una matriz como la siguiente:

$$\begin{array}{c|cccc} & T_1 & T_2 & & T_t \\ D_1 & d_{11} & d_{12} & \ldots & d_{1t} \\ D_2 & d_{21} & d_{22} & \ldots & d_{2t} \\ \vdots & & \vdots & & \\ D_{n1} & d_{n1} & d_{n2} & \ldots & d_{nt} \end{array}$$

Podemos utilizar una función de similaridad $S(D_i, D_j)$ que refleje las similaridades de términos índice, ésto lo hacemos mediante la comparación de pares de renglones de la matriz de documentos. Mientras los renglones de la matriz representan vectores de documentos individualmente, las columnas identifican la asignación de términos a los documentos. Esto es, una columna, j, de la matriz de vectores de documentos, refleja la asignación del término j a los documentos de la colección. Dados dos vectores de términos de la forma $T_k = (t_{1k}, t_{2k}, \ldots, t_{nk})$ donde t_{ik} indica el peso o valor del término k en el documento i y suponiendo n documentos en la colección, una medida de similaridad puede ser definida como sigue:

$$S(t_k, t_h) = \sum_{i=1}^{n} t_{ik}\, t_{ih}$$

o utilizando un factor de normalización para limitar los resultados calculados a valores entre 0 y 1,

$$S(t_k, t_h) = \frac{\sum_{i=1}^{n} t_{ik}\, t_{ih}}{\sum_{i=1}^{n} (t_{ik})^2 + \sum_{i=1}^{n} (t_{ih})^2 - \sum_{i=1}^{n} t_{ik}\, t_{ih}}$$

Cuando todos los pares de columnas distintas de la matriz son comparados con cada uno de los otros, se construye una matriz T de similaridad término a término, en la cual el elemento ubicado en el renglón k y la columna h es igual a $S(t_k, t_h)$.

$$\begin{array}{c|cccc} & T_1 & T_2 & \cdots & T_t \\ T_1 & S(T_1,T_1) & S(T_1,T_2) & \cdots & S(T_1,T_t) \\ T_2 & S(T_2,T_1) & S(T_2,T_2) & \cdots & S(T_2,T_t) \\ \vdots & \vdots & & & \vdots \\ T_t & S(T_t,T_1) & S(T_t,T_2) & \cdots & S(T_t,T_t) \end{array}$$

A partir de esta matriz se pueden elaborar métodos de classificación automática, para construir clases de términos similares (categorías de thesaurus), colocando en una misma clase todos los términos cuyo coeficiente de similaridad sea suficientemente grande. Por ejemplo la similaridad entre T_k y todos los miembros de su clase debe exceder un número estipulado (umbral)

Los métodos de clasificación asumen la pre-existencia de clases de términos y proceden refinando el estado inicial de la clasificación, por ejemplo: una clase de términos determinada puede ser definida como el conjunto de términos asignados a un documento en particular, o conjunto de documentos; ésto genera un número de clases de términos inicial igual al número de documentos utilizado como conjunto de arranque.

Para cada clase existente, se puede definir un centroide (C) tal que

$$C = (\bar{t}_1, \bar{t}_2, \ldots, \bar{t}_m)$$

donde C es el vector promedio de los vectores de términos de esa clase. Esto es, cada término del centroide se define como el valor promedio de todos los valores de T_k en los documentos individuales de la clase:

$$\bar{t}_k = \frac{1}{m}\sum_{i=1}^{n} t_{ik}$$

para una clase con m vectores de términos. El refinamiento de las clases de términos consiste ahora en calcular la similaridad entre cada vector de términos T_k y cada centroide de clase para todas las clases existentes. Suponiendo t vectores de términos y p clases, el proceso requiere la generación de t x p coeficientes de similaridad $S(T_k, C_h)$ para k entre 1 y t, y h entre 1 y p. Cada vector de términos es ahora incluído en la clase cuya similaridad con respecto al centroide sea más grande. Si ésto implica el paso de un vector de términos dado de una clase a otra, los centroides de esas clases deben ser recalculados.

De esta manera las posibilidades de búsqueda para el usuario pueden ampliarse, al tener acceso a los términos que están fuertemente asociados con los términos de su pregunta.

CAPITULO 3

ALGUNOS MODELOS FORMALES PARA LA RECUPERACION DE INFORMACION.

Existen varios diseños propuestos como estructuras lógicas de un sistema de recuperación de información, creemos conveniente presentar las características y diferencias de algunos de ellos y ese es el propósito del presente capítulo En la figura 3.1 se muestra con más detalle la estructura básica de los sistemas que estamos estudiando:

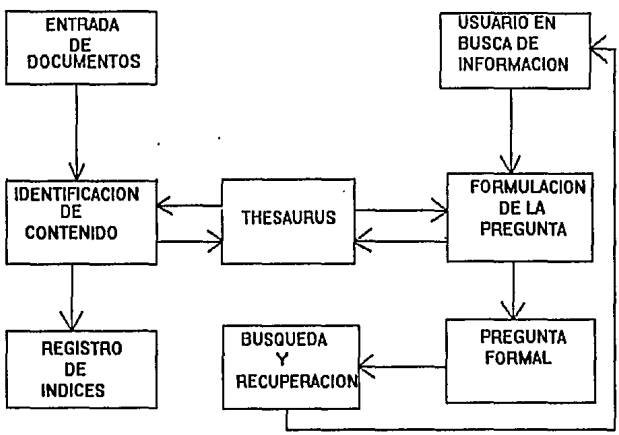


FIG. 3.1

los modelos que habremos de presentar, describen las principales formas en que dichos sistemas pueden operar. De esta manera podremos establecer comparaciones útiles entre ellos e identificar la problemática que se presenta en el diseño de tales sistemas.

3.1 MODELO 1

1. Cada pregunta consta de un solo descriptor.

2. Los documentos son descritos mediante un conjunto de descriptores (términos índice)

3. Regla de recuperación: si el descriptor de la pregunta es un elemento del conjunto de descriptores del documento, el documento es recuperado.

Por ejemplo: Pregunta = D_k

Documento A = $\{D_a, D_b, D_c\}$ no recuperado

Documento B = $\{D_b, D_k, D_m\}$ recuperado

El modelo 1 es el más simple en recuperación de documentos, a causa de ello es uno de los más comunes y precisamente esa es su mayor ventaja, pero debido a que utiliza un sólo descriptor como pregunta formal, los resultados de la recuperación tienden a ser deficientes cuando el tamaño de la colección aumenta, es decir, se pueden recuperar tantos documentos, que el usuario no podrá revisarlos en su totalidad. Además este modelo no permite hacer búsquedas a partir de varios descriptores.

3.2 MODELO 2

1. Pregunta: Un conjunto de descriptores.

2. Los documentos son descritos mediante un conjunto de descriptores

3. Regla de recuperación: un documento es recuperado si todos los descriptores que forman la pregunta pertenecen al conjunto de términos índices asignados al documento.

Ejemplo: Pregunta = D_k y D_j

Documento A = $\{D_a, D_k, D_m\}$ no recuperado

Documento B = $\{D_a, D_k, D_j\}$ recuperado

El modelo 2 es ligeramente mejor que el modelo 1, ofrece al usuario mayor flexibilidad para la formulación de preguntas, ya que permite incluir en cada una de ellas varios descriptores, lo cual es deseable en bases documentales grandes, porque permite superar el problema del punto inútil.

Podría objetarse que este modelo no permite el manejo de preguntas disyuntivas, no es así, una pregunta de este tipo como A U B podría ser transformada en $(A' \cap B')'$. Aunque esto podría no ser tan fácil de implementar para el usuario.

Los dos modelos anteriores recuperan documentos solamente si y solo si la totalidad de los descriptores que forman la pregunta fueron asignados a cada uno de los documentos recuperados. El siguiente modelo difiere un poco, veamos:

3.3 MODELO 3

1. Pregunta: un conjunto de índices más un valor de corte.
2. Los documentos son descritos mediante un conjunto de descriptores
3. Regla de recuperación: un documento es recuperado si comparte con la pregunta un conjunto de índices cuyo tamaño es mayor que el valor de corte.

Ejemplo: Pregunta = D_k y D_j y D_m

Valor de corte = 1

Documento A = $\{D_a, D_k, D_o\}$ no recuperado

Documento B = $\{D_a, D_k, D_j, D_m\}$ recuperado

Documento C = $\{D_b, D_c, D_j, D_m\}$ recuperado

El modelo 3, a diferencia de los dos primeros, permite que sean recuperados documentos cuando sus descriptores asignados coinciden con un subconjunto del conjunto de términos de búsqueda especificados, lo cual significa que la comparación entre la pregunta y el conjunto de descriptores asignado a los documentos recuperados no necesita ser

exácta. Este modelo permite al usuario expander ó limitar su búsqueda modificando el valor de corte. Por ejemplo: el usuario podría empezar introduciendo la misma pregunta del modelo 3 pero con un valor de corte igual a 2 en lugar de 1, esta pregunta recuperará solamente aquellos documentos a los cuales los descriptores D_k, D_j y D_m han sido asignados. Si el usuario desea hacer una búsqueda más exhaustiva, podría volver a someter su pregunta con valor de corte igual a 1, esto haría que el sistema recupere todos los documentos que contengan cualquiera de las siguientes combinaciones de descriptores:

$D_k \cap D_j \cap D_m$

$D_k \cap D_j$

$D_k \cap D_m$

$D_j \cap D_m$

Observemos que los documentos recuperados con $D_k \cap D_j \cap D_m$ serán los mismos que aquellos recuperados en la búsqueda previa.

Llevando a cabo las combinaciones de términos de búsqueda automáticamente, el modelo 3 elimina en gran medida el trabajo de la reformulación de preguntas y consecuentemente, se hace más amable la interfase con el usuario. Pero esta flexibilidad da al mismo tiempo una desventaja para este modelo; un usuario ingenuo podría incluir un gran número de descriptores en su pregunta junto con un valor de corte muy bajo. De hecho, para una pregunta de n descriptores y un valor de corte igual a r, el número de subconjuntos-pregunta que debe generar el sistema es:

$$\binom{n}{r}+\binom{n}{r-1}+\ldots+\binom{n}{1}$$

Por ejemplo si se incluyen 7 descriptores en la pregunta y un valor de corte igual a 3 tendríamos que el número de subconjuntos-pregunta está dado por:

$$\binom{7}{3}+\binom{7}{2}+\binom{7}{1}=42+21+1=64$$

Otro problema que resulta de este procedimiento es el siguiente: muchos documentos serán recuperados más de una vez. Por ejemplo el documento que contiene los términos de búsqueda D_a, D_b, D_c, D_d con un valor de corte igual a 1 será recuperado por las siguientes combinaciones de términos de búsqueda generadas como preguntas:

$D_a \cap D_b \cap D_c \cap D_d$

$D_a \cap D_b \cap D_c$

$D_a \cap D_b \cap D_d$

$D_a \cap D_c \cap D_d$

$D_b \cap D_c \cap D_d$

$D_a \cap D_b$

$D_a \cap D_c$

$D_a \cap D_d$

$D_b \cap D_c$

$D_b \cap D_d$

$D_c \cap D_d$

Así pues, este documento podría aparecer 11 veces en el conjunto recuperado, lo cual sólo sirve para aumentar la cantidad de material irrelevante que el usuario debe revisar.

Existen dos maneras de eliminar esta redundancia:

1.- Un registro puede ser conservado cuando se detecte la primera ocurrencia de un documento, si bajo una nueva combinación de términos el documento vuelve a ser recuperado ya no será tomado en cuenta.

2.- Otro método es cambiar cada una de las combinaciones de términos a su forma normal conjuntiva completa, aclaremos ésto con un ejemplo:

$D_a \cap D_b \cap D_c \quad ---> \quad D_a \cap D_b \cap D_c - D_d$

El conjunto de documentos que satisfacen las combinaciones en esta forma serán, por definición, ajenos. Así, ningún documento será recuperado más de una vez. Obviamente esta alternativa no es muy práctica.

Existe otra desventaja del modelo 3, supone que todos los descriptores tienen el mismo valor o peso, esto significa que los documentos en los conjuntos recuperados no son categorizados. En otras palabras, para una pregunta de 5 descriptores con un valor de corte igual a 1, los documentos que coinciden con los cinco descriptores no están en un orden de importancia mayor que los documentos que contienen solamente dos descriptores.

El modelo 4 categoriza los documentos recuperados de acuerdo al grado de coincidencia entre los descriptores de documentos y el conjunto de términos de pregunta, así que los documentos que contengan a todos los descriptores de la pregunta serán categorizados en primer lugar, los documentos que coincidan con un término menos, en segundo lugar y así sucesivamente.

3.4 MODELO 4

1. Pregunta: un conjunto de índices más un valor de corte.
2. Los documentos son descritos mediante un conjunto de descriptores
3. Los documentos recuperados son categorizados.
4. Regla de recuperación: los documentos que comparten con la pregunta un número de descriptores mayor que el valor de corte son categorizados en orden decreciente.

Ejemplo: Pregunta = D_k y D_j y D_m

Valor de corte = 1

Documento A = $\{D_a, D_j\}$ - no recuperado

Documento B = $\{D_b, D_j, D_m\}$ recuperado. Categoría 2

Documento C = $\{D_a, D_j, D_k, D_m\}$ recuperado. Categoría 1

En este modelo, aunque el número de documentos recuperados puede rebasar el punto inútil del usuario, se pueden revisar las categorías más altas, ya que es de esperarse que que la categorización coloque a los documentos por grado de importancia, aunque, hay que aclarar, esto no tiene porqué ser siempre cierto. Veamos sus desventajas: en primer lugar se consume más tiempo de proceso debido a los procedimientos de categorización de los documentos recuperados. Además, al igual que en el modelo 3 muchos documentos podrían ser recuperados más de una vez. Otro problema es que todos los términos son considerados como si fueran igualmente importantes, lo cual no es recomendable ya que la mayoría de los usuarios podrían percibir algunos términos de pregunta como más importantes y consecuentemente, los documentos que los contienen deberían ser categorizados en un nivel más alto que los documentos recuperados que contienen el mismo número de términos, pero los cuales son menos importantes.

El modelo 5 permite al usuario asignar a cada término de búsqueda un valor de ponderación (peso) que representa la importancia relativa del término para realizar la búsqueda, los documentos recuperados serán categorizados de acuerdo ese peso.

3.5 MODELO 5 (Preguntas Ponderadas)

1. Pregunta: un conjunto de índices cada uno de los cuales tiene un número positivo (ponderación) asociado a él.

2. Los documentos son descritos mediante un conjunto de descriptores

3. Los documentos recuperados son categorizados.

4. Regla de recuperación: Los documentos son categorizados en orden decreciente con respecto a la suma de las ponderaciones de los índices comunes entre la pregunta y el documento.

Ejemplo: Pregunta = $D_k(7)$ y $D_j(4)$ y $D_m(2)$

CATEGORIA		VALOR
Ultima	Documento A = (D_a, D_b)	0
2a	Documento B = (D_a, D_k, D_o)	7
1a	Documento C = (D_k, D_j, D_o)	11
3a	Documento D = (D_b, D_j, D_m, D_o)	6

Debe tenerse cuidado de no otorgar demasiado peso a algunos términos y poco a los demás, por ejemplo:

Si $D_a(500)$, $D_b(10)$, $D_c(3)$, $D_d(2)$, habría dos grupos de categorías diferentes; los indexados por Da y los no indexados por Da. La presencia de este término en un documento opaca cualquier otra combinación de términos de búsqueda. Al asignar un peso tan grande a un sólo término, se estaría cayendo en algo muy parecido al modelo 1, es por eso que se debe tener cuidado en la elección del peso.

A menudo, durante los procedimientos de indexación, se debe tomar la decisión de asignar o no asignar (indexación binaria) un término índice (I_j), que describe sólo de manera secundaria al documento Di, que está siendo indexado, Si no se asigna el término al documento, entonces existe la posibilidad de que un usuario para el cual D_i es útil, use I_j como término de búsqueda, así D_i no sería recuperado, es decir el usuario no podría satisfacer el criterio de predicción. Por otro lado, si se asignan términos índice a documentos aún cuando sólo los describan de manera secundaria, es de esperarse que el tamaño de los conjuntos recuperados cuando esos términos se usen en una pregunta aumentará considerablemente, es decir, no se satisfará el criterio del punto inútil.

El modelo 6 proporciona una solución a este problema producto de la indexación binaria, permitiendo ponderar cada término con relación al documento al cual ha sido asignado.

3.6 MODELO 6 (Indexación Ponderada)

1. Pregunta: un conjunto de índices
2. Los documentos son descritos mediante un conjunto de descriptores cada uno de los cuales tiene un número positivo (ponderación) asociado con él.
3. Los documentos recuperados son categorizados
4. Regla de recuperación: Los documentos son categorizados en orden decreciente con respecto a la suma de las ponderaciones de los índices comunes entre la pregunta y el documento.

Ejemplo: Pregunta = D_k y D_j y D_m

CATEGORIA		VALOR
Ultima	Documento A = $\{D_a(7), D_b(3)\}$	0
1a	Documento B = $\{D_a(4), D_k(6), D_o(1)\}$	6
2a	Documento C = $\{D_a(1), D_k(4), D_o(6)\}$	4
3a	Documento D = $\{D_a(3), D_k(3), D_o(2)\}$	3

Aquí el problema es seleccionar la escala mediante la cual asignar peso a los términos, sin embargo la ponderación ofrece varias ventajas: permite categorías de recuperación más refinadas que los modelos de recuperación no ponderados, otra ventaja es que permite introducir preguntas que consten de un sólo término, aún si la amplitud de ese término rebasa el punto inútil, y como los documentos recuperados son categorizados de acuerdo al peso de sus términos asignados, los documentos con más peso aparecerán al principio del conjunto recuperado, entonces el usuario puede revisarlos hasta que el grado de utilidad vaya decreciendo y considere

que el conjunto recuperado ya no necesita ser examinado. Pero veamos las desventajas del modelo 6: el proceso de ponderación de términos índice puede incrementar el tiempo requerido para indexar nuevos documentos, aunque existen procesos de indexación automática que pueden aligerar éste problema, otra desventaja es que el usuario puede no estar de acuerdo en la magnitud del peso que se le ha dado a un término índice asignado a un documento en particular, de tal forma, el usuario no estará seguro que la primera parte del conjunto recuperado contenga documentos de utilidad.

El modelo booleano (modelo 7) es probablemente el más popular en sistemas de recuperación documental, permite construir preguntas formales a partir de combinaciones lógicas de términos índice, se ha utilizado como interfase básica en algunos sistemas, mejorándolo con esquemas de ponderación de términos índice, procedimientos de búsqueda asociativa, técnicas de retroalimentación, etc., pero en cualquiera de esos casos, se puede objetar que un usuario clásico, puede no conocer la lógica proposicional (Booleana) suficiente para formar preguntas, resultando demasiado complejo para tal usuario.

MODELO 7 (Preguntas Booleanas)

1. Las preguntas son cualquier combinación Booleana de índices, de los siguientes tipos:

$D_a \cap D_b$	Intersección
$D_a \cup D_b$	Unión
$-D_a$	Negación

2. Los documentos son descritos mediante un conjunto de descriptores

3. Reglas de recuperación:

Si la pregunta es $D_a \cap D_b$ se recuperan los documentos que tienen asignados a D_a y a D_b como términos índice.

Si la pregunta es $D_a \cup D_b$ se recuperan los documentos que han sido indexados por D_a o D_b (al menos uno de los dos)

Si la pregunta es $-D_a$ se recuperan todos los documentos que no tienen a D_a como término índice.

El modelo 8 (full text), presenta notables mejoras sobre el modelo anterior; en lugar de representar al documento mediante un conjunto de descriptores, el texto completo del documento es almacenado. Para efectos de búsqueda, se considera el conjunto de palabras significativas que contienen los documentos recuperables, durante la recuperación el usuario intenta predecir las palabras que aparecen en los documentos que le pueden ser útiles

MODELO 8 (Recuperación en texto completo)

1. Las preguntas son cualquier combinación Booleana de índices.

$D_a \cap D_b$	Intersección
$D_a \cup D_b$	Unión
$-D_a$	Negación

2. La búsqueda se realiza sobre el texto completo, excluyendo las palabras no significativas.

3. Reglas de recuperación:

Si la pregunta es $D_a \cap D_b$ se recuperan los documentos que contienen a D_a y a D_b
Si la pregunta es $D_a \cup D_b$ se reuperan los documentos que contienen a D_a o a D_b (al menos uno de los dos)
Si la pregunta es $-D_a$ se recuperan todos los documentos que no contienen D_a.

Los sistemas que trabajan bajo este modelo son afectados por cuestiones tales como el precio, potencialidad del equipo en cuanto a velocidad y almacenamiento, etc. lo cual implica mayores costos que los modelos anteriores, pero tiene la ventaja de que libera al diseñador de los problemas que ocasiona la indexación. Se argumenta que al almacenar el texto del documento en su totalidad, la indexación no es necesaria porque los usuarios son capaces de recuperar documentos útiles tratando de anticipar las palabras ó frases que aparecen en esos documentos.

Uno de los problemas que presenta el modelo de full text es cuando se contruye una pregunta; el usuario no solo debe seleccionar términos que cree aparecen en los documentos que desea recuperar, debe además, anticipar que esas palabras no aparezcan en documentos que no desea, es decir, el usuario debe ser muy sensitivo para hacer una adecuada elección de las palabras que formarán su pregunta, de otra manera, el sistema recuperará documentos no deseables. Así, la riqueza del lenguaje natural, hace que sea demasiado dificil predecir las palabras que contienen los documentos que se quieren y que además no ocurran en aquellos que son irrelevantes.

Otra consecuencia es que habrá demasiados términos disponibles en el vocabulario de búsqueda de un sistema de full text (vocabulario no controlado), lo cual incrementa de manera considerable el número de preguntas diferentes que pueden ser formuladas durante una búsqueda.

Una última desventaja del full text simple: al tratar de anticipar las palabras que fueron utilizadas en el documento que trata sobre un tópico en particular, es que la informacion necesaria para la recuperación no está contenida dentro del documento, esto es, a menudo existe una relación implícita entre el texto de un documento y el tema que éste aborda, es decir, el tema del documento no está explícitamente descrito en su texto.

Modelo 9 (Thesaurus Simple)

1. Las preguntas son descriptores simples

2. Los documentos son descritos mediante un conjunto de descriptores

3. Regla de recuperación: El término de la pregunta es buscado en un thesaurus y los términos semánticamente relacionados son agregados disyuntivamente.

Thesaurus binario típico:

	T_A	T_B	T_C	T_D	T_E	.	.	.	T_N
T_A	*	0	0	1	0	.	.	.	0
T_B		*	1	0	1	.	.	.	0
T_C			*	1	0	.	.	.	0
T_D				*	0	.	.	.	0
T_E					*	.	.	.	0
.									.
.									.
.									.
T_N									*

Cabe hacer la aclaración de que el modelo 8 es conocido como full text simple. Existen sistemas de recuperación de información full text a los que se han hecho adiciones enfocadas a mejorar la calidad de la recuperación. Un desarrollo realizado en esta área, es la construcción de un sistema de full text basado en los principios "coneccionistas" del diseño de sistemas. El gran problema de los sistemas de full text simple es que saturan al usuario con alternativas de búsqueda al representar a los documentos con un excesivo número de palabras y frases. Uno de los mayores obstáculos a que se enfrenta el usuario es que puede resultarle difícil pensar en todos los posibles términos de búsqueda que posiblemente han sido asignados a los documentos que desea. Es bien sabido que para el ser humano es más difícil recordar que reconocer, de tal manera que es más fácil reconocer descriptores que están semánticamente

relacionados, pero puede tener problemas al tratar de recordar esos mismos términos sin ninguna ayuda. El agregar un thesaurus a un sistema de recuperación de información puede ser útil ya que proporciona al usuario términos adicionales de búsqueda los cuales están relacionados semánticamente con los términos que son de su interés. Normalmente, esta implementación no es hecha automáticamente, pero podría haber una opción la cual sería seleccionada por el usuario si su conjunto inicial de términos de búsqueda no es suficiente al tratar de recuperar documentos. Un thesaurus es útil al proceso de búsqueda sólo si la adición de términos semánticamente relacionados a los iniciales es hecha selectivamente. El usuario no necesita todos los términos que se relacionan con su pregunta inicial, más bien necesita aquellos que según su criterio representen su necesidad de recuperación apropiadamente. Una manera de hacerlo es desplegar dichos términos con ayuda del thesaurus, entonces el usuario podrá elegir libremente los que considere convenientes agregar a su pregunta inicial. Así un thesaurus sirve más que nada, para que el usuario pueda recordar términos que le son útiles en su búsqueda.

Después de haber discutido las ventajas y desventajas de los más importantes modelos para la recuperación de información, cabe preguntarse ¿Cuál de ellos es el mejor?. Deben tomarse varios factores en cuenta, como son: el tamaño de la colección de documentos, el nivel de eficacia requerido, el grado de sofisticación deseado, etc. Por otro lado debe considerarse la posibilidad de desarrollar un sistema el cual combine dos o más de los modelos expuestos, a un sistema diseñado así se le conoce como híbrido. En todo caso, el determinar cual de todas las opciones con que se cuenta es la mejor, no es un problema fácil de resolver y puede variar para cada caso particular, en el Capítulo 1 se han expuesto algunos intentos realizados con el objeto de evaluar la eficacia de estos sistemas.

Para terminar este capítulo presentamos un cuadro sinóptico donde se pueden comparar los modelos expuestos anteriormente.

NUM	PREGUNTA	REPRESENTACION	REGLA DE RECUPERACIÓN	OBSERVACIONES
1	Un solo término índice	Documentos descritos por un conjunto de términos índice.	El descriptor de la pregunta debe pertenecer al conjunto de descriptores del documento.	Modelo de recuperación simple. El conjunto recuperado tiende a rebasar el punto inútil cuando el tamaño de la colección crece.
2	Un conjunto de términos índice.	Documentos descritos por un conjunto de términos índice.	Todos los descriptores que forman la pregunta deben pertenecer al conjunto de términos índice asignados al documento.	Ofrece mas posibilidades para formulación de preguntas que el modelo 1
3	Un conjunto de términos y un valor de corte	Documentos descritos por un conjunto de términos índice.	Los documentos recuperados comparten con la pregunta una cantidad de índices mayor que el valor de corte.	La comparación entre la pregunta y el conjunto de términos índice asignado a los documentos recuperados no necesita ser exácta.
4	Un conjunto de términos y un valor de corte	Documentos descritos por un conjunto de términos índice.	Los documentos recuperados comparten con la pregunta una cantidad de índices mayor que el valor de corte.	Los documentos recuperados son categorizados.
5	Un conjunto de términos índice con coeficientes de ponderación.	Documentos descritos por un conjunto de términos índice.	Se utiliza la suma de las ponderaciones de los índices comunes entre la pregunta y el documento.	Los documentos recuperados son categorizados.
6	Un conjunto de términos índice.	Documentos descritos por un conjunto de términos índice con coeficientes de ponderación.	Se utiliza la suma de las ponderaciones de los índices comunes entre la pregunta y el documento.	Los documentos recuperados son categorizados.
7	Combinación Booleana de términos índice.	Documentos descritos por un conjunto de términos índice.	Los documentos recuperados tienen términos índice que satisfacen la proposición lógica formada por la pregunta.	Es el modelo más popular. Gran flexibilidad para la formulación de preguntas formales.
8	Combinación Booleana de términos índice.	Búsqueda sobre texto completo.	Los documentos recuperados tienen términos índice que satisfacen la proposición lógica formada por la pregunta.	Lo afectan factores inherentes a la potencialidad del equipo. necesita un vocabulario controlado. Se debe almacenar el texto completo del documento.
9	Un solo término índice	Documentos descritos por un conjunto de términos índice.	Se agregan a la pregunta los términos índice asociados semánticamente con el término inicial.	Utiliza vocabulario controlado (thesaurus).

CAPITULO 4.
PROPUESTA DE NUESTRO SISTEMA.

4.1 DESCRIPCION GENERAL.

Una vez que hemos analizado de manera general (ya que un estudio profundo requiere de más tiempo y espacio) de la problemática de la recuperación de información, nos disponemos a presentar la manera en la cual pretendemos darle solución

Nuestro propósito es crear un sistema de recuperación de información no muy complejo, que sea de uso fácil y que pueda emplearse por cualquier persona, sin importar el área de conocimiento que se intente cubrir, ni el volumen de la colección de documentos.

Como se vió en el capítulo 3, existe una gran variedad de modelos para diseñar sistemas de recuperación de información que podrían ser implementados. En primera instancia, nuestro sistema no presenta muchas alternativas de búsqueda, ni mucho menos, pero puede servir como base para trabajos posteriores en los que se pueda llegar a contar con un sistema más completo , con un mayor número de alternativas y como consecuencia, más efectivo.

a) Almacenamiento y Caracterización.

En esta fase se va a definir la manera en que se almacenarán y caracterizarán los documentos para su posterior recuperación.

La entrada al sistema será el texto completo del documento (Full Text) , acompañado por una breve descripción del mismo (encabezado) , el cual tendrá una función parecida al "abstract" , en el sentido de que en él , se puede hacer

referencia a aspectos que trata el documento y que no aparecen en forma explicita en el texto del mismo.

La caracterización del documento estará basada en un mecanismo de indexacion. Veamos: Habrá dos modos de almacenar los documentos: el modo "ON LINE" y el modo "BATCH"; en el primero se captura el documento y al término se inicia un proceso de análisis en el cual se eliminan todas las palabras que no son significativas, como lo son los artículos, pronombres, preposiciones, conjunciones, etc. Estas palabras han sido consideradas previamente y durante el análisis se establece una comparación entre este grupo y las palabras que integran el documento. Al conjunto de palabras restantes o significativas se les aplicará un proceso de filtración en el cual el usuario tiene como opción aceptar o rechazar palabras nuevas, aquellas que sean aceptadas se guardarán en un archivo aparte y se les denominará términos índice.

El modo "BATCH" es llamado así por aceptar captura de un lote o conjunto de documentos sin someterlos en ese momento al proceso de análisis mencionado antes. Estos documentos deberán ser analizados posteriormente. La finalidad de usar este modo de captura es que si la persona que la realiza no está capacitada para efectuar el proceso de filtración que se aplica después, entonces mediante otra opción del sistema, alguna persona mejor informada, puede hacer el análisis de las palabras para caracterizar el documento, tal análisis se puede hacer en cualquier momento, ya que se tienen diferenciados los documentos que fueron capturados con cada uno de los modos descritos.

Durante el proceso anterior cada palabra significativa es almacenada en un archivo aparte, en el cual se le asigna un número para identificarla, además se lleva el conteo de los documentos en los cuales aparece. Al mismo tiempo, en un

tercer archivo se guarda la información referente a cada uno de los documentos en los cuales está contenida.

B) Estrategias de búsquedas

Una vez que el documento ha sido caracterizado, y que sus palabras clave han sido detectadas, podemos realizar las búsquedas. En primer término tenemos que la búsqueda de nivel más simple es la **BUSQUEDA POR PALABRA** (o subpalabra), aquí se pide al usuario que introduzca una palabra con el propósito de saber en cuales documentos aparece de manera significativa y si se desea, tener acceso inmediato a los mismos; una vez que se ha introducido la palabra, se le busca en el archivo de palabras y si es encontrada, nos remitimos al tercer archivo para saber exáctamente en que documentos se encuentra, de esta manera regresamos al archivo de documentos y con la información anterior, se despliega en pantalla una lista de los documentos (ENCABEZADOS) que la contienen, seleccionando un elemento de la lista se podrá revisar el documento completo para decidir si es relevante o no. Por otro lado si la búsqueda es por subpalabra, se introducen los caracteres que aparecen a la izquierda (raíz) de la palabra , no importando el número de ellos, como el archivo de palabras se ha ordenado alfabéticamente de antemano, al buscar la subpalabra (si existe), localizamos la primera palabra que la contiene, de esta manera se recuperan todas las palabras cuyos primeros caracteres coinciden con la subpalabra especificada, de tal modo que el usuario puede escoger entre ellas la que más le convenga. Por ejemplo: si el usuario pregunta por la subpalabra "GOB" el sistema deberá regresar un conjunto de palabras parecido al siguiente:

GOBERNACIÓN
GOBERNANTE
GOBERNANTES
GOBERNATURA
GOBIERNO

En este caso, la lista anterior está formada por todas las palabras significativas que existen en la colección y que comienzan con "GOB". Entonces se deberá escoger de la lista la palabra deseada y el sistema presentará otra lista con todos los documentos a los cuales pertenece, junto con una descripción de cada uno de ellos, posteriormente el usuario podrá seleccionar el o los documentos que juzgue de interés para consultarlos y comprobar la relevancia de los mismos. Podríamos decir que este tipo de búsqueda es alfabética o por raíz.

Aunque esta manera de búsqueda no es adecuada cuando las colecciones de documentos adquieren tamaños considerables (ya que el tamaño de los conjuntos recuperados tiende a crecer demasiado), se le puede considerar como una herramienta útil, ya que nos da con precisión las palabras que figuran como significativas dentro del sistema, por las cuales se puede preguntar. Ademas equivale a tener agrupados los accidentes gramaticales de las palabras (género y número) así como las diferentes formas verbales que indican una acción (infinitivo,participio,gerundio, etc.)

La otra forma que será utilizada para recuperar documentos es la conocida como **BUSQUEDA BOOLEANA** y en el caso de nuestro sistema, viene siendo una extensión de la anterior; una vez que se ha recuperado un conjunto de documentos en respuesta a una búsqueda por palabra, conservamos ese conjunto, después hacemos otra búsqueda para una nueva palabra, de tal manera que obtenemos otro conjunto de documentos recuperados en respuesta a la segunda palabra, entonces, si nos fijamos en la intersección de ambos conjuntos (los documentos que pertenecen a ambos), el resultado será un nuevo conjunto formado por los documentos que contienen ambas palabras. Si se desea aumentar otra palabra a la pregunta, tomamos como base el nuevo conjunto recuperado y se sigue el mismo procedimiento, con esto obtendremos conjuntos más pequeños y

por lo tanto, más específicos. Adicionalmente se implementarán otros tipos de búsqueda alternativos a ésta, de tal suerte que no sólo se podrán realizar intersecciones de conjuntos, sino uniones, complementos, etc. Así,.se dice que la estrategia de búsqueda booleana recupera aquellos documentos que satisfacen (o"son ciertos") a la pregunta, la cual debe ser presentada en términos de palabras clave y combinada mediante los operadores lógicos usuales: AND, OR Y NOT .

Veamos un ejemplo: supongamos que el término I_1 se encuentra en los documentos D_1, D_3, D_5, D_8; I_3 está en D_3, D_5, D_8, D_9 e I_5 pertenece a D_1, D_3, D_7, D_9. Supongamos además que la pregunta Q = I_1 e I_3, entonces el conjunto recuperado será $\{D_3, D_5, D_8\}$. Por otro lado si Q = (I_1 e I_3 e I_5) intersectamos el conjunto anterior con el conjunto de documentos que contienen a I_5 $\{D_1, D_3, D_7, D_9\}$ para obtener finalmente $\{D_3\}$, es decir solamente el documento D_3 contiene a los tres términos índice considerados en la pregunta.

Para la programación del sistema se escogio el lenguaje **Clipper**, versión 5.01. la razón principal de esta elección es qúe Clipper es un sistema manejador de bases de datos (.dbf), con un complilador (Clipper.éxe), cuenta con la capacidad de generar código ejecutable. Su lenguaje de programación es estructurado y dispone de una buena cantidad de comandos y funciones, algunas de las cuales facilitan la manipulación de cadenas de caracteres (Texto) suficientemente grandes.

4.2 DISEÑO DEL SISTEMA.

DISEÑO DE LAS BASES DE DATOS DEL SISTEMA.

A continuación se presenta la descripción de cada una de las bases de datos con que cuenta el sistema, así como de los campos que las componen.

La base de datos DOCTOS.DBF tiene como propósito principal almacenar el texto completo de cada documento, así como algunas referencias para poder identificarlo. Su estructura es la siguiente:

CAMPO	TIPO	TAMAÑO
REF	Caracter	6
CONT	Memo	10
FECHA	Date	8
ENCAB	Caracter	60
OL	Lógico	1

El campo REF guarda un número para hacer referencia al documento en cuestión, es para control interno del sistema, es decir, el usuario no lo puede modificar.

El campo CONT es de tipo memo, en Clipper estos datos son usados para representar datos de tipo caracter de longitud variable, un campo tipo memo ocupa 10 caracteres en la base de datos, este campo es utilizado como un apuntador al dato (texto) propiamente dicho, el cual es almacenado en un archivo (.DBT) por separado. Es aquí donde se almacenará el texto completo del documento, la limitante que tiene Clipper es que en cada registro de estos, la longitud no puede ser mayor que 65,535 caracteres, es decir 64 Kb.

El campo FECHA, obviamente, guardará la fecha en que el documento fué dado de alta en el sistema, proporcionando así

una alternativa más de búsqueda, por omisión se asigna la fecha del sistema.

El campo ENCAB es un campo que sirve para almacenar una brevísima descripción del documento (encabezado). Este campo es importante, ya que la información que contiene será desplegada en la pantalla cuando se recuperen documentos en respuesta a una busqueda, de esta manera el usuario podrá revisar los que juzgue pertinentes.

Finalmente el campo OL sirve para detectar qué tipo de captura se utilizó para dar de alta el documento, si fué captura ON LINE el valor del campo será verdadero (T), si fué captura BATCH, su valor será falso (F), de esta manera, cuando se realice la indexación y se den de alta las palabras en la base PALABRAS mediante el módulo de ALTAS, sólo se considerarán aquellos registros cuyo valor del campo OL sea falso (F).

La base PALABRAS contiene las palabras que son significativas (términos indice) dentro de los documentos. Su estructura es la que sigue:

CAMPO	TIPO	TAMAÑO
PALABRA	Caracter	15
NO_DOCTOS	Numérico	6
NO_PALAB	Numérico	6

El campo PALABRA es utilizado para guardar la palabra significativa.

El campo NO_DOCTOS contiene el número de documentos para los cuales la PALABRA es significativa.

El campo NO_PALAB guarda un número que permite identificar la PALABRA, y por medio de este campo se relaciona con la base REF.DBF.

La base de datos llamada REF.DBF tiene como función guardar las referencias de las PALABRAS y los DOCUMENTOS en los cuales se encuentran las mismas. Al realizar una búsqueda, ésta se efectúa sobre la base PALABRAS.DBF, si es encontrada entonces nos remitimos a la base de referencias (REF.DBF) para buscar el número que le corresponde a la PALABRA ya que en su registro se ubica la información referente a los documentos que la contienen. Su estructura es la siguiente:

CAMPO	TIPO	TAMAÑO
NO_PALAB	Numérico	6
RD01	Numérico	6
.	.	.
.	.	.
.	.	.
RD15	Numérico	6

El campo NO_PALAB tiene como función almacenar el número de la PALABRA con el cual es identificada.

Los campos RD01 al RD15 son de referencia a los documentos donde se encuentra una determinada palabra, cada campo guarda el número de un documento que la contiene. Cuando esa palabra se encuentra en más de 15 documentos se crea un nuevo registro, en el cual se siguen almacenando las referencias a los documentos posteriores, este proceso se repite cada vez que el número de documentos en los que se encuentra la palabra rebasa un múltiplo de 15. No existe una razón especial para crear una estructura con 15 referencias, por lo tanto, al modificar un parámetro en el código del programa el número de referencias en el registro se puede modificar.

La figura 4.1 muestra gráficamente las relaciones entre las bases de datos.

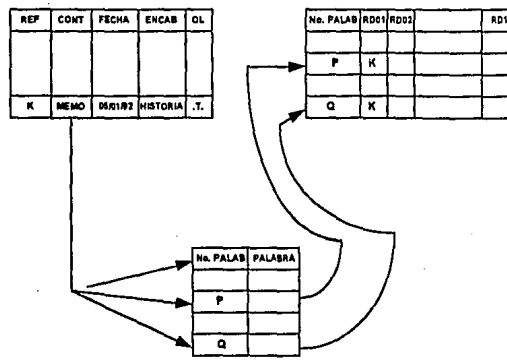


fig 4.1 Representación grafica de las relaciones entre las bases de datos del sistema

La figura 4.2 muestra de forma gráfica los menús del sistema.

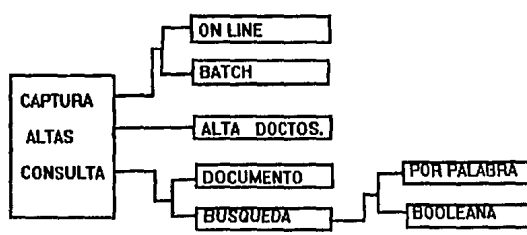


FIG. 4.2

4.3 ESTRATEGIA DE CLASIFICACION Y BUSQUEDA.

El primer paso que se lleva a cabo en cualquier sistema de recuperación de información , es el almacenamiento y clasificación de documentos, de ello se encarga el modulo de captura,aquí se captan en primer término los campos de información general del documento como son: el número de referencia, la fecha y el encabezado; posteriormente se permite la edición del texto del documento mediante la función Memoedit() de Clipper, la cual permite llevar a cabo la interfase con el usuario para la edición del texto que puede ser usados en campos tipo memo y cadenas largas de caracteres (tamaño máximo de 64 kb-1). Lo que en realidad se esta editando en ese momento es un "buffer" de texto cuyo contenido será almacenado en la base DOCTOS. Al terminar la edición ,posteriormente es asignado a una variable de memoria la cual es pasada como parámetro a una función que se encarga de analizar el texto de la siguiente manera:

1) Extrae del texto una cuerda con palabras significativas . Por ejemplo, si el texto original fuera:

"La Universidad Nacional Autónoma de México es una corporación pública dotada de plena capacidad jurídica y que tiene por fines impartir educación superior para formar profesionistas, investigadores, profesores universitarios y técnicos útiles a la sociedad".

El primer paso en el análisis seria obtener la cuerda:

Universidad/Nacional/Autónoma/México/corporación/pública/dotada/plena/capacidad/jurídica/tiene/fines/impartir/educación/superior/formar/profesionistas/investigadores/profesores/universitarios/técnicos/útiles/sociedad.

2) Con la cuerda de palabras significativas obtenida , se checa si cada una de estas ha sido dada de alta con

anterioridad en la base PALABRAS, las palabras nuevas son desplegadas dando al usuario la facilidad de eliminar aquellas que no desee incorporar a la base, incluso puede regresar a la edición del documento para hacer correcciones en caso de que se detecte una palabra mal escrita.

3) Una vez que se han aprobado las palabras que describirán al documento, queda conformada la cadena con las palabras que serán agregados a la base PALABRAS, se incrementa la CANTIDAD de documentos en que aparece cada una y se registra en cada caso en la base REF el número del documento al que ha sido asignado como palabra clave.

En el módulo consultas es donde se realizarán las búsquedas de documentos relevantes a una pregunta, para ello seleccionamos el modelo booleano. Antes de ver la forma en que implementamos este tipo de búsqueda en nuestro sistema, ahondaremos un poco en el mismo:

Después de haber almacenado y clasificado los documentos en la manera descrita, lo que hemos hecho es: para una palabra dada, determinar el conjunto de documentos que la contienen. La clasifacación efectuada equivale a construir una matriz en la cual cada renglón corresponde a un documento en la colección y cada columna a una palabra que ha sido asignada a al menos un documento. Veámoslo con un ejemplo: supongamos que nuestra colección consta de seis documentos como se muestra a continuación .

DOCUMENTO	TEXTO DEL DOCUMENTO
1	Control de la contaminación en ríos.
2	Reducción de la contaminación de la atmósfera.
3	La contaminación y sus efectos en los ríos.

4 Efectos de la contaminación por humo.

5 Contenido del humo en la atmósfera.

6 Control de la contaminación de los ríos.

Eliminando las palabras no significativas y anotando en qué documentos aparece cada palabra, formamos una matriz con seis renglones y ocho columnas en la cual el elemento en el i-ésimo renglón y la j-ésima documento contiene o no a la j-ésima palabra. En nuestro ejemplo:

	ATMOSFERA	CONTENIDO	CONTROL	EFECTOS	CONTAMINACIÓN	REDUCCIÓN	RÍOS	HUMO
1	0	0	1	0	1	0	1	0
2	1	0	0	0	1	1	0	0
3	0	0	0	1	1	0	1	0
4	0	0	0	1	1	0	0	1
5	1	1	0	0	0	0	0	1
6	0	0	1	0	1	0	1	0

De esta manera la búsqueda booleana puede hacerse mediante la aplicación de operaciones lógicas apropiadas con las columnas de la matriz de PALABRAS/DOCUMENTOS. Las reglas de dichas operaciones lógicas para un par de columnas de elementos ui y vi para producir una columna de elementos w_i, se enuncian a continuación :

1.- Para la operación AND :

$$w_i = \begin{cases} 1 & \text{si } u_i = v_i = 1 \\ \text{o en cualquier otro caso} \end{cases}$$

2.- Para la operación OR :

$$w_i = \begin{cases} 0 & \text{si } u_i = v_i = 0 \\ 1 & \text{en cualquier otro caso} \end{cases}$$

3.- Para la operación NOT aplicada a la columna de elementos u_i :

$$w_i = \begin{cases} 1 & \text{si } u_i = 0 \\ 0 & \text{si } u_i = 1 \end{cases}$$

Para encontrar los documentos que contienen CONTAMINACIÓN AND RÍOS , en nuestro ejemplo , realizamos la operación AND para las columnas correspondientes :

CONTAMINACIÓN		RÍOS		
1		1		1
1		0		0
1	AND	1	=	1
1		0		0
0		0		0
1		1		1

De aquí se desprende que los documentos que satisfacen la pregunta son el número 1 , el 3 y el 6.

Si aceptamos esta solución, se presentan ahora dos problemas:

1) Establecer un mecanismo para poder utilizar preguntas un poco más complejas, es decir, en donde se combinen varios operadores lógicos con varios términos índice.

2): Encontrar la manera de implementar o adecuar dicha solución al diseño de nuestro sistema.

Una manera de intentar dar solución al primer problema es mediante el uso de preguntas formadas a partir de "OR" lógicos simplemente anidados como parámetros de un operador "AND". Veamos:

Supongamos que las únicas operaciones lógicas permitidas son: AND, OR, y NOT.

Definimos un PARAMETRO como una expresión booleana con una y sólo una de las siguientes propiedades:

i) Consiste solamente de una palabra clave.
ii) Consiste de la negación de una palabra clave.
iii) Consiste de un conjunto de palabras clave conectadas por el operador OR.
iv) Consiste de un conjunto de negaciones de palabras clave conectadas por el operador OR.

Una PREGUNTA será un conjunto de varios parámetros conectados por el operador AND. Por ejemplo.

I_1AND(I_2ORI_3)AND(I_4ORI_5ORI_6)AND(NOTI_7)ANDI_8

Es una pregunta con cinco parámetros donde $I_1,\ldots,I_8$ son palabras clave (o términos índice).

Cuando se introduzca una pregunta de este tipo, el sistema deberá realizar la lectura de la misma, en este momento se creará una cadena (DESPAR) de ceros y unos, cuya función será describir los parámetros de la pregunta; la i-ésima posición de DESPAR será puesta en "1" si la pregunta contiene un i-ésimo parámetro formado por uno ó varios términos índice (i,

iii) y será puesta en "0" si la pregunta contiene un i-ésimo parámetro formado por la negación de uno ó varios términos índice (ii, iv). Por ejemplo, para la pregunta:

I_1AND(I_2ORI_3)AND(I_4ORI_5ORI_6)AND(NOTI_7)ANDI_8

su DESPAR correspondiente será:

11101

Entonces una manera de buscar los documentos que satisfacen la pregunta es la siguiente: Para cada documento en la base se crea una cadena (DESDOC) del mismo tamaño que DESPAR, que será inicializada con cero en todas sus posiciones, cada palabra clave del documento en turno es buscada en la pregunta, si aparece, entonces nos fijamos en el parámetro en el cual esta contenida y la posición correspondiente a ese parámetro en DESDOC es puesta en "1", si DESPAR=DESDOC, después de haber leído todos los términos del documento, éste es recuperado. Veamos un ejemplo:

Consideramos la pregunta mostrada anteriormente, entonces:
DESPAR = 11101

Supongamos que nuestra colección de documentos está indexada como sigue:

D_1 -	I_1,	I_2			I_5,		I_7,		I_9
D_2 -		I_2,	I_3,	I_4,	I_5				
D_3 -	I_1,		I_3,			I_6,		I_8	
D_4 -		I_2,		I_4,		I_6,		I_8	
D_5 -	I_1,	I_2	I_3,				I_7		

Construyamos ahora la cadena DESDOC para D_1: Inicializamos DESDOC=00000. El término I_1 en D_1 aparece en el primer parámetro de la pregunta, por lo tanto DESDOC=10000, algo

análogo sucede con I_2, I_5 e I_7, como I_9 no aparece en ningún parámetro de la pregunta, su ocurrencia en D_1 no afecta el valor final de DESDOC. Al terminar con D_1 tenemos que DESDOC=11110, que es diferente de DESPAR, de modo que D_1 no es recuperado.

Efectuando el mismo procedimiento para los documentos de nuestra colección tendríamos:

Para D_2, DESDOC=01100, diferente de DESPAR, no recuperado.
Para D_3, DESDOC=11101, igual que DESPAR, recuperado.
Para D_4, DESDOC=01101, diferente de DESPAR, no recuperado.
Para D_5, DESDOC=11010, diferente de DESPAR, no recuperado.

Esta parace ser una alternativa aceptable para resolver el primer problema, desgraciadamente, tiene la desventaja de que cualquier consulta requiere analizar todos y cada uno de los términos de todos y cada uno de los documentos, lo cual implica una gran cantidad de tiempo, además no resuelve el segundo problema, es decir, no se adapta al diseño de nuestros archivos. Sin embargo retomaremos esta idea, adaptándola a nuestro sistema y sin necesidad de revisar la totalidad de la colección, sino únicamente los registros en nuestros archivos (PALABRAS Y REF), de aquellos términos que figuran en la pregunta.

Supongamos que el primer parámetro de la pregunta es del tipo i), o sea que no está formado de negaciones y consta de un solo término (I_1): dicho de otra manera, la primera condición que deben de satisfacer los documentos a buscar es que contengan la palabra I_1. Proponemos el siguiente algoritmo:

0) Inicio.

1) Sean :

T=2

P=2

i=1

q=0

N=Número de términos que contiene la pregunta.

M=Número de parámetros que contiene la pregunta.

AP=(1,2,P3, 4,...,P_N) un arreglo donde P_T es el número de parámetro que contiene al T-ésimo término de la pregunta. $P_T <= P_{t+1}$.

2) Leer el único término (I_1) del primer parámetro.

3) Buscar la palabra I_1 en el archivo de PALABRAS, si no es encontrada, ir al paso 7), no existen documentos en la colección que satisfagan la pregunta. Si se encuentra, sea q el número que le corresponde a la palabra, continuar.

4) Buscar el registro de la palabra número q en REF (por construcción, debe encontrarse), leer los números de los documentos a que ha sido asignada.

Sea D={D_1, D_2,...,D_k} el conjunto de los números de los documentos en la colección que contienen a I_1. Para cada i en el conjunto {1,2,..k}, formamos una cadena (S_i) de ceros y unos de la siguiente manera:

Si=100...0

Cuyo tamaño es igual a M. El "1" de la primera posición obedece a que I_1 está en Di para toda i

5) Si P es menor o igual a M:

5.1) Si T es menor o igual que N:

5.1.1) Si P= AP[T]

Buscar It (el T-ésimo término de la pregunta) en el archivo de PALABRAS.

Si existe, buscar su registro en el archivo REF.

Sea:

AT= {Dj | el T-ésimo término de la pregunta está en Dj}

Es decir AT es el conjunto de documentos que contienen al T-ésimo término de la pregunta.

5.1.2 Si i <= k

Si Di pertenece a AT asignar 1 a la p-ésima posición de Si

Hacer i=i+1

Ir a 5.1.2

5.1.3 Hacer T=T+1, i=1

Ir a 5.1

5.2) Hacer P=P+1

Ir a 5

6) Si Si=DESPAR, entonces Di es recuperado. Para cada i en el conjunto {1,2,..k}

7) Fin

A continuación ilustramos el algoritmo con un ejemplo; consideremos la pregunta del ejemplo expuesto anteriormente:

I_1AND(I_2ORI_3)AND(I_4ORI_5ORI_6)AND(NOTI_7)ANDI_8

Como vimos su DESPAR correspondiente es:
11101

Habíamos supuesto que nuestra colección de documentos estaba indexada como sigue:

$D_1 - I_1, I_2, I_5, I_7, I_9$
$D_2 - I_2, I_3, I_4, I_5$
$D_3 - I_1, I_3, I_6, I_8$
$D_4 - I_2, I_4, I_6, I_8$
$D_5 - I_1, I_2, I_3, I_7$

Bajo nuestro sistema, tal indexación equivaldría a tener el archivo REF como se muestra en la siguiente tabla:

REF

NUM PAL	RDO1 1	RDO2	RDO3	RDO4	RDO5	RDO6	RDO7	...	RD15
1	1	3	4						
2	1	2	4	5					
3	2	3	5						
4	2	4							
5	1	2							
6	3	4							
7	1	5							
8	3	4							
9	1								

El primer campo indica el número que ha sido asignado a la palabra correspondiente al registro en cuestión, los campos siguientes indican los números de los documentos a que ha sido asignada la palabra cuyo número aparece en el primer campo.

La base de palabras muestra cómo se han asignado números a las palabras:

PALABRAS		0) Inicio 1) Sean
NUM PALAB	PALAB	T = 2
1	I_1	P = 2
2	I_2	I = 2
3	I_3	q = 0
4	I_4	N = 8 términos de la pregunta.
5	I_5	M = 5 parámetros que contiene la
6	I_6	pregunta

7 I7
8 I8
9 I9

AP = (1,2,3,2,3,3,3,4,5) un arreglo donde la T-ésima posición indica el número de parámetro que contiene al T-ésimo término de la pregunta. Por ejemplo, la cuarta posición indica que el cuarto término de la pregunta pertenece al tercer parámetro.

2) Leer el único término (I_1) del primer parámetro.

3) Buscar la palabra I_1 en el archivo de PALABRAS, existe, entonces q = 1

4) Buscar el registro de la palabra número 1 en REF, leer los números de los documentos a que ha sido asignada. D={1,3,5} el conjunto de los números de los documentos en la colección que contienen a I_1. Para cada i en D, formamos una cadena (S_i) de ceros y unos de la siguiente manera:

S_1 = 10000 S_3 = 10000 S_5 = 10000

5) P=2 es menor o igual a M=5 :

5.1) Si T=2 es menor o igual a N=8:

5.1.1) Si P=AP[2]=2
Buscar I_2 (el segundo término de la pregunta) en el archivo de PALABRAS.
Existe, buscar su registro en el archivo REF.
Sea:

AT={1,2,4,5}
Es decir AT es el conjunto de documentos que contienen segundo término de la pregunta.

5.1.2 i=1 <= 3 documentos que contienen a I_1
Como el documento 1 pertenece a AT asignar 1 a la segunda posición de S_1
S_1=11000
Hacer i=i+1=2
Ir a 5.1.2

5.1.2 i=2 <= 3 documentos que contienen a I_1
Como el documento 3 no pertenece a AT, S_3 no es modificada
S_3=10000
Hacer i=i+1=3
Ir a 5.1.2

5.1.2 i=1 <= 3 documentos que contienen a I_1
Como el documento 5 pertenece a AT asignar 1 a la segunda posición de S5
S5=11000
Hacer i=i+1=4
Ir a 5.1.2
Como i=4 > 3 documentos que contienen a I_1, voy a 5.1.3

5.1.3 Hacer T=T+1=3 , i=1
Ir a 5.1

5.1) Si T=3 es menor o igual que N=8
5.1.1) Si P=AP[3]=2
Buscar I_3 (el segundo término de la pregunta) en el archivo de PALABRAS.
Existe, buscar su registro en el archivo REF.
Sea:

AT={2,3,5}
Es decir AT es el conjunto de documentos que contienen tercer término de la pregunta.

5.1.2 i=1<=3 documentos que contienen a I_1
Como el documento 1 no pertenece a AT, S1 no es modificada
S1=11000
Hacer i=i+1=2
Ir a 5.1.2

5.1.2 i=2 <=3 documentos que contienen a I_1
Como el documento 3 pertenece a AT asignar 1 a la segunda posición de S3
S3=11000
Hacer i=i+1=3
Ir a 5.1.2

5.1.2 i=1 <=3 documentos que contienen a I_1
Como el documento 5 pertenece a AT asignar 1 a la segunda posición de S5
S5=11000
Hacer i=i+1=4
Ir a 5.1.2
Como i=4 > 3 documentos que contienen a I_1, voy a 5.1.3

5.1.3 Hacer T=T+1=4 , i=1

ir a 5.1

5.1) Si T=4 es menor o igual que N=8:

5.1.1 Como P=2 es diferente de AP[4]=3, voy a 5.2

5.2) Hacer P=P+1=3
Ir a 5

5) P=3 es menor o igual a M=5:
5.1) Si T=4 es menor o igual que N=8:
5.1.1 Si P=AP[4]=3
Buscar I_4 (el cuarto término de la pregunta) en el archivo de PALABRAS.
Existe, buscar su registro en el archivo REF.
Sea:
AT=(2,4)
Es decir AT es el conjunto de documentos que contienen al cuarto término de la pregunta

5.1.2 i=1 <= 3 documentos que contienen a I_1
Como el documento 1 no pertenece a AT, S_1 no es modificada.
Hacer i=i+1=2
Ir a 5.1.2

5.1.2 i=2 <= 3 documentos que contienen a I_1
Como el documento 3 no pertenece a AT, S_3 no es modificada.
S_3=11000
Hacer i=i+1=3
Ir a 5.1.2

5.1.2 i=1 <= 3 documentos que contienen a I_1
Como el documento 5 no pertenece a AT, S_5 no es modificada.
S_5=11000
Hacer i=i+1=4
Ir a 5.1.2

Como i=4 > 3 documentos que contienen a I1, voy a 5.1.3

5.1.3 Hacer T=T+1=5 , i=1
Ir a 5.1

5.1) Si T=5 es menor o igual que N=8:
5.1.1 Si P=AP[5]=3
Buscar I_5 (el quinto término de la pregunta) en el archivo de PALABRAS.
Existe, buscar su registro en el archivo REF.
Sea:
AT={1,2}

Es decir AT es el conjunto de documentos que contienen al quinto término de la pregunta.

5.1.2 i=1 <= 3 documentos que contienen a I_1
Como el documento 1 no pertenece a AT, asignar 1 a la tercera posición de S_1
S_1=11100
Hacer i=i+1=2
Ir a 5.1.2

5.1.2 i=2 <= 3 documentos que contienen a I_1
Como el documento 3 no pertenece a AT, S_3 no es modificada.
S_3=11000
Hacer i=i+1=3
Ir a 5.1.2

5.1.2 i=1 <= 3 documentos que contienen a I_1
Como el documento 5 no pertenece a AT, S_5 no es modificada.
S_5=11000
Hacer i=i+1=4
Ir a 5.1.2

Como i=4 > 3 documentos que contienen a I_1 , voy a 5.1.3

5.1.3 Hacer T=T+1=6 , i=1
Ir a 5.1

5.1) Si T=6 es menor o igual que N=8:

5.1.1 Si P=AP[6]=3
Buscar I_6 (el sexto término de la pregunta) en el archivo de PALABRAS.
Existe, buscar su registro en el archivo REF.
Sea:
AT={3,4}
Es decir AT es el conjunto de documentos que contienen al sexto término de la pregunta.

5.1.2 i=1 <= 3 documentos que contienen a I_1
Como el documento 1 no pertenece a AT, S_1 no es modificada.
S_1=11100
Hacer i=i+1=2
Ir a 5.1.2

5.1.2 i=2 <= 3 documentos que contienen a I_1
Como el documento 3 pertenece a AT, asignar 1 ala tercera posición de S_3
S_3=11100
Hacer i=i+1=3
Ir a 5.1.2

5.1.2 i=1 <= 3 documentos que contienen a I_1
Como el documento 5 no pertenece a AT, S_5 no es modificada.
S_5=11000
Hacer i=i+1=4
Ir a 5.1.2

Como i=4 > 3 documentos que contienen a I_1 , voy a 5.1.3

5.1.3 Hacer T=T+1=7, i=1
Ir a 5.1

5.1) Si T=7 es menor o igual que N=8

5.1.1) Como P=3 es diferente de AP[7]=4, voy a 5.2

5.2) Hacer P=P+1=4
ir a 5

5) P=4 es menor que o igual a M=5:

5.1) Si T=7 es menor o igual que N=8

5.1.1) Si P=AP[7]=4
Buscar I7 (el séptimo término de la pregunta) en el archivo PALABRAS.
Existe, buscar su registro en el archivo REF.
Sea:
AT={1,5}
Es decir AT es el conjunto de documentos que contienen el sexto término de la pregunta.

5.1.2) i=1 <= 3 documentos que contienen a I1.
Como el documento 1 pertenece a AT, asignar 1 a la cuarta posición de S1.
S1=11110
Hacer i=i+1=2
ir a 5.1.2

5.1.2) i=2 <= 3 documentos que contienen a I1.
Como el documento 3 no pertenece a AT, S3 no es modificada.
S3=11100
Hacer i=i+1=3
ir a 5.1.2

5.1.2) i=3 <= 3 documentos que contienen a I1.
Como el documento 5 pertenece a AT, asignar 1 a la cuarta posición de S5.
S5=11010

Hacer i=i+1=4
ir a 5.1.2

Como i=4 > 3 documentos que contienen a I1:
Voy a 5.1.3

5.1.3 Hacer T=T+1=8, i=1
ir a 5.1

5.1) Si T=8 es menor o igual que N=8:

5.1.1) Como P=4 es diferente de AP[8]=5, voy a 5.2

5.2) Hacer P=P+1=5
Ir a 5

5) P=5 es menor que o igual a M=5:

5.1) Si T=8 es menor o igual que N=8

5.1.1) Si P=AP[8]=5
Buscar I8 (el octavo término de la pregunta) en el archivo PALABRAS.
Existe, buscar su registro en el archivo REF.
Sea:
AT={3,4}
Es decir AT es el conjunto de documentos que contienen el octavo término de la pregunta.

5.1.2) i=1 <= 3 documentos que contienen a I1.
Como el documento 1 no pertenece a AT, S1 no es modificada.
S1=11110
Hacer i=i+1=2
ir a 5.1.2

5.1.2) i=2 <= 3 documentos que contienen a I1.
Como el documento 3 pertenece a AT, asignar 1 a la quinta posición de S3.
S3=11101
Hacer i=i+1=3
ir a 5.1.2

5.1.2) i=3 <= 3 documentos que contienen a I1.
Como el documento 5 pertenece a AT, asignar 1 a la cuarta posición de S5.
S5=11010
Hacer i=i+1=4
ir a 5.1.2

Como i=4 > 3 documentos que contienen a I1:
Voy a 5.1.3

5.1.3 Hacer T=T+1=9, i=1
ir a 5.1

5.1) Como T= 9 no es menor o igual que N=8, ir 5.2.

5.2) Hacer P=P+1=6
ir a 5

5) Como P=6 no es menor o igual a M=5, ir a 6.

6) S1=11110, diferente de DESPAR. no recuperado.

S3=11101, igual que DESPAR recuperado.

S5=11010, diferente de DESPAR. no recuperado.

7) Fin.

El sistema regresa una lista con los documentos que satisfacen la pregunta y permite su consulta.

Es así como implementamos la búsqueda booleana en nuestro sistema. Quisiera hacer notar que la restricción hecha al primer parámetro de la pregunta, en el sentido de que debe ser simple, es decir no admite un operador OR o NOT, podría ser resuelta si se llevan a cabo búsquedas complementarias, aunque, claro está, ésto aumenta el número de operacioines lógicas que el usuario debe realizar. Sin embargo, si se deseara mejorar el sistema, se podría implementar un thesaurus para establecer asociaciones entre términos índice, en particular con el primer parámetro de la pregunta y de esta manera, reducir el trabajo para el usuario.

Una última implementación hecha al sistema, es la relativa a la manera en que el usuario introduce su pregunta, como se dijo, las preguntas deben estar en un lenguaje "comprensible" para la máquina y los usuarios no siempre están están familiarizados con la construcción de preguntas formales. Así, se implementó un módulo para la formulación de preguntas, donde el usuario es "llevado de la mano", tiene acceso a todas las palabras registradas en la base de datos e

introduce los operadores permitidos utilizando ventanas y teclas de dirección, de tal forma que no es posible cometer errores de sintaxis en la construcción de la pregunta formal. Luego el sistema convierte la pregunta en una cadena de ceros y unos para formar DESPAR y continua con la búsqueda.

CONCLUSIONES.

Dada la naturaleza de los sistemas de recuperación de información, dificilmente se puede dar por terminada la tarea del desarrollo de un sistema de este tipo. Por ejemplo, para perfeccionar nuestro sistema, sería conveniente que realizara la ponderación de términos índice con respecto a cada documento, de esta forma sería posible crear un algoritmo que realizara la contrucción automatizada de un thesaurus (capítulo 2), pero bien podría dedicarse otro trabajo completo para resolver éste problema.

Otro punto que debe considerarse es el problema de la captura, sería conveniente que un sistema de éste tipo contara con la opción de reconocer archivos previamente mediante otros medios como un *scanner*.

Una vez que el sistema estuvo listo para admitir preguntas y efectuar búsquedas, había que buscar dónde aplicar el sistema, un primer intento se realizó en la biblioteca del colegio Green Hills en la ciudad de México, por razones laborales sólo se pudo implementar el módulo de búsquedas por palabra, y se hizo un programa para la indexación automática de los títulos que habían sido previamente capturados. Actualmente se encuentra funcionando y se han observado resultados aceptables en la búsqueda de material bibliográfico. La biblioteca cuenta con alrededor de cinco mil títulos.

Para el módulo de Búsqueda Booleana se están realizando búsquedas sobre archivos de prueba y los primeros resultados indican que es posible alcanzar un buen nivel de eficacia y eficiencia, pero como vimos, no es fácil evaluar la eficacia de éstos sistemas.

Las implementaciones hechas para formular preguntas, garantizan el cumplimiento del criterio de predicción y permiten cumplir con el criterio del punto inútil.

Considero que mínimamente se cumple el objetivo en cuanto a la elaboración del sistema y creo que están sentadas las bases para continuar en el perfeccionamiento de este trabajo.

La investigación bibliográfica proporciona un panorama básico sobre los sistemas de recuperación de información.

En el anexo se mencionan algunos sistemas realizados en nuestro país y creo que todavía hay mucho por hacer en este campo. Existe la intención por parte de un servidor de continuar con este trabajo y que sea de utilidad general, la opción existe y la disponibilidad también. Las colecciones de documentos esperan ser explotadas.

ANEXO

SISTEMAS DE RECUPERACION EXISTENTES EN MEXICO

Actualmente, el uso de la automatización en centros de información en México, va en aumento, se observa un desarrollo importante en el uso de medios electrónicos y ópticos. Se ha comprendido que la creación de bases de datos automatizadas constituyen una herramienta de apoyo para los servicios de información en lo que concierne al almacenamiento y recuperación de la información y a los usos múltiples que puede hacerse de ella.

En México existen dos tipos de sistemas de información que auxilian en el proceso de documentos, los sistemas administradores de bibliotecas, como los son **SIABUC** y **LogiCat**, que esencialmente manejan monografías pero en general cualquier tipo de material documental y los recuperadores de información, como **MICRO CDS/ISIS**. Los primeros fueron diseñados empleando el manejador de base de datos dBase III y el último desarrollado en el lenguaje de programación Pascal.

Los sistemas de información automatizados se encuentran constituidos principalmente por una base de datos, procedimientos y programas de computadora que en su conjunto permiten la captura, almacenamiento, manejo, actualización y recuperación de la información.

La elaboración de LogiCat y SIABUC, corrió a cargo de personal mexicano, encontrándose su domicilio en el D.F. y Colima, respectivamente. El sistema MICRO CDS/ISIS generado a nivel internacional es distribuido a través del CONACYT, el cual tiene sus oficinas en el Distrito Federal.

Dentro de las características del sistema MICRO CDS/ISIS se encuentran:

- Definición de Bases de datos
- Ingreso de nuevos registros en una base ya existente
- Modificación, corrección y eliminación de registros ya existentes
- Construcción y mantenimiento de archivos de acceso rápido a cada base de datos en forma automática
- Recuperación de registros por su contenido, a través de operaciones booleanas de búsqueda.
- Ordenación de los registros en cualquier secuencia deseada
- Capacidad de manejo de archivos ANY
- El número máximo de registros en una Base de datos es de 16 millones
- Tamaño de registros individuales hasta un largo total de 8K
- 200 caracteres como máximo para cada campo
- 200 campos para ser indexados
- Número máximo de campos contenidos en una pagina de la hoja de trabajo, 19
- Número máximo de páginas en una hoja de trabajo, 20
- Número máximo de palabras no significativas (Stopword), 799
- Número de máximo de caracteres por registros para despliegue es de 4000

El sistema MICRO CDS/ISIS tiene como requerimiento de equipo el siguiente:

- MicroComputadora IBM-PC, PS, AT, XT o compatible que trabaje bajo el sistema operativo MS-DOS.
- Monitor (cromático o monocromático)
- Unidad de disco duro con 20 MB como mínimo.
- Unidad de disco flexible

- 640 Kb de memoria RAM

En cuanto a la realización de búsquedas tenemos que el sistema está basado en el algebra booleana, la cual utiliza operadores lógicos representados por los signos (+, *, ^), equivalentes a: OR, AND y NOT.

Existen 4 tipos para la formulación de búsquedas:

- Términos Precisos
- Términos Truncados a la derecha
- Términos ANY
- Campos no indexados.

El menú de búsqueda permite visualizar los términos del diccionario, así como visualizar el resultado de la búsqueda.

Por otra parte, tenemos a los sistemas LogiCat y SIABUC, los cuales manejan esencialmente monografías, sin que esto signifique que no puedan manejar otro tipo de material. Lo que cambia es el tratamiento distinto que se le da a cada material documental. Lo que no disminuye la variedad de productos y reportes impresos que ofrecen estos productos.

En el caso de Logicat (versión más reciente, 4.1), la empresa que lo genero ha desarrollado sistemas complementarios que permiten el tratamiento de más tipos de documentos.

El sistema LogiCat funciona por medio de menús, lo cual simplifica su uso, pues no es necesario memorizar las instrucciones para su operación. A lo largo de todos los procesos LogiCat guía al usuario, indicándole siempre el comando que se está ejecutando; además, despliega las explicaciones necesarias para su operación y solicita en forma clara al usuario las órdenes para realizar un proceso.

Dentro de las características del sistema estan las siguientes:

- Creación y modificación de base de datos
- Captura de datos
- Organización de los índices de recuperación
- Recuperación de datos
- Procesos administrativos

La recuperación en Logicat se realiza a partir de todas las palabras o parte de ellas, incluidas en cualquier campo registrado en la ficha catalográfica: autor, título, casa editorial, encabezamiento de materia, país de edición y derechos de autor, idioma, tipo de material, forma de reproducción, número de clasificación, ISBN y número de acceso de la ficha.

Dentro de LogiCat existen 3 formas para recuperar información, cada una de ellas diseñada para ofrecer al usuario la posibilidad de localizar exactamente el conjunto de fichas que necesita.

Búsqueda Especial, permite recuperar información de las fichas, a partir de cualquiera de los datos ya mencionados.

Búsqueda Normal, en esta técnica, las palabras o parte de ellas que se dan para recuperación, se pueden encontrar en cualquier parte de la cadena de caracteres que forman el campo, es decir, que no importa la posición en que se encuentre en la cadena original.

Búsqueda Directa, se realiza en función de índices que se constituyen o actualizan después de la sesión de captura, es más rápida que las otras dos búsquedas. Las llaves de recuperación son: autor, título, temas, clasificación y número de acceso de la ficha, pudiéndose combinar con el operador lógico "Y" (AND).

Cuando no se requiera gran sofisticación en la estrategia de búsqueda es preferible realizarla con el comando de búsqueda

directa, ya que el tiempo de recuperación se reduce considerablemente.

El requerimiento de equipo para trabajar con LogiCat es:

- Una microcomputadora XT, AT y compatibles
- 512 Kb de memoria (RAM)
- Unidad de disco duro con 20 MB como mínimo.
- Sistema Operativo MS DOS 2.0 o mayor
- El máximo de fichas en una base de datos es de 100,000

El sistema LogiCat está desarrollado con el manejador de base de datos Dbase III plus, lo que permite a los usuarios aumentar las opciones que ofrece, ya que los datos capturados se pueden manipular directamente utilizando Dbase III plus, con lo cual se logran consultas no previstas por el sistema.

El sistema en sus versiones estandarizadas no contempla la estructuración de un thesaurus, sin embargo a solicitud expresa del Centro de Estudios Linguísticos y Literarios (UNAM), se modificó para ofrecer tal opción.

El Sistema Integral Automatizado de Bibliotecas de la Universidad de Colima, SIABUC, fue realizado en la Dirección General de Desarrollo Bibliotecario de la Universidad mencionada a partir de 1983.

Los creadores de SIABUC fijaron como objetivo que el sistema se tornara en una herramienta para la administración bibliotecaria, en razón de que está diseñado para apoyar todas las funciones o un conjunto de ellas a través del uso de una microcomputadora. El sistema fue desarrollado con el manejador de base de datos Dbase III.

Los requerimientos de equipo para el sistema SIABUC son:

- Una microcomputadora AT

- 512 Kb de memoria RAM
- Sistema operativo MS DOS

La capacidad de almacenamiento estará en función del disco duro con que se cuente; un disco de 20 Mb aproximadamente nos permitirá almacenar 30 mil registros. El tamaño máximo por registro es de 2,346 caracteres.

El sistema SIABUC se presenta en 6 módulos de entre los cuales existe el de Control de Archivos de Consulta. Su diseño corresponde a los esquemas tradicionales de recuperación por autor, título y tema. Siendo esta última forma de recuperación la más socorrida por los usuarios. Durante la recuperación por tema pueden utilizarse hasta cuatro descriptores en cada búsqueda.

Dentro del SIABUC, la información se almacena en diversas bases de datos y se generan índices de acuerdo a los requerimientos de recuperación del usuario. El modo de acceso es secuencial-indexado.

Dentro las opciones de este módulo tenemos:

- Consulta al catálogo de descriptores
- Recuperación de información bibliográfica por temas
- Recuperación de información bibliográfica por autor
- Recuperación de información bibliográfica por título
- Consulta al catálogo de topográfico
- Recuperación de bibliografía por temas principales

4.1 SERVICIOS DE CONSULTA

En cuanto a servicios de consulta a bancos de información tenemos que existe conexión a más de 400 bases de datos extranjeros vía telefónica y alrededor de 14 bases nacionales[1] a través de SECOBI del CONACYT (SECOBI es el servicio de consulta a bancos de

[1] GARDUÑO VERA, Roberto. Los formatos MARC y CCF y su aplicación en unidades de información mexicanas. -- p. 155.

información, creado por el CONACYT en 1976). De entre las bases extranjeras que pueden consultarse tenemos a ORBIT, Dialog, New York, Data Resources, Bibliographic Retrieval Services, Questel, G'CAM, Sligos y CISI. De las bases de datos nacionales tenemos ANAFACTA, ARIES, BIBLAT, CLASE, DESA, PERIODICA, SIE-BANXICO, UNAM-JURE, VIVE, LIME, LIBRUNAM, TESIUNAM, e INFOBILA. También existe dentro de la Universidad Nacional el Centro de Información Científica y humanística (CICH), que haciendo una solicitud previa, podemos consultar varias bases de datos sobre diversos temas.
Existen empresas que bajo cierto costo nos permiten consultar bases de datos nacionales e internacionales. Un ejemplo es la empresa de consultoria Información tecnológica INFOTEC, en donde se pueden consultar cerca de 750 bases de datos sobre distintos temas.

Actualmente estan proliferando los productos de texto completo (full text) ya sea por medios ópticos o por consulta en línea por computadora (en la Biblioteca Central de la Universidad Nacional ya existe la consulta a bancos de información almacenados en disco compacto).

Hasta los últimos años para el almacenamiento de bancos de información se toman datos importantes en cuanto al documento o documentos de que se trata, como por ejemplo, autor, título, año de publicación, etc., una breve descripción del tema que trata el documento y un pequeño resumen del contenido, en un sistema de texto completo se almacena el contenido completo del documento. Con los medios de almacenamiento comunmente usados hasta antes del advenimiento del disco compacto (conocido como CD ROM), el hecho de guardar todo el contenido de cada documento nos causa serios problemas al agotarse el espacio disponible para nuestros datos. Problema que es menos sencillo suceda al usar un disco compacto dada la gran capacidad de almacenamiento que tiene este medio (660 Megabytes, aproximadamente 1500 floppys de 5 $^{1}/_{4}$) y la velocidad con que se pueden accesar los datos. Con el fin de

divulgar las bases de datos que se tienen disponibles en disco compacto, existe una publicación llamada Difusión Científica Latinoamericana Cd-Rom, en la que aparecen nombre de la base, nombre de la institución que la elaboro, periodo en que se actualiza y el tipo de información que proporciona. De hecho uno de estos discos contiene información sobre los bancos bibliográficos mexicanos y fue realizado por la Universidad de Colima.

Algunas de las bases de datos hechas en méxico son: BIBLAT, CIIN, CIME, CIME B, CLASE, DESA, MEXICO ARTE, MEXINV, PERIODICA; realizadas por el CICH y la UNAM; DIALEX, producida por el Archivo General de la Nación; SIBA, hecha por Difusión Científica Latinoamericana; LIBRUNAM, SERIUNAM, TESIUNAM, realizadas por la Universidad Nacional; LIME, realizada por el Instituto Nacional de Bellas Artes y CONACYT.

BIBLIOGRAFIA

H.S. Heaps
Information Retrieval.Computational and Theoretical Aspects
Ed. Academic Press, 1978

Dietschmann, Hans J.
Representation and Exchange of Knowledge a Basis of Information Process
Proceedings of the Fifth International Research Forum Science North - Holland 1983

D.C. Blair
Language and Representation in Information Retrieval
Ed. Elsevier Science Publisher, 1990

Van Rijsbergen C.J.
Information Retrieval
Ed. Butterworths, Segunda Edición 1979

Gerard Salton, Michael J. Mc. Gill
Introduction to Modern Information Retrieval
Ed. Mc. Graw - Hill, 1983

Gerard Salton
Automatic Text Processing.The Transformation, Analysis, and Retrieval of Information By Computer
Ed. Addison - Wesley Publishing Company 1989