(10) 2 ejem.

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

UN SISTEMA DE RECUPERACION DE INFORMACION

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A

FERNANDO BOTAS HERRERA

México, D. F. 1981

I N D I C E

INTRODUCCION.

I. RESEÑA HISTORICA.

Las matemáticas y la necesidad de resolver problemas matemáticos han sido siempre un reto para el hombre. La historia del proceso - de datos es una compilación del esfuerzo contínuo del hombre para - encontrar métodos mejores y más eficientes de reunir y procesar -- datos útiles a medida que sus problemas a resolver aumentaban, tanto en complejidad como en tamaño.

Así mismo a través del tiempo, el hombre ha estado acostumbrado a almacenar su conocimiento debido a varias razones, él puede juzgar la probabilidad de que el material continuará interesandole en un período, su decisión de guardar el material ó tirarlo es a menudo difícil. ¿Querrá él usar el material antes de que sea reemplazado por algo más nuevo y de más valor para él?..

Antiguamente los cálculos eran muy complicados debido a que tenían que hacerse "a mano" lo que se debía primordialmente a la escasez de los materiales de escritura, el hombre hacía la mayor parte de sus cálculos mentalmente, tal vez con la ayuda de sus dedos. Una vez ideadas formas de cálculo más complejas, el entrenamiento del uso de los dedos era tan importante que se enseñaba en las escuelas romanas y se idearon varios métodos para las operaciones "avanzadas" tales como multiplicación y división.

. . .

El hombre quedó limitado hasta donde podía llegar con el conteo de sus dedos por lo que ideó un dispositivo manual que contenía cuentas ensartadas en una cuerda, al cuál se le llama ábaco o tabla de contar.

Posteriormente a partir del siglo XIV, vino el desarrollo de métodos auxiliares manuales en los cálculos escritos, entre este tipo de métodos podemos mencionar el método de contabilidad por partida doble, el cuál daba un balance de pérdidas y ganancias, otro método es el método del "emparrillado", el cual servía para hacer multiplicaciones. Se ideó también el método de "alineamiento" cuya función era reducir los cálculos de los números grandes. El método del "holgazán" sirvió para facilitar el hacer multiplicaciones y también en esos tiempos se desarrolló el sistema de "números arábigos".

DESARROLLO DE AUXILIARES MECANICOS PARA LOS CALCULOS ESCRITOS.

La calculadora de rueda numérica.- Debido a la expansión del sistema arábigo en Europa, las matemáticas empezaron a desarrollar dispositivos de computación para calcular a un nivel más alto que el ábaco y debido a esto se creó esta calculadora, así por el año de 1642, Blas Pascal fabricó una calculadora por medio del giro de una a nueve etapas de un sistema de ruedas dentadas.

. . .

Máquina de "cuatro funciones".- A esta máquina se le llamó así debido a que ejecutaba las cuatro operaciones aritméticas: suma, resta, multiplicación y división, fué diseñada por Leibniz.

Máquinas calculadoras accionadas por teclas.- La invención de las máquinas accionadas por teclas tales como las máquinas de escribir y máquinas registradoras tuvieron un papel muy importante en el avance del proceso de datos sobre todo en lo referente a la emisión de reportes.

Desarrollo de la tarjeta perforada.- En 1801, Joseph Marie - - - Jacquard, perfeccionó la primera máquina de tarjetas perforadas, cuya función era tejer diseños complicados sobre las telas. En -- 1833, Charles Babbage, inspirado en los mecanismos de los telares de Jacquard, concibió la idea de construír una máquina que fuera capaz de ejecutar cualquier cálculo, lo que sería la primera - - computadora digital para fines generales, pero que nunca completó. En 1890 Herman Hollerith diseñó un juego completo de máquinas para procesar toda la información del censo en Estados Unidos, se usó la que fué la primera instalación de máquinas de proceso de datos : con tarjetas perforadas, estas máquinas incluían una perforadora - de tarjetas, contadores electromagnéticos alimentados a mano y - - una caja clasificadora, usando esta instalación se logró un ahorro de 5 años para la emisión de los resultados del censo.

. . .

Todas las máquinas contadoras descritas anteriormente fueron digitales y aún cuando muchos dispositivos analógicos se conocieron en los principios de la historia de Europa Occidental se cree que la primera computadora analógica usada amplia y extensivamente fue la regla de cálculo inventada a principios del siglo XV por William Oughtred basado en los estudios de Napier. Sin embargo este tipo de computadoras no son de gran importancia para este desarrollo histórico, por lo que no profundizaré más en ellas.

Desarrollo moderno.- Computadoras de primera generación.- En el año de 1937 la Universidad de Harvard, bajo la dirección de Aiken y en colaboración con un equipo de ingenieros de la IBM, inició la investigación para el desarrollo de una computadora secuencial, siete años más tarde entró en servicio una calculadora electromecánica de secuencia controlada llamada Harvard Mark I, ésta era capaz de hacer sumas, restas, multiplicaciones, divisiones y comparar cantidades, así mismo podía hacer referencia a tablas almacenadas en ella previamente. Posteriormente, en el año de 1946 se terminó la construcción de la primera computadora totalmente electrónica, a la cual llamaron ENIAC (Electronic Numerical Integrator and Calculator), ésta fue desarrollada por Mauchly y Eckert en la escuela Moore de Ingeniería Eléctrica de la Universidad de Pennsylvania y bajo contrato con el ejército de Estados Unidos, entre sus características encontramos:

a) una suma tardaba 5 milisegundos.

b) pesaba 30 toneladas.

c) tenía 18, tubos de vacio (bulbos).

El mismo equipo que construyó la ENIAC, antes de que ésta entrara en operación, se reunió con John Von Neumann para diseñar una nueva computadora que se llamaría EDVAC (Electronic Discrete Variable - Computer). Esta máquina nunca fué construida, pero su diseño marcó el diseño futuro de todas las computadoras que le siguieron, siendo su principal característica la de incluir en su memoria el programa a ejecutar, teniendo además un controlador central que interpretaba las órdenes del programa.

A partir de 1952 se dió un gran auge al desarrollo de estas máquinas - primeramente decenas y luego cientos de éstas grandes y pequeñas fueron vendidas para fines comerciales.

Las computadoras de primera generación eran de un tamaño enorme bastante inflexibles y requerían un estricto control sobre las necesidades de aire acondicionado. Otras de sus características; fueron - construidas a base de bulbos, su programación era en lenguaje de -- máquina y posteriormente en lenguaje ensamblador, su memoria era a base de tableros de ferritas y tambores o cilindros magnéticos.

Vemos que dadas las limitaciones de estos equipos, debido al proceso distribuido y la falta de relación en línea entre el procesador central y los periféricos, el proceso de recuperación de información no era electrónico y por lo tanto lento.

COMPUTADORAS DE SEGUNDA GENERACION.

En esta etapa, gracias a que se introdujo al transistor dentro de las computadoras, el tamaño físico de éstas se redujo considerablemente sin disminuír su efectividad, el consumo de energía y la disipación de calor fueron abatidos drásticamente, de modo que los requerimientos de aire acondicionado fueron menos estricos, se mejoró el equipo secundario, tales como lectoras o impresoras y las técnicas más sofisticadas de programación avanzaron de manera significativa, gracias al desarrollo de los primeros lenguajes de alto nivel: Fortran en 1957, Algol, etc. En este tipo de computadoras electrónicas se empiezan a desarrollar algunos sistemas de recuperación de información, entre los que sobresale SABRE, sistema de reservaciones aereas de American Air Lines, desarrollado en 1964.

COMPUTADORAS DE LA TERCERA GENERACION.

Están caracterizadas por un mayor refinamiento en la programación y equipo periférico, así como en una más grande miniaturización del equipo, El uso más efectivo de los dispositivos de entrada y salida permite que las organizaciones almacenen todas sus operaciones y datos de funcionamiento. Algunas de sus características son:

a) Están construídas a base de circuitos integrados.

b) Su programación a nivel de máquina y superlenguajes posee compiladores muy rápidos.

c) Queda establecido el sistema operativo, el cual controla el ambiente del computador.

Dado que estos equipos ya son bastante flexibles y poderosos, la recuperación de información ya toma mucho más auge, debido a que en -- ellos se almacena y actualiza gran cantidad de información.

Algunos autores hablan de las computadoras de cuarta generación, las cuales son la última palabra en tecnología de computadoras y estan -- construídas a base de circuítos integrados.

Las fechas de nacimiento y fin de cada generación son las siguientes:

1a. generación: de 1944 a 1958

2a. generación: de 1958 a 1965

3a. generación: de 1965 a 1972

4a. generación: de 1972 a ?

. . .

II. PROPOSITOS DEL ALMACENAMIENTO DE CONOCIMIENTO.

El hombre casi siempre ha estado acostumbrado a hacer todas sus operaciones aritméticas manejando números decimales (base 10). Al empezar a usar las computadoras, éstas trabajan con números binarios (base 2), por lo que fué necesario desarrollar traductores de números decimales a binarios y viceversa. En un principio las computadoras tuvieron aplicaciones de cálculo numérico, sin embargo rápidamente se empezó a sentir la necesidad de manejar en ellas todo tipo de información, por lo cual se idearon códigos para manejar caracteres alfanuméricos, o sea que los caracteres más usuales del idioma, pudieran ser traducidos a un código binario, y de ahí -- que actualmente la función básica de una computadora es el cálculo numérico y el manejo de información.

La experiencia en tecnología, industria y gobierno prueba que la disponibilidad de los conocimientos es esencial para el mantenimiento de nuestra civilización, el almacenamiento y manejo de conocimiento en bibliotecas y archivos se ha transformado en un problema práctico, por los volúmenes de datos que ello implica.

Actualmente el proceso de toma de decisiones se está transformando esencialmente en un proceso de análisis de información, así mismo cualquier persona estudiante, profesionista, etc. en cualquier momento podrá tener requerimientos de información sobre un determinado tema, por lo tanto es importante proporcionar información selecta a grandes velocidades así como tener actualizada la misma, -

que produce la necesidad de tener un control sobre el volumen y tipo de información que se maneja, la forma de actualizarla, etc. La respuesta a este problema parece estar en estos equipos electrónicos modernos, así también estos equipos abren la posibilidad de contar con más información por lo que aceleran el desarrollo de esta problemática.

Recuperación mecanizada de información.-

El desarrollo efectivo y uso de cualquier sistema de cómputo de recuperación de información, debe estar basado en el claro entendimiento del material y los problemas que el sistema manejará. La experiencia en la recuperación de información indica que en general se deben manejar los siguientes tipos de material:

1. Archivos grandes, usualmente conteniendo más que algunos documentos. Por ejemplo: Titulos de libros y autores en una biblioteca pública.
2. Potencial de la información.- Dado que en la actualidad cualquier institución tiene requerimientos de información esto ha hecho que se creen bancos de datos de muy diversa índole, que además de contener lo más relevante para la institución, contiene información latente para explotarse.
3. Información de la cual se requerirá selección de entre muchos documentos y subsecuentemente su correlación.

Reconociendo esto como el material del sistema, nosotros podemos aceptar como su propósito, la identificación de documentos de acuerdo

a varios criterios dados por los requerimientos de la empresa ó institución.

¿Porqué el análisis y la recuperación mecanizada de información? - El campo de la recuperación de información se deriva de los requerimientos referentes a información de un tema determinado. Por ejemplo: Para un gerente de una empresa sus requerimientos de información pueden ser: volúmenes de ventas mensuales y acumuladas etc. Un estudiante puede tener requerimientos de información como: Características de la cultura Maya, lo que encontrará en una biblioteca.

Por lo tanto es muy importante la calidad y relevancia de la información así como también es fundamental tener acumulada y organizada la información con la que cuenta una institución.

La recuperación de información particularmente usando computadoras se ha desarrollado muchísimo en las últimas décadas debido a una serie de factores:

1. El lapso para disponer de información ha sido reducido drásticamente en estos últimos tiempos.
2. Ha habido un cambio drámático en la cantidad de información que está disponible, resultando la caracterización de la situación como una explosión de la informa ción, lo cual genera la imposibilidad de procesar para recuperación posterior la mayoría de la literatura de posible interés, así como hace inoperantes los métodos tradicionales de copiado con los requerimientos de detalle para los lectores en la

identificación de información pertinente hacia un problema dado.

Como resultado de esto, nuevos sistemas de comunicación, nuevos métodos de organización de información han sido propuestos y desarrollados. La velocidad con que las computadoras pueden buscar sobre archivos grandes textos completos, trae consigo un alto costo en el desarrollo de algoritmos que permitan a los programas identificar información significativa.

En un intento por amortizar el costo sobre muchos usos, se ha tendido a utilizar proceso en hornada * para manejar el mayor número de preguntas al mismo tiempo. Pero como resultado de esto, se ha decrementado la velocidad efectiva de las búsquedas. Esto ha llevado a consideraciones de cómo computadoras de tiempo compartido pueden ser utilizadas para dar resultados de búsquedas en tiempo real.

Así mismo la tecnología moderna de comunicación ofrece la oportunidad de transmitir información en forma de datos, voces e imágenes, usando esta tecnología las fuentes de información de las diferentes organizaciones e instituciones pueden ser compartidas, permitiendo el acceso mediante un sistema de redes adecuado.

III. PRESENTACION DEL PROBLEMA A RESOLVER.

Este trabajo de tesis, pretende presentar el análisis, diseño y desarrollo de un sistema computarizado de recuperación de información, recuperando textos y particularmente nombres de empresas, usando una computadora Burroughs B6700.

* Hornada. La traducción al español de la palabra batch en inglés.

Esto se hace trabajando sobre el Catálogo Básico de Empresas del - Instituto del Fondo Nacional de la Vivienda para los Trabajadores - (INFONAVIT), el cual es un archivo que cuenta con 400,000 regis-- tros, y contiene la información de identificación y contable de las - empresas que pagan aportaciones al Instituto y tienen trabajadores con crédito de vivienda. Por lo tanto existen algunas variables que son de interés, como los tiempos de respuesta de la consulta y el espacio en disco utilizado para los directorios de acceso.

ANALISIS Y RECUPERACION DE INFORMACION

PRINCIPIOS DE ANALISIS

Debe ser recordado que el propósito de un sistema de recuperación de información es facilitar la identificación de algunos registros que tienen una o varias características comunes, tales características son determinadas por la persona que especifica la configuración de la información.
Los principios del análisis son básicamente los mismos ya sea que el sistema de recuperación usado sea mecanizado o no mecanizado, las técnicas básicas pueden ser adaptadas a fin de llevar a cabo una adecuada penetración dentro de la situación del registro o se pueden usar características especiales de algún recuperador a fin de ahorrar dinero, tiempo o facilitar búsquedas efectivas. Inevitablemente se introducen algunas variaciones en las técnicas del análisis, a fin de controlar una o más deficiencias de un recuperador. Echando un vistazo a varios problemas particulares de recuperación de información podemos llegar a la siguiente conclusión: ningún análisis puede predecir todos los posibles puntos de vista o usos que pueden ser pedidos sobre una información dada, entonces, la mayoría de las veces podemos decir que prevee los casos más usuales.

TECNICAS DE ANALISIS

Los procedimientos de análisis pueden ser considerados en las siguientes categorías:

Indexamiento.

Clasificación.

Procesamiento de textos completos.

Indexamiento.

Un índice puede ser definido como un dispositivo que sirve como un apuntador ó indicador. Más a menudo es una lista alfabética que incluye temas y nombres de gente ó lugares ligados a una referencia, y que son considerados de especial interés en un archivo.

El propósito del indexamiento es facilitar la identificación o selección de documentos después que han sido clasificados y almacenados.

Las técnicas de indexamiento las podemos dividir en dos tipos:

1) Indexamiento de palabras

2) Indexamiento controlado

1.- Indexamiento de palabras.

Esta técnica debe representar un tipo de indexamiento que una computadora pueda ejecutar con precisión y consistencia

a) Concordancia. Es un indexamiento alfabético de palabras y no se da ningún tipo de discriminación en este tipo de indexamiento. Por lo tanto las decisiones tomadas, que se hacen en este tipo de indexamiento no son muy difíciles y puede ser ejecutado muy bien por computadoras.

Una técnica de computación que se usa haciendo concordancia de textos contínuos implica los siguientes pasos:

1) Se marca el texto indicando cómo se perfora en tarjeta haciendo notar el principio y fin de párrafos y oraciones.

2) Ninguna palabra se puede perforar en dos tarjetas diferentes, --

una palabra será perforada en una nueva tarjeta en caso de que no quepa en la precedente.

3) En la perforación se verifica y se corrigen los errores.

4) Las tarjetas se clasifican en orden alfabético y se imprimen si así se desea, estando así el indexamiento listo para ser usado.

En este tipo de indexamiento, el tipo de búsqueda que haremos sobre nuestro archivo será manual, dado que buscaremos sobre un listado.

Ejemplo:

Dos títulos de dos documentos se prepararon para una concordancia.

1) Inspección visual y fotografía de sitios.

2) Una descripción de programas espaciales.

Palabra	Número de documento
De	1,2
Descripción	2
Espaciales	2
Fotografía	1
Inspección	1
Programas	2
Sitios	1
Una	2
Visual	1

Analizando el ejemplo, vemos que es posible que fuese una herramienta más útil si el contexto de cada elemento estuviese dado junto con la palabra indexada.

Por ejemplo:

Palabra	Contexto	No. de documento
Una	Una descripción de programas espaciales	2
De	Inspección visual y fotográfica de sitios	1
Descripción	Una descripción de programas espaciales	2
Inspección	Inspección visual y fotográfica de sitios	1

La palabra indexada de la izquierda pudiera ser omitida si se pudiera consultar el texto.

b) Palabra-clave en contexto, o indexamiento de permutación.

Esta es una técnica que es posible mecanizar. En este caso sería muy costoso perforar artículos o textos enteros suponiendo que los autores de materiales fuente se esfuerzan considerablemente en preparar un título informativo, entonces la mayoría de las palabras llave usadas por los usuarios estarían contenidas en el mismo, por lo que sería recomendable sólo perforar el título.

La utilidad de un indexamiento depende de la manera en que se organiza la información. El establecimiento de categorías por temas u otras características apropiadas es el medio convencional gracias al que tal organización es llevada a cabo.

Puede haber diferencias de opinión en cuanto a la efectividad de un esquema u otro, pero el hecho más importante parece ser que cualquier esquema debe localizar la información deseada en el menor tiempo posible.

Este método está basado en la permutación cíclica de palabras.

Se han establecido reglas para tratar de diferenciar en los títulos las pala-

bras que son significativas de las palabras que no lo son para propósitos de recuperación. Dado que es muy difícil predecir cuales palabras serán significativas y cuales no, se ha decidido omitir aquellas tales como artículos, preposiciones, conjunciones, ciertos adjetivos y algunos nombres. Esta lista de palabras llamada "lista de exclusión" es almacenada en la memoria de la computadora.

Se perfora el título del documento y cada palabra del título se compara automáticamente contra la lista de exclusión, aquellas palabras que no están en ésta, serán consideradas como significativas y estarán listas -- para futuros procesos como palabras "llave".

Por ejemplo, tomamos el título INSPECCION VISUAL Y FOTOGRAFICA DE SITIOS. En este caso, las posibles palabras "llave" que tendríamos serían INSPECCION VISUAL FOTOGRAFICA SITIOS.

Estos registros son ordenados alfabéticamente por palabra llave, impresos si así es deseado, y quedan listos para ser consultados.

También se puede proveer una lista de títulos ordenada por número de - documento a fin de proveer una información bibliográfica completa. Para dar mayor facilidad en la búsqueda de artículos es conveniente imprimir la palabra "llave" a la izquierda del reporte.

Para facilitar el proceso de clasificación sería conveniente definir zonas en el registro a digitar o sea que cada palabra solamente será digitada - en cierta parte del registro.

Ejemplo:

← ZONA 1 →	← ZONA 2 →	← ZONA 3 →
INSPECCION VISUAL	FOTOGRAFICA	
DESCRIPCION	PROGRAMAS	ESPACIALES

Y de esta forma también podemos ejecutar la clasificación de varias formas o sea tomando diferentes campos como llave.

c) Indexamiento de términos unitarios.

Esta técnica implica el análisis del contenido de registros en términos de palabras "llave" que representan el contenido del registro que está siendo indexado. Estas palabras "llave" incluyen no solamente palabras comunes del idioma, también números de serie y otros símbolos, si éstos son encontrados en el texto y si, a criterio del analista son significativos para el registro.

Demos un ejemplo: este es un reporte que ha sido analizado y los términos unitarios se han puesto a la derecha: LAS MATRICES D^{-1} y G^{-1} EN LA TEORIA DE VIBRACIONES MOLECULARES.

Reporte	Términos Unitarios
Un método vectorial es dado para determinar los elementos de las matrices --	Matriz D^{-1}

Reporte	Términos unitarios.
y G^{-1} el cual ocurre en el estudio del	G^{-1}
espectro de la vibración rotacional de	Teoría
moléculas poliatómicas. D^{-1} es la -	Moléculas
matriz de transformación dando el des	Vibración
plazamiento cartesiano de masa y G^{-1}	Vector
tiene como elementos los coeficientes	Poliatómico
en la expresión de la energía vibracio-	
nal en términos de velocidades.	

El analista puede añadir libremente a la lista de términos unitarios otros términos; el analista no especifica el orden de las palabras "llave" o la relación entre ellas.

El usuario de este sistema debe estar suficientemente familiarizado con los temas, de tal forma que pueda tener control sobre los términos unitarios, o sea el usuario debe compensar la falta de orden ó relación entre las palabras usadas en este indexamiento.

Vemos sin embargo que para poder aplicar este tipo de indexamiento, se requiere de un análisis de cada uno de los textos, el cual llevaría mucho tiempo efectuarlo y sería muy costoso, además de que la información únicamente la podrían utilizar personas familiarizadas con los temas para poder entender la relación entre las palabras.

2.- Indexamiento controlado.

Este es opuesto a la técnica de indexamiento de palabras e implica una selección cuidadosa de la terminología usada en indexamientos a fin de

evitar lo más posible la dispersión de temas bajo diferentes cabezas. Se cree que el aspecto más atractivo de este indexamiento es identificar aspectos específicos de la información que pueda ser discutida en un documento. Pero también existe el deseo de combinar alguna de las ventajas de la clasificación. En un intento por conciliar estos dos conceptos en un indexamiento, hay una tendencia que trata de establecer control y más control sobre el uso del lenguaje "permisible" en el almacenamiento de resultados del análisis de registros. Se usan varios métodos a fin de establecer "control" sobre la operación de indexamiento:

a) En los temas que puedan ser escogidos.

b) En el número de aspectos que pueden ser escogidos.

c) En el lenguaje usado para expresar los resultados del análisis.

a.- Control sobre cuales temas serán escogidos.

Este tipo de control puede ser establecido en base a dos criterios:

1) Temas de interés existentes o predecidos de aquellos que serán usuarios potenciales del indexamiento.

2) Naturaleza de los registros que serán analizados.

El primer criterio puede ser establecido mediante una lista de temas que dé puntos de vista y requerimientos de cómo éstos deben ser expresados.

El control impuesto por esta lista puede ser dado de varias formas.

1. Al usuario se le puede pedir que use la lista de temas como una guía para toma de decisiones en la selección de registros.

2. Al usuario se le puede permitir grabar registros, llevando un control de lo grabado en forma de diccionario que se consultará manualmente o bien con la ayuda de la máquina.

Otra forma de implantar el primer criterio es el llamado "indexamiento probabilístico". En este método se decide, durante la operación de indexamiento que sólo hay una probabilidad de que el usuario encontrara que el documento en cuestión es relevante para un cierto tema. Puede resultar mucho más real y razonable aceptar que cada tema puede caer dentro de un cierto grado o puede tener un cierto peso. Desarrollando la habilidad de darle peso a cada tema, el codificador puede caracterizar más la información de cada documento, entonces el codificador puede asignar un peso bajo tal como 0.1 ó 0.2 a un término, más que decir que el término no es importante para el documento.

El codificador también puede asignar un peso alto tal como 0.8 ó 0.9 a un término dentro del documento, más que decir que éste es definitivamente importante dentro del documento usado. Con estos "pesos" es posible caracterizar

mejor la información contenida en un documento y este "peso" puede ser explotado durante la búsqueda que la máquina hace.

A pesar de que existen criterios bien definidos, vemos que dado que esta labor es hecha por un codificador, es una labor subjetiva, o sea depende del criterio de la persona y las condiciones en las que la hace.

Otro método de control de temas escogidos durante el análisis está basado en la frecuencia de la ocurrencia de las palabras llave en los textos; aquellas que ocurran más frecuentemente serán consideradas las más significativas para el análisis de los textos:

Este tipo de análisis puede ser llevado al cabo de la siguiente manera:

Se perforan los documentos de tal forma que estos puedan ser analizados por la computadora y ésta pueda sacar estadísticas acerca de la frecuencia de las palabras llaves. Las palabras encontradas analizando los textos perforados son comparadas contra una lista de palabras con el fin de ver cuales no son suficientemente discriminatorias, o sea, que son demasiado comunes para ser significativas; cuando éstas son descartadas, entonces la frecuencia de ocurrencia del resto de los términos puede ser usada como base para el indexamiento controlado. El segundo criterio para el control de temas escogidos es implantado permitiendo que los registros " se indexen ellos mismos". Esta expresión implica que la naturale-

za del tema dictará el desarrollo de los controles en el indexamiento.

Esto se logra solamente después de acumular una colección representativa de entradas, de tal forma que el tema probable del que trata el archivo puede ser descubierto y una política de indexamiento puede ser formulada y llevada a cabo.

b.- Control sobre el número de temas escogidos.

Un texto sujeto a ser catalogado y clasificado, nos da el siguiente resumen:

Una tarjeta al catalogarse nos da una descripción concisa de un libro o un conjunto de libros. Dado que el tamaño standard de las tarjetas es pequeño, la descripción del libro se limita a algunos términos, se dan términos convencionales y se arregla de una cierta manera. Los primeros términos usualmente dados son:

1) Número del anaquel donde se encuentra el libro.
2) Nombre del autor
3) Título de libro, incluyendo editorial
4) Fecha y lugar de la publicación y
5) La descripción del libro, dando el número de páginas o número de volumen si es que lo hay, mencionando ilustraciones, mapas y portadas.

Entonces tendremos un buen catálogo donde podremos encontrar un li-

bro por medio de diferentes formas según la que queramos. Aunque la aplicación de recuperación de información en computadoras nos - da "muchas formas" para seleccionar la información deseada, usualmente hay un límite en los criterios a ser usados y por lo tanto se hacen explícitos como una referencia para la búsqueda.

c.- Control sobre el lenguaje usado.

Otra variable independiente en el indexamiento controlado es el lenguaje a ser usado en la grabación de los resultados del análisis de registros.

En muchas formas ésta variable es completamente análoga al control de temas escogidos.

Esto puede ser pensado, dando un indexamiento con un punto de vista especial, por ejemplo, un documento que trate sobre metalúrgica puede ser indexado independientemente desde dos puntos de vista: documentación y metalúrgica, como resultado el documento puede tener dos conjuntos de entradas.

Independientemente del control de temas escogidos o del "punto de - vista" de un analista,es útil regularizar la manera mediante la cual se expresan las entradas a un indexamiento.

Algunos de estos métodos son idénticos a aquellos usados en el control de temas escogidos (lista de temas que utiliza el usuario). Sin embargo un exámen cuidadoso de la lista de temas algunas veces hace posible comparar el grado de control de la lista contra el control

del lenguaje inherente a la lista.

El control sobre temas escogidos es llevado al cabo por medio de referencias de sinónimos, casi sinónimos y términos similares sobre un solo encabezado, el control sobre el lenguaje se hace referenciando únicamente por medio de sinónimos sobre un solo encabezado.

Otra técnica que ha sido usada es describir los límites de los temas de cada encabezado mediante un "indicador de su función". Estos indicadores son útiles para limitar el significado de cada entrada, de acuerdo a la función que cada entrada juega en un contexto particular.

Consideramos un documento que discute las formas en las cuales se procesa el mineral de hierro a fin de producir hierro puro.

Un analista lee este documento a fin de dar una entrada al indexamiento.

Mineral de hierro, uso en la preparación de hierro puro, 78.

El término "mineral de hierro" fué puesto como la notación principal y se pudo haber asignado "materia prima" como indicador de su función, lo que especifica un punto de vista particular a partir del cual el mineral de hierro es tratado en ese documento.

3.- Indexamiento de referencias bibliográficas.

La filosofía de este método es la suposición de que el autor de un artículo cita el trabajo de alguien más, en una referencia que tiene alguna relación entre su trabajo y el de la persona citada. Por lo tanto, todas las referencias citadas en la literatura son incluídas juntas en el directorio, cada referencia acompañada por una lista de documentos fuentes citados.

El arreglo elemental del indexamiento puede ser por el autor de la referencia, creando un indexamiento de autores citados; por publicación en la cual la referencia aparece, creando un indexamiento de artículos citados, ó por fecha de publicación de la referencia, creando un indexamiento cronológico de las referencias.

Por ejemplo:

Una publicación de Richard Roe:

Richard Roe Biol. Chem., 6, 103 (1978)

citado por John Doe en un artículo publicado en Biol. Chem, 7, 55 -- (1979), aparecería con un indexamiento de referencia bibliográfica del autor, listado como:

Roe, Richard, y con el autor que lo cita (John Doe) listado inmediatamente después, junto con algunos otros autores que pudieron haber citado su trabajo. Las referencias bibliográficas para cada elemento del indexamiento serán dadas también.

La totalidad de las referencias bibliográficas, junto con una referencia al artículo en el cual aparecen almacenadas dentro de una computadora, permiten la emisión de listados para ser consultados.

Clasificación.

La clasificación puede ser definida como un arreglo sistemático de términos acorde a un plan definido ó a una secuencia definida.

Dicho de otra forma, la clasificación puede comprenderse como el arreglo ó colocación en una clase o clases sobre la base de semejanzas.

Una clase puede ser definida como algo que consiste de cualquier elemento específico. Cualquier cosa puede ser elemento de una clase. Por ejemplo, podemos tener tales clases como:

1. La clase de comida que es color verde
2. La clase de comida que contiene más del 10% de proteína.
3. La clase de objetos verdes que son comestibles.
4. La clase de proteínas que son comestibles.

Una clase puede ser definida aún cuando sus elementos existan o no. Por ejemplo: la clase de planetas que estan después de la tierra donde hay vida. También es posible reconocer la existencia de clases, las cuales no tienen elementos por ejemplo, la clase de mujeres de más -- de 10 m. de altura.

Una clase también puede ser miembro de otra clase definida previamente.

Se han usado dos tipos de clasificación: la rígida o monodimensional y la no rígida o multidimensional.

1. Clasificación Rígida.

   Esta clasificación es la caracterización de cada registro desde un solo punto de vista.

   Cuando un registro va a ser almacenado, solo una localidad física puede ser dada para este registro. Por supuesto siempre se reconoce que los registros son multidimensionales por naturaleza.

   De aquí que la clasificación de registros a menudo tiende a suponer

las características de una clasificación no rígida.

2. Clasificación Multidimensional

La clasificación multidimensional comprende la caracterización de cada registro desde más de un punto de vista ó llave. Este método sólo es permisible cuando existan más de un punto de vista ó llaves en especial iguales.

Se han hecho intentos experimentales para lograr la clasificación - automática. Por ejemplo:

Un documento que contiene las palabras: niño, niña, maestra, escuela, aritmética, lectura: probablemente trate de educación.

Los experimentos se hacen de la siguiente manera:

Se hace un conjunto de procedimientos por medio de los cuales los documentos pueden ser clasificados automáticamente dentro de sus categorías, y se determina la precisión de la clasificación mediante la comparación, dado un criterio.

Un conjunto de palabras se usaron para el experimento: las categorías de la clasificación fueron dadas por medio de una técnica de análisis; esta técnica se hace ordenando las palabras mediante un programa de computadora, a fin de saber su frecuencia en los textos. Las palabras con más frecuencia de ocurrencia son analizadas produciendo así las categorías de la clasificación. Se determinó que el 48.9 por ciento de las palabras fueron colocadas en su categoría correcta por este método, sin embargo se sigue experimentando para lograr mayor precisión en las técnicas de clasificación automática.

Abstracción.

Tradicionalmente la abstracción ha sido considerada como aquello - que concentra las cualidades esenciales de algo. Una abstracción es un resumen de una publicación o un artículo acompañado por una descripción bibliográfica adecuada a fin de que la publicación o artículo pueda ser rastreada o encontrada fácilmente. Dado que generalmente la abstracción es llevada al cabo por analistas de información, es entonces subjetiva.

Es posible identificar 3 tipos de abstracción:

1) Tradicional
2) Extractada o resumida
3) Estilizada.

1. Abstracción Tradicional.

En la práctica, se distinguen dos tipos de abstracción tradicional: la descriptiva y la informativa. La descriptiva incluye un texto general de la naturaleza y alcance del documento. No se pretende que este tipo de abstracción pueda servir como un sustituto para la lectura del documento original. Por otra parte, la abstracción informativa tiene el propósito de presentar información significativa que probablemente contendrá el registro original.

Idealmente, la abstracción informativa producirá la necesidad de referirse al registro original.

La abstracción descriptiva a menudo consiste de una sola frase u -- oración elaborada a partir del título del documento. Cuando se trata con temas de considerable complejidad y a fin de lograr que la - abstracción tenga significado, la referencia se hace a alguno de los

procedimientos, conclusiones y resultados de las investigaciones más significativas.

La abstracción informativa se escribe a fin de proveer un resumen conciso y significativo del contenido del documento. Un conjunto de reglas para preparar una abstracción informativa es la siguiente:

a. Indicar el alcance y objetivos del título (si no son evidentes)
b. Resumir toda la nueva información.
c. Establecer los principios esenciales de nuevos métodos o equipos y de las conclusiones, pero sin irse al detalle.
d. Citar aplicaciones nuevas o especiales.
e. Referirse a los hechos y no ser crítico.
f. Ser específico e informativo.
g. Ser breve.

Actualmente las abstracciones varían desde un título muy pobre hasta uno muy completo, y desde una o dos líneas de anotación hasta una anotación larga. El tercer tipo de abstracción tradicional es la pseudo abstracción y es descrita como la abstracción de un texto que no ha sido y tal vez nunca será escrito. Esto se deriva del hecho de invitar oradores y hacer resúmenes de sus pláticas, las cuales posiblemente nunca sean publicadas.

Las funciones de la abstracción tradicional son las siguientes:

a. Servir como un substituto del material fuente para el lector.
b. Servir como una forma de predicción con el fin de valorar si la lectura del documentos fuente valdrá la pena para el lector,

Esto es, poder separar documentos que son relevantes de -- aquellos que son irrelevantes para un interés particular del lector.

Es la segunda de estas funciones la más considerada cuando el sistema de recuperación de información es diseñado.

Daremos un ejemplo de abstracción informativa y descriptiva del mismo artículo:

W.L. Wyman, "vacío generado por la fusión de rayos de electrones", publicado en febrero de 1958.

El artículo es el siguiente:

Un nuevo proceso de fusión es realizado mediante el bombardeo - con un rayo de electrones en una cámara de alto vacío. Los elementos básicos son un cátodo de tungsteno para emitir un gran número de electrones.

Alta potencia: varios miles de volts, entre el cátodo y el plato para acelerar los electrones, un sistema de foco para formar los - electrones en el rayo.

Abstracción informativa: una cámara de vacío para mantener la presión de 5 x 10^{-3} mm de mercurio. La mayor parte del trabajo se ha dirigido hacia el desarrollo del procedimiento de soldado para el entubamiento zircaloy-2, molibdeno, titanio, nickel y aleación de acero inoxidable.

Abstracción descriptiva: "Fusión de zircaloy-2, por medio de un rayo de electrones en una cámara de alto vacio. La técnica también -

se ha aplicado a las aleaciones de tugsteno, molibdeno, titanio, nickel y acero inoxidable".

2. Extracción.

Una extracción es análoga a una abstracción, dado que ellas representan qué será considerado por el analista a ser un tema importante de un registro.

Algunas personas piensan que el uso de extracciones da un mejor servicio al lector que una abstracción. Sin embargo, en una abstracción es posible presentar usualmente, en una forma más concisa, una descripción más completa del contenido de un registro, que lo que nos presenta una extracción del contexto.

La extracción puede ser hecha aplicando técnicas de computación ya sea automatizadas o por medio de analistas, cuando estas técnicas se usan para la extracción, al resultado se le ha llamado "auto-abstracción".

Las técnicas usadas por analistas para preparar extracciones son subjetivas,y se basan en juzgar el documento a fin de determinar cual porción de este es suficientemente significativo para garantizar su almacenamiento.

Cuando las técnicas de computación se usan para la extracción, todo texto es convertido en una forma tal que la máquina lo puede leer, entonces se analiza por la computadora digital y se asume que aplicando estos métodos, la frecuencia y distribución de las palabras llave en el texto pueden ser usadas como una base para determinar el significado de las oraciones dentro de un texto.

Siguiendo esta suposición, las oraciones que tienen más significado son impresas a fin de producir una extracción.

Ejemplo: Este ejemplo es hecho a partir de un artículo del New - York Times.

"La química es empleada en la investigación de nuevos métodos para la conquista de las enfermedades mentales.

Por coincidencia, este fin de semana en Nueva York finaliza la Convención Anual de la Asociación Psicológica Americana y es el principio de la Convención Anual de la Sociedad Química Americana.

Los psicólogos y químicos nunca han tenido mucho en común, como ahora tienen en los nuevos estudios de las bases químicas sobre el comportamiento humano.

Fueron hechos nuevos descubrimientos muy importantes la semana pasada, en la Convención Anual de la Sociedad Psicológica Americana y en Zurich Suiza, en el Segundo Congreso Internacional de Psiquiatría.

Dos desarrollos recientes sobre enfermedades mentales han llamado la atención de los químicos, psicólogos, físicos y otros científicos. Se ha descubierto que pequeñísimas cantidades de productos químicos pueden provocar alucinaciones y disturbios en gente normal y drogas que alteren el humor de las personas, por ejemplo tranquilizantes, han hecho que la gente se someta a la terapia.

El dinero para financiar la investigación de los factores físicos en las enfermedades mentales está disponible.

Los estudios están siendo encaminados hacia el entendimiento de la química del cerebro.

Se han trazado nuevas metas.

En el Congreso de Psiquiatría en Zurich la semana pasada, cuatro físicos neoyorkinos pidieron a sus colegas ampliar su concepto de "enfermedad mental" y probar más exhaustivamente la química y metabolismo del cuerpo humano para responder a desórdenes mentales y su previsión.

El Doctor Félix Marti Ibañez y tres hermanos, Doctor Martiner D. Sackler, Doctor Raymond R. Sackler y Doctor Arthur M. Sackler dieron evidencia de que la química de la sangre de víctimas de esquizofrenia es diferente a la de gente normal. Sugirieron que posiblemente factores biológicos múltiples son responsables de éste - cambio químico.

Dijeron: la enfermedad mental es un "proceso de desarrollo mental" y una larga duración de un desorden, puede dar como resultado una alteración permanente en la anatomía y fisiología. Pidieron que a las pruebas de nuevas drogas que afecten el cerebro se les apliquen estudios sobre su mecanismo de acción. La variación de substancias capaces de producir efectos mentales es un nuevo arsenal de herramientas para la investigación de los mecanismos biológicos en los cuales se fundamentan las enfermedades mentales. Las fuentes que provocan los disturbios en el comportamiento son muchas y

pueden venir de fuentes externas o internas. Este concepto ya fué probado en la práctica, cuando se capacitó a los psiquiatras para verificar que la administración de ACTH y cortisona produciría psicósis.

También dijeron que hace algunos años se desarrolló un exámen de sangre, que con un 80 por ciento de precisión permitía la identificación de pacientes esquizofrénicos.

Es permitido en el campo de la psicología negar que la psicósis tenga mayores o menores grados del proceso de enfermedad.

Química del cerebro.

En la convención de Psicólogos, una técnica para el señalamiento de la actividad eléctrica en porciones específicas de cerebros animales, fueron descritas por investigadores de la Universidad de California. Reportaron que se usaron cerebros de gatos para registrar descargas eléctricas.

De esta forma el grupo de California reportó que es posible rastrear la secuencia en la cual el cerebro obtiene fases para el aprendizaje. Se pueden localizar áreas específicas de memoria en el cerebro.. Las rutas eléctricas rastreadas pueden ser bloqueadas temporalmente mediante el uso de productos químicos. Esto da nuevas posibilidades para el estudio de los cambios químicos del cerebro en la salud y enfermedad y su curación, enfatizaron los investigadores de California.

Los nuevos estudios de la química del cerebro han dado resultados prácticos en terapia y una tremenda motivación para aquellas personas que deben cuidar pacientes mentales.

Una evidencia que el conocimiento en campos interdisciplinarios es acumulada más rápidamente vino la semana pasada en un anuncio de Washington.

El Instituto Nacional de Salud Mental, estableció una casa de información de Psicofarmacología. La literatura sobre esto sería clasificada y codificada de tal forma que los miembros pudieran contestar una amplia variedad de preguntas científicas y técnicas.

La gente que trabaja en este campo está invitada a enviar tres copias de artículos u otro material, aún siendo cartas informales que describan el trabajo que están desarrollando, a la unidad de información técnica".

Los párrafos subrayados fueron los que el analista tomó para hacer la extracción. Esta extracción no parece una mala representación del artículo de la cual fué tomada.

El análisis en computadora debe estar basado en el análisis de aquellas palabras que el escritor repite conforme avanza el argumento de su artículo y elaborado sobre varios aspectos del tema. Proposiciones de este sistema asumen una relación directa entre la frecuencia de ocurrencia de las palabras y el significado de las oraciones que dan el contexto.

Dado que los primeros experimentos de auto-abstracción se hicieron sobre material periodístico, por ejemplo: artículos de periódico, es posible asumir cierta relación entre la frecuencia de ocurrencia de palabras y el significado del texto de las oraciones.

Puede ser recalcado que los periodistas tienen la tendencia a repetir el mensaje básico de su historia tres veces: de una forma concisa en el primer párrafo, desarrollando el tema en los siguientes y expandiéndolo en el último párrafo de la historia. La tendencia es entonces: las palabras y términos del mensaje básico, serán usadas varias veces después.

Debe ser recalcado que los resultados obtenidos a partir de los artículos periodísticos no necesariamente serán los mismos que los obtenidos a partir de los escritos por personas especializadas. Los autores en esos campos no necesariamente emplean las mismas reglas para la exposición.

3. Abstracción estilizada.

a. Introducción.- En una sección anterior la abstracción fué discutida en términos de su uso por seres humanos teniendo varios propósitos. Esta sección presenta la discusión de algunos esfuerzos que se han hecho a fin de lograr el desarrollo de la abstracción estilizada, logrando así:

I. Incrementar la consistencia, con lo cual las abstracciones son preparadas.

II. Facilitar su comprensión por los lectores.

III. Servir como un indexamiento, particularmente para los -

sistemas mecanizados de recuperación de información.

b. Abstracción formateada.- La abstracción tradicional es escrita en una forma narrativa. Es instructivo intentar escribir una -- abstracción de un papel y descubrir qué difícil es decidir que parte es suficientemente significativa para ser incluída, qué fácil sería si no fuera necesario encararse a una hoja totalmente en blanco - donde la abstracción sería escrita y tener algunos encabezados co- como guía.

Algunas guías generales serían propósitos, procedimientos, búsquedas, las cuales serían usadas como encabezados.

Ejemplo: Consideremos el siguiente artículo:

Ronald J. McBeath un estudio comparativo del efecto de películas, películas con sonido y película con diagramas para la enseñanza de hechos y conceptos.

Proyecto No. 462. Universidad del Sur de California.

Las guías serían entonces:

Propósito del estudio:

Comparar la efectividad de varios medios para la enseñanza de hechos y conceptos.

Procedimientos.

Una lección de estudios sociales fué preparada por medio de una película sin sonido, una película con una narración grabada y una película con diagramas de 16 mm. Se hicieron pre-exámenes, - post-exámenes y exámenes de retención a 558 personas del condado de los Angeles, agrupados de acuerdo a su IQ, edad, sexo

y situación socio-económica del padre.

Investigaciones.

El efecto del sexo e IQ en la ejecución y la interacción entre sexo y método de presentación son discutidos.

Es importante hacer notar que el uso del formato no altera la naturaleza de la abstracción, esto es, sigue siendo una abstracción indicativa más que una informativa.

Es evidente que cuando se da un formato demasiado detallado, la información no puede estar disponible en conexión con cada encabezado. También el material fuente a ser procesado por el analista debe ser bastante similar en caracter y forma a fin de hacer factible el uso de tal formato.

c. Abstracción Telegráfica.- Otro paso en la evolución de la abstracción estilizada puede ser definido como la abstracción "Telegráfica". En este tipo de abstracción se dan un conjunto de encabezados ó indicadores de su función a ser usados en conjunción con un tema seleccionado por el analista del material fuente. Uno ó más del conjunto de encabezados deben ser usados en el orden que el analista considere apropiado, dependiendo del tipo de material fuente.

Usando indicadores de función, un analista puede preparar una abstracción telegráfica del mismo ejemplo dado para la abstracción formateada.

Indicador de Función	Descripción
Tipo de literatura	Investigación
Fuente del soporte financiero	Acción nacional de defensa de la educación.
Población.	Estudiantes (558)
Localidad de la investigación	Escuela elemental intermedia en el Condado de los Angeles.
Atributos de las formas de estudio	IQ Edad Sexo Estado socioeconómico
Tema de enseñanza	Estudios sociales
Medio	Películas Películas con diagramas
Atributos del medio	Sin sonido Con sonido.
Etc.	

La abstracción "Telegráfica" está compuesta de:

1) Palabras significativas seleccionadas a partir del documento.
2) Encabezados, que proporcionan un contexto de las palabras seleccionadas.
3) Símbolos de puntuación, los cuales separan y agrupan las palabras y encabezados en varias unidades.

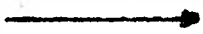
La abstracción "Telegráfica" es un método para almacenar características importantes de la información contenida en documentos, tal que estas características sean procesadas por un computador a fin de hacer un indexamiento. De esta forma el documento será habilitado para poder ser identificado por la computadora en respuesta a

los requerimientos de información.

El propósito de la abstracción "Telegráfica" es entonces proveer una entrada a la computadora en una forma consistente a fin de que la computadora pueda ser programada para buscar arreglos de información de esa entrada.

d. Abstracción esquematizada.- Otro paso evolucionario en la abstracción formateada pero de una manera flexible, es la abstracción esquematizada. Es bastante similar a la abstracción "Telegráfica" en la filosofía básica. Las diferencias son en formato, definición de los encabezados y el simbolismo usado para representarlas. Para la abstracción esquematizada, los arreglos de la entrada y encabezados, representan un intento para proveer más exactitud y rápido entendimiento de las relaciones discutidas sobre el material fuente.

Podemos considerar los siguientes indicadores de función:

Indicadores de función	Representación simbólica
Dirección	
Influencia	
Sin influencia	
Incremento	
Decremento	
Relacionado	
No correlación	
Igual	=

Indicadores de función	Representación simbólica
No igual	ϴ
Menos	-
Más	+
Mayor que	>
Menor que	<

4. Indexamientos Estilizados.

a. Introducción. Tal como las abstracciones han sido estilizadas para incrementar consistencia y confiabilidad en la forma que su explotación se da, y mejorar el procedimiento automático, los indexamientos se han transformado más estilizados por razones análogas. Las técnicas del indexamiento estilizado han sido tomadas de las técnicas de la abstracción estilizada.

El razonamiento para el indexamiento estilizado puede ser derivado de los problemas encontrados y estuvo en operación como un -- sistema de recuperación de información no convencional.

Como puede ser obvio, del término "indexamiento en profundidad" implica la selección, durante el análisis de materiales fuente, de tantas entradas que pueden ser consideradas útiles como puntos de referencia en la búsqueda. Los sistemas de recuperación de información no convencionales consideran la facultad de buscar convenientemente sobre una base multidimensional.

Desarrollemos un ejemplo de un problema:

En un sistema de recuperación, tres documentos contienen la siguiente información.

Documento 1. Discute la cría de borregos en Australia.

Documento 2. Discute la cría de caballos en EE. UU.

Documento 3. Compara la cría de borregos en Australia con la cría de caballos en EE.UU.

Se puede listar un indexamiento alfabético convencional para los tres documentos de la siguiente manera:

AUSTRALIA

Cría

de caballos 3.

de borregos 1.

Cría

Caballos

en Australia 3.

en EE. UU. 2.

Borregos

en Australia 1.

en EE. UU. 3.

Caballos

Cria

en Australia 3.

en EE. UU. 2.

Borregos

Cría

en Australia 1.

en EE. UU. 3.

ESTADOS UNIDOS.

Cría

de caballos 2.

de borregos 3.

En este tipo de indexamiento, refiriéndonos al documento número 3 la información se cruza, o sea es posible formular una pregunta acerca de: cría de caballos en Australia, obtener como respuesta la referencia del documento 3, sin que dicha respuesta sea la adecuada, dado que tal documento no trata ese tema.

Por otra parte si se deseara descubrir en cual documento se discute la cría de borregos en Australia, sería posible referirse alternativamente a las siguientes cabezas:

Australia

Cría

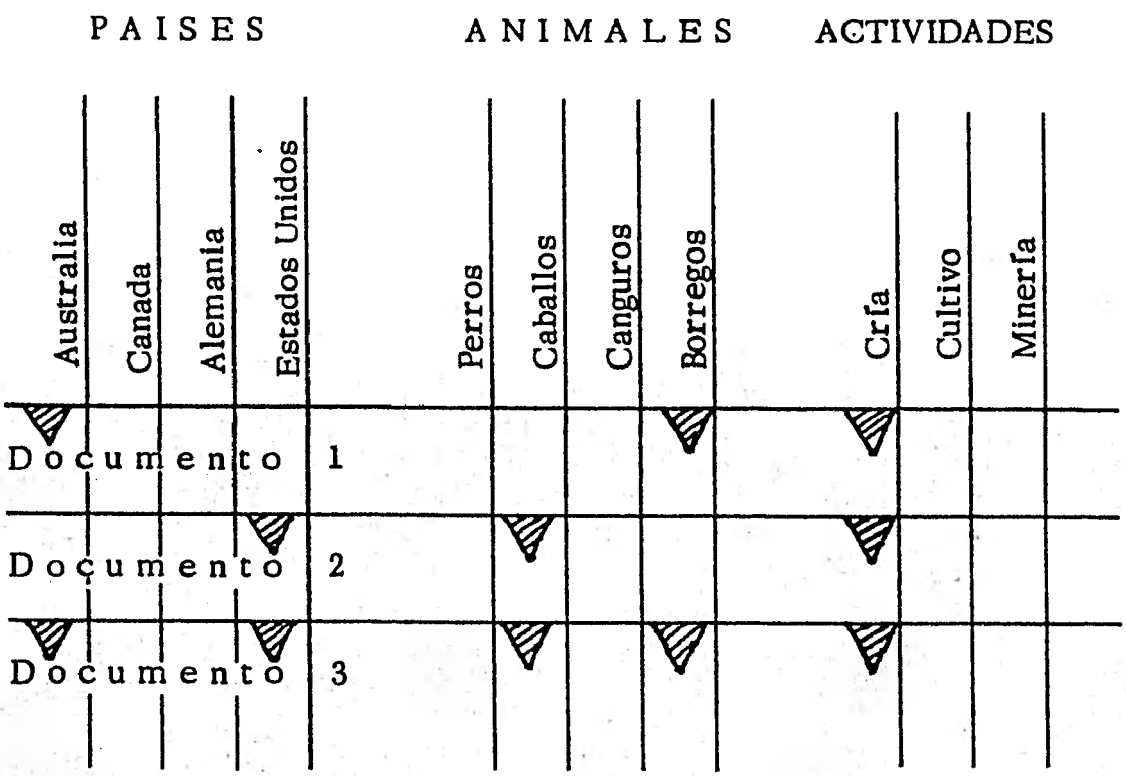
ó

Borregos

y al llegar rápidamente a la conclusión de que solamente el documento 1 sería de interés para la pregunta.

De otra forma, si se implantaran tarjetas perforadas para resolver este problema de recuperación de información, los documentos se

digitarían de la siguiente manera:



Si desearamos descubrir en cual documento se discute la cría de - borregos en Australia, el documento 1 sería identificado, pero el documento 3 también sería identificado aún cuando no contenga una respuesta propia a la pregunta hecha.

Se han desarrollado técnicas para controlar este tipo de "ruido" en sistemas de recuperación de información basados en indexamiento en profundidad.

b. Encadenamientos.- Otra forma de indexamiento estilizado propone el "control" de este tipo de "ruido" en las operaciones de recuperación por medio de una técnica llamada "encadenamientos". Esta técnica análoga a la "puntuación" de la abstracción telegráfica provee un método para el almacenamiento del hecho que en el documento 3 del ---

Ejemplo anterior, la cría de borregos en Australia, se discute aparte de la cría de caballos en Estados Unidos.

Esto puede ser almacenado de la siguiente manera:

Entrada	Identificadores de encadenamiento	
Cría	A	B
Caballos	A	
Borregos		B
Australia		B
Estados Unidos	A	

Si esto fuera grabado en una cinta magnética para que se efectuara la búsqueda por medio de la computadora, se haría de la siguiente manera:

3; Cría A, B; Caballos A; Borregos B; Australia B; Estados Unidos A.

Entonces sería posible programar una computadora para buscar las tres entradas deseadas (Cría, Borregos, y Australia).

La técnica de encadenamiento también puede ser explotada en los -- sistemas de recuperación manuales. En el ejemplo dado, cría, borregos y Australia puede ser considerado como la entrada para el - subdocumento 3 B, mientras que cría, caballos y Estados Unidos -

sería considerado como la entrada para el subdocumento 3 A. Desde el punto de vista de sistemas, cada número de subdocumento puede ser considerado como si fuera un número independiente de documento. Cuando se usa una computadora, el número de subdocumento no necesita ser usado; más bien, los encadenamientos pueden ser explotados mediante la programación.

c. Funciones.- Como se discutió anteriormente, es posiblemente obvio que las entradas puedan tener diferentes funciones dentro del documento.

Por ejemplo: Una entrada de un compuesto químico particular puede representar el compuesto químico cuando se discute como una materia prima o cuando se presenta como un producto terminado.

Como un ejemplo adicional, el nombre de una personalidad mencionada a un documento puede haber sido indexado por nombre sin importar la función de esta personalidad en el documento en cuestión. Esta personalidad puede ser el autor del documento, puede ser el tema del documento o puede ser discutida por otra personalidad, la cual es el tema principal del documento.

Si una búsqueda se hace sobre el nombre de la personalidad, puede darse que se puedan identificar muchas referencias inapropiadas, - si el usuario desea localizar solamente aquellas referencias de una personalidad particular en las cuales tiene una función específica. En estos casos, puede ser factible establecer procedimientos de -

análisis, los cuales harán explícitas las funciones de la entrada. Un conjunto de encabezados ó indicadores de función han sido propuestos para usarse con los encadenamientos. Algunos de estos - son los siguientes:

1. Tema principal
2. Materia prima
3. Producto
4. Desperdicio
5. Aplicaciones posibles
6. Medio ambiente
7. Causa
8. Efecto
9. Medio para llevar a cabo un objetivo primario
10. Bibliografía y otros datos fuente de identificación.

Para ilustrar cómo los indicadores de una función pueden usarse, asumamos que un artículo habla del proceso "fogón abierto" para fabricar acero a partir del hierro.

ENTRADA	F U N C I O N	
	NOMBRE	NUMERO
Fabricación de acero	tema principal	1
hierro	materia prima	2
acero	producto	3
"Fogón abierto"	proceso o medio para llevar a cabo el objetivo primario.	

Procesamiento de Texto Completos.

1. Para recuperación de información.

   Un texto puede considerarse que va a ser procesado para propósitos de recuperación siempre que haya sido leído e indexado, o bien, que se le han aplicado métodos como clasificación, abstracción, extracción u otro tipo de análisis. En general, tal procedimiento implica que ciertos temas y puntos de vista han sido seleccionados del texto por humanos ó bien, por análisis del computador a fin de almacenar decisiones como qué parte del texto de un documento es de mayor importancia para propósitos de recuperación.

Sin embargo algunas investigaciones se han dirigido hacia el almacenamiento para propósitos de recuperación, esencialmente todo lo que está disponible en el texto completo de un documento fuente. El raciocinio de esta idea es que las necesidades potenciales son tan diversas que solamente el almacenamiento de un documento completo dará los servicios adecuados.

Para lograr el almacenamiento de textos completos para propósitos de recuperación, debemos hacer las siguientes suposiciones:

a) El texto será leído por una computadora. No es evidente todavía que el rango completo de estilos y técnicas de reproducción, ahora en uso, permitirá programación efectiva y económica para el reconocimiento de textos por medio del computador. De otra forma resultaría necesario procesar todo el texto en una forma que la computadora lo pueda leer.

b) Será económico procesar por medio de computadoras el tremendo volumen de textos completos, sin intentar "comprimir" el -- material disponible por medio de un análisis al documento.

c) Muchas preguntas se pueden analizar con tal precisión que una - selección igualmente precisa de información en el texto de un documento dará resultados útiles.

Es sabido que el principal problema al que nos enfrentaremos es la "normalización" o análisis sintáctico de un texto hecho por una computadora.

2. Traducción Automática.

En la búsqueda de conclusiones razonables para la pregunta de determinar factibilidad para el procesamiento automático de textos para propósitos de análisis, es tal vez prudente -- examinar el progreso hecho en métodos de procesamiento - automático en el campo, traducción automática.

Este es un problema muy complicado y consiste en desarrollar un - traductor automático de un idioma a otro. Obviamente esto requiere de un análisis detallado de la gramática y semántica de cada idioma, tener guardado en la memoria de la computadora un diccionario y es muy factible caer en el problema de que las traducciones sean literales. Ahora bien, generalmente las traducciones literales de un idioma a otro pueden cambiar el significado de la idea original, por lo - tanto,éstas no son confiables.

PRINCIPIOS DE BUSQUEDA

El abogado, el químico, el físico, en general todo profesionista tienen en un momento dado que enfrentar un problema común y tomar una decisión común. Para cada uno de ellos el problema de su literatura ha tomado proporciones - cruciales y cada uno busca nuevos métodos para búsqueda y correlación de la literatura de su disciplina. Cada uno siente que su problema es único y para el cual una solución única debe desarrollarse. Y es cierto que cada rama del conocimiento tiene, por varias razones, desarrollados ciertos problemas peculiares, los cuales pueden ser solucionados por medio de un diseño habilidoso de un sistema de recuperación de información.

Sin embargo, la experiencia en el desarrollo de sistemas de recuperación, - ha probado que algunos principios básicos de búsqueda son comunes a muchos temas. Existen mecanismos de búsqueda que pueden ser usados sin importar los temas de que se trate. Estos mecanismos operan basados en ciertos procedimientos de lógica común. Por lo tanto los pasos básicos que pueden ser tomados conduciendo una operación de búsqueda son:

1.- Un problema debe existir y ser reconocido, y debe ser registrado - para comunicación al sistema de búsqueda.

2.- El problema debe analizarse a fin de seleccionar guías que serán -- útiles al formular la estrategia de búsqueda.

3.- Las guías seleccionadas deben transformarse a un lenguaje y a una configuración que se adecúe a aquellas del sistema usado para análisis y almacenamiento de registros de un archivo.

4.- Las guías y las estrategias de búsqueda seleccionadas deben ser formalizadas en términos de un lenguaje y programa que se adecuará a aquellas del dispositivo usado para la búsqueda.

Puede ser útil trabajar a través de estos pasos en una búsqueda, que pudiera ser conducida en una biblioteca, usando un catálogo de tarjetas como un dispositivo de búsqueda.

1.- Un usuario de una biblioteca pudiera plantear una pregunta a un bibliotecario.

2.- Después del diálogo con el usuario, el bibliotecario determinaría -- qué conceptos de la pregunta pudieran ser guías.

3.- El bibliotecario traduciría estos conceptos:

a) Temas estandares.

b) A una estrategia de búsqueda, por ejemplo: consultar cada tema sucesivamente hasta identificar suficiente material de interés.

4.- Los temas deseados serían ordenados en una secuencia alfabética - para compararlos contra el catálogo.

5.- Se haría la consulta, seleccionando la tarjeta apropiada y comparándolo contra el catálogo de tarjetas hasta que el tema apropiado fuera localizado.

6.- Se obtendría una respuesta copiando los números encontrados de los materiales fuente identificados.

Si usaramos una computadora los pasos a seguir serían los mismos en general más no a detalle.

Existen bastantes principios los cuales pueden ser aplicados a la sistematización de procesos de búsqueda.

Primero hipoteticemos un sistema ideal de recuperación de información, consistente de 3 áreas operacionales:

1. Entrada.
2. Salida.
3. Conjunción auxiliar.

Cada una de estas áreas deben ser consideradas más a detalle.

1. Las operaciones de entrada incluyen las actividades de análisis, control de vocabulario y almacenamiento de información. En el curso de las operaciones de entrada se deben tomar decisiones en base a:
   a) Escoger aquellos términos del tema que se habilitarán para operaciones de recuperación. Por ejemplo : De los títulos de los temas, se pueden omitir artículos, preposiciones, conjunciones, etc. para formar nuestros directorios de términos para las operaciones de recuperación.
   b) Escoger la forma en la cual los términos seleccionados serán almacenados. Por ejemplo: Definir cómo serán las actualizaciones a los directorios y cómo se generarán éstos, o sea su organización, su formato, sus campos, etc.
   c) Escoger la localidad y la forma en la cual el material fuente será almacenado. Por ejemplo: Definir cómo serán los archivos que contendrán el material fuente, esto es su formato, su organización, sus campos, etc.
2. Las operaciones de salida comprenden la actividad de recuperación. Durante estas operaciones se intenta identificar guías usadas durante la entrada, las cuales relacionan a las guías de una pregunta dada.

Por ejemplo: Hecha una pregunta, a ésta se le omiten artículos, preposiciones, conjunciones, etc., con el fin de tomar algún término y poder buscarlo en el directorio, en caso de encontrarlo, la relación estaría dada.

3. La conjunción auxiliar puede ser imaginada como la función de un -- observador ideal, quien conociendo tanto la entrada como la salida - apareja tal información como si se necesitara un mantenimiento de - consistencia entre las operaciones de entrada y salida. En un sistema real de recuperación de información, la conjunción auxiliar es -- llevada a cabo por medio de una o más técnicas cada una de las cuales contribuye a la tarea del mantenimiento de consistencia.

   Las más usuales de estas operaciones son:

   a) Lista de temas: Una tabulación de temas estándares usados en el sistema de Información con el propósito de llevar a cabo la consistencia en el análisis de información y usar en el almacenamiento - los resultados de tal análisis. Por ejemplo: Emitir una lista de -- los términos usados para efectos de búsqueda en el Sistema de In-- formación, con el fin de que los nuevos documentos fuente que entren al sistema, vayan adecuados a los términos ya usados para efectos de búsqueda.

   b) Referencias cruzadas. Una notación que haga explícita una relación entre 2 ó más términos usados para designar los temas contenidos en los registros. Por ejemplo: Cuando recuperemos el término "museo", se puede dar como referencia el término "museología", en caso de existir algún documento que trate el segundo tema.

Dos formas de referencia cruzadas son:

I. "Vea referencia", para términos que son sinónimos para un sistema particular de recuperación de información.

II. "Vea referencia" para términos que están sumamente relacionados.

c) Entrada duplicada. Uno de varios temas similares los cuales referencían un documento puede ser almacenado redundantemente. El propósito de esta redundancia es impedir incertidumbre al analista, permitiéndole tener referencias en más de un lugar, de tal manera que la búsqueda puede acercarse a cualquiera de las entradas y por lo tanto lograr el material deseado.

d) Tesauro: Un libro de palabras que muestra explícitamente la relación entre las palabras que contiene, las relaciones pueden ser:

- Sinonimia
- Específica a genérica (a menudo llamado término amplio).
- Genérica a específica (a menudo llamado término angosto).
- Relación general no especificada (a menudo llamado término relacionado).

Una relación genérica a específica sería: A partir de los términos usados para recuperación de información hacer un directorio, el cual contenga para cada tema un encabezado, e inmediatamente a continuación de éste los diferentes subtemas que contiene. Por ejemplo:

Biología
Zoología
Botánica.

II ALMACENAMIENTO DE UNA PREGUNTA.

Analicemos el primer punto básico para llevar al cabo una búsqueda que es el reconocimiento de una pregunta y su almacenamiento como base para la búsqueda. Al formular una pregunta es necesario verbalizar o registrar el tema del que trate, para comunicárselo al operador del sistema de información. No es difícil llevar a cabo esta tarea si el cuestionador está familiarizado con:

1. Los temas contenidos en un archivo sobre el cual efectuaremos las búsquedas.
2. La política para el análisis del contenido de los registros.
3. Las herramientas de control de vocabulario usadas.

Sin embargo, debe recordarse que la persona que formula una pregunta puede nunca haber visto o leído alguno de los registros que desea localizar. Por lo tanto es necesario para el cuestionador:

1. Predecir la forma en la cual los autores, posiblemente desconocidos para él, han escrito acerca de los conceptos, ideas o temas que son de interés.
2. Predecir la forma para la cual, el personal del Centro de Investigación, posiblemente desconocido para él, ha analizado esos registros.

Dado que es bastante difícil para el cuestionador, hacer estas predicciones con precisión, él debe formular su pregunta al suficiente nivel de generalidad ó con la suficiente variación de expresiones, de tal forma que se tenga una seguridad de poder localizar el mayor número de registros que contienen información referente a su pregunta. No obstante el problema antes planteado, al desarrollar sistemas de ésta índole, se incluyen manuales para el

usuario, en los cuales se les explica la forma de cómo deben hacer sus preguntas, a fin de que el sistema pueda localizar el mayor número de registros referentes al tema deseado. Así mismo, también existen personas que orientan a los usuarios de cómo hacer sus preguntas.

Ejemplo:

Pregunta Genérica:

Estoy interesado en libros sobre la exploración del espacio, cuando realmente estoy interesado en el viaje del Apolo II.

Pregunta con variación de expresiones:

Deseo saber acerca de los materiales de joyería de metales, accesorios para hombre y metales preciosos, cuando realmente estoy interesado en pulseras de oro.

III SELECCION DE GUIAS A PARTIR DE UNA PREGUNTA.

Dado que estamos discutiendo acerca de recuperación mecanizada de información, esperamos que estos sistemas sean diseñados para dar un acceso multidimensional a la recuperación de registros. Tal diseño implica que los registros fuente serán caracterizados desde más de un punto de vista. Por lo tanto será conveniente recuperar los registros de interés especificando en la pregunta más de un aspecto del tema.

En sistemas tradicionales, que pueden ser monodimensionales, por naturaleza, la búsqueda generalmente es direccionada a un tema único. Cuando un indexamiento alfabético tradicional se consulta, es conveniente referirse a una entrada principal única, y leer las referencias identificadas por ella.

Por ejemplo:

Errores de puntería en armamentos antitanque; sería necesario referirse a una de las dos entradas dadas en el indexamiento alfabético:

Artillería, ejecución humana

ó

Armas antitanque, artillería

Entonces si se desea la otra entrada se consultaría.

En sistemas no convencionales que permiten búsquedas multidimensionales sería posible especificar:

Artillería, ejecución humana

y

Armas antitanque

Ambas estarían disponibles en un documento fuente dado, a fin de que el documento sea identificado.

Si se usa un acceso indexado, entonces corresponde al cuestionador dar guías en términos de ideas que pudieron haber sido indexadas. Si se usa un acceso clasificado, entonces el cuestionador, debe seleccionar la categoría o clase que cree que ha sido usada para caracterizar información que le interesa a él. Si se almacena un texto en lenguaje natural de tal forma que sea recuperable, entonces es necesario para el cuestionador predecir las formas mediante las cuales las guías pueden expresarse en lenguaje natural, a pesar del hecho de que no pudo haber visto el texto antes.

El análisis de preguntas para identificar guías como puntos de referencia en búsquedas, es análogo al análisis de material fuente. Por ejemplo:

Existen documentos que discuten el uso de fibras naturales tales como Abacá, y fibras sintéticas tales como dacrón en textiles.

Puede haber algunas guías identificadas como: fibras naturales, abacá, fibras sintéticas, dacrón y textiles.

En algunos tipos de tesauros se usa un código para los sistemas mecanizados de recuperación de información, el término abacá es listado con un código de la siguiente manera:

Término	Código
ABACA	FABR-TUTL-001

El código FABR significa miembro de la clase tejido y el código TUTL significa "usado en textiles". Cuando se consulta cada uno de estos códigos en el diccionario de códigos, un conjunto de términos, todos con el mismo código son listados.

Entonces, bajo el código FABR tendremos los siguientes términos:

FIBRA

FIBROSO

LINO

HENEQUEN

CAÑAMO

FIBRA DE VIDRIO

ABACA

SEDA ARTIFICIAL

Estos términos pueden ser considerados disponibles, adicionales a la búsqueda. Similarmente bajo el código TUTL, encontramos los siguientes términos:

TEXTILES

ALGODON ABSORBENTE

ABACA

Los cuales están disponibles para consideraciones en una manera similar. Cuando un sistema de clasificación es la base de un sistema de recuperación, un registro puede servir como una entrada al sistema. Consideramos la siguiente pregunta:

¿Hay algún material sobre turbinas hidráulicas?. La guía seleccionada como un punto de referencia para la búsqueda puede ser turbinas hidráulicas. Pero si la pregunta no considera la organización del sistema de clasificación de los documentos fuente en los cuales la respuesta apropiada puede ser localizada, es posible que no se pueda localizar la cabeza apropiada en el sistema:

Porción del Sistema de Clasificación.

62	Ingeniería
621	Ingeniería mecánica
621.2	Poder hidráulico, máquinas hidráulicas.
621.24	Turbinas hidráulicas.

En un indexamiento alfabético sería de la siguiente manera:

Porción del indexamiento del sistema de clasificación:

Ingeniería, clase 6 2

mecánica, clase 6 2 1

Máquinas hidráulicas, clase 6 2 1 . 2

Poder hidráulico, clase 6 2 1 . 2

Turbinas hidráulicas, clase 6 2 1 . 2 4

IV ORGANIZACION DE GUIAS DE BUSQUEDA

A. INTRODUCCION.

Cuando una pregunta se ha planteado y analizado, y las guías que serán útiles en la búsqueda han sido seleccionadas, es necesario organizar esas guías en una oración del "lenguaje de consulta del sistema", o sea esto significa usar la sintáxis que el sistema mecanizado de información pide para poder efectuar la búsqueda de esas guías, esto se hace usando los manuales del sistema, o bien mediante una persona que ha sido adiestrada previamente para efectuar esa labor.

Es importante recalcar que en las operaciones de sistemas de recuperación mecanizados, existen tres parámetros que pueden ser llevadas a cabo sobre sistemas convencionales de búsqueda:

(a) Encabezados de temas convencionales.

(b) Características de los temas escogidos para el sistema computarizado.

(c) Características referentes a la pregunta.

Ejemplo:

Pregunta: Selección de documentos referentes a la herrería de magnesio.

Título del documento: Desarrollo de lubricantes para herrería de materiales ferrosos y no ferrosos.

Considerando a la operación de herrería como la operación de formación.

En base al ejemplo planteado y utilizando nuestros tres parámetros obtenemos lo siguiente:

(a)	(b)	(c)
Extrusión de metales	lubricantes herrería	
Desarrollo de lubricantes	material ferroso	herrería
	material no ferroso aluminio titanio	
	magnesio	magnesio
	acero	

CONSIDERACIONES.-

1. Las búsquedas pueden más convenientemente estar basadas en la coordinación de varios aspectos o guías de los registros fuente. Este tipo de búsqueda puede ser ejecutada más efectivamente con sistemas en computadora dada la conveniencia de las guías de registros que son consideradas importantes.
2. Es generalmente económico almacenar un número grande de términos (guías) de los registros, dado el bajo costo para tal operación. Obviamente, esto habilita para búsqueda y correlación a un número mayor de términos de los registros.
3. Es conveniente efectuar las búsquedas más eficientemente en los casos donde una prescripción de búsqueda (especificación de guías) no es idéntica a las guías resultantes del análisis de los registros, dado que las guías de la prescripción de búsqueda y las almacenadas en el archivo no necesariamente son iguales. En los sistemas tradicionales de búsqueda, es conveniente requerir que los medios de identificación de registros no necesariamente se apliquen a las reglas de la búsqueda. En sistemas mecanizados es posible seleccionar registros aplicando características parciales de lo requerido a las prescripciones de búsqueda cuando:

a) Hay una sinonimia parcial entre características.

b) Hay un traslape parcial en la escala de general a específico entre características.

c) Hay un traslape parcial entre características buscadas y disponibles.

d) Hay información disponible en dos o más registros los cuales tomados juntos satisfacerán una prescripción de búsqueda.

Ejemplo:

Consideremos el ejemplo dado en la página 49 . El título del documento es:

Desarrollo de lubricantes para herrería de materiales ferrosos y no ferrosos.

Asumamos que el analista que incluyó el documento en el sistema de recuperación de información seleccionó las siguientes entradas:

Lubricantes

Herrería

Materiales ferrosos

Materiales no ferrosos

Aluminio

Titanio

Magnesio

Acero.

Podemos considerar que la operación de herrería es realmente la

operación de formación. Un sistema que haga explícita esta relación entre dos términos permitirá la identificación de uno de estos en un documento. Este es un ejemplo de sinonimia parcial entre características, siendo explícitas para propósitos de búsqueda.

Ahora consideremos el siguiente ejemplo: seleccionar todos los documentos que traten de la formación de materiales ligeros y vanadio.

Las guías o palabras llave identificadas como puntos de referencia para la búsqueda son:

formación

materiales ligeros

vanadio.

Si preguntamos por materiales ligeros, se puede satisfacer la pregunta mediante un documento que trate acerca de magnesio (que puede ser caracterizado como un metal ligero), si el sistema de recuperación hiciera explicita la relación general a específico entre los dos términos. Este es un ejemplo de traslape parcial en la escala general a específico entre características.

B. TRANSFORMACION DE GUIAS EN LENGUAJE DE UN SISTEMA DE BUSQUEDA.

Las guías seleccionadas para una pregunta son usualmente palabras, términos, frases y relaciones entre ellos. Usualmente están expresadas en len-

guaje natural, y su transformación a un lenguaje para búsqueda depende de las peculiaridades del sistema, y del código usado si existe.

Algunas formas de llevar al cabo esta transformación son:

1. Palabras índices. En palabras índices las guías seleccionadas de registros durante el análisis son palabras y por lo tanto palabras deben ser escogidas como guías de preguntas que servirán para - posteriormente encontrar palabras semejantes. Ayuda para la elección de palabras guías pueda ser obtenida usando un catálogo que -- pueda sugerir palabras como guías de búsqueda que pueden no aparecer en la búsqueda. Algunos tipos de palabras índices son:

a) Concordancias. Dado que una concordancia es un índice alfabético de palabras en un libro, cualquier pregunta dirigida a este tipo de índice debe tener sus guías hechas de palabras.

Consideramos como ejemplo una búsqueda de un texto completo para todas las palabras relacionadas con:

ANUNCIO

Para asegurar que una búsqueda exhaustiva será ejecutada sería apropiado listar otras formas del término:

ANUNCIAR	ANUNCIO
ANUNCIANDO	ANUNCIOS
ANUNCIADOR	ANUNCIABLES
ANUNCIADORES.	

Si tal búsqueda fuera ejecutada por una computadora, podría ser conveniente conducir la búsqueda para todas las palabras que comienza con los caracteres.

ANUNCI

Sin interesarnos cuales caracteres finalizarían la palabra. En otras palabras todos los caracteres seguidos de la "I" son arbitrariamente truncados, a esto se le llama truncación de la derecha.

Si nosotros truncaramos dejando las letras "AN" tendríamos un número tremendo de palabras inapropiadas, por lo tanto debemos reconsiderar si la búsqueda truncada sería útil y si no pudiera haber sido mejor especificar que la búsqueda considerara cada una de las palabras específicas de interés como puntos de referencia independientes de búsqueda. La truncación puede transformarse aún más interesante cuando consideremos la truncación de izquierda o central.

b) Palabras clave en contexto. Esta es basada en la permutación cíclica de palabras en las cuales cada sustantivo es puesto en una posición predeterminada. Dado que las palabras clave en contexto están generalmente basadas en títulos de autores, el uso de palabras es difícil de predecir. Con un indexamiento de este tipo el análisis de preguntas producirá guías de búsqueda las cuales deben resultar a partir de la selección de palabras. La discusión acerca de concordancias es por lo tanto apropiada aquí.

c) Indice de un sólo término. Este sistema es otro tipo de palabras índices el cual implica la selección de palabras llave a partir de los registros. Cuando una pregunta se analiza, estas palabras -

llave deben ser predecidas para dar guías de búsqueda. A menos que un tipo de control se haga durante el análisis de los materiales fuente, para asegurar que las formas variantes de una palabra dada no son usadas, la discusión acerca de concordancias sería apropiada aquí.

d) Sistemas basados en lenguaje natural. En este tipo de sistemas, el texto total ha sido almacenado en una máquina. Si una búsqueda es dirigida a guías basadas en palabras, todas las condiciones concernientes a las palabras índice se vuelven verdaderas aquí

2. INDICES CONTROLADOS Y CLASIFICACIONES.

a) Sin código. En índices controlados las guías seleccionadas a partir de registros durante el análisis se almacenan como palabras (o términos), que son escogidas para regularizar la manera en la cual se expresan - los resultados del análisis. Si esta regularización es llevada a cabo por medio del uso de listas de temas o sistemas de clasificación, las guías seleccionadas de una pregunta deben ser expresadas en términos de lenguaje aceptado de clasificación.

Mediante la siguiente tabla, veamos el comportamiento de una pregunta donde sus guías ya fueron seleccionadas.

Guías de las preguntas.	Porción de la lista de temas	Porción de la clasificación
(a) Indices clasificados ——→	Indices clasificados Ver aspectos de un sistema. ↓	2 Sistema de búsqueda mecanizados.
(b) Aspectos de un sistema. ——→	Aspectos de un sistema. - definición - relaciones - organizacion de almacenamiento.	21 Documentación de sistemas. 22 Aspectos de un sistema.

Explicación de la tabla:
En los indexamientos controlados o clasificaciones, las guías deben ser expresadas en una terminología cuidadosamente regularizada, como la dada en una lista de temas en la clasificación. La guía (a) no se encontró en la lista de temas como un encabezado aceptado, en lugar de eso se dió una referencia cruzada para suplir el encabezado. La guía (b) fué encontrada en la lista de temas tal como se expresó en la pregunta.
Ambas guías (a) y (b) se refieren a porciones específicas de un sistema de clasificación, determinado por los índices de referencia de una clasificación, ó a partir de saber el alcance de cada encabezado.

b) Con código. Si se usa un código para expresar los términos -- aceptados en la lista o sistema de clasificación, las guías deben expresarse en términos de los mismos códigos.

Es necesario almacenar la relación entre guías y códigos en un diccionario en el cual las guías son representadas en algún arreglo ordenado por ejemplo alfabéticamente o en secuencia numérica.

Cuando se usan códigos para expresar encabezados en una lista de temas, los mismos códigos se asignan a los sinónimos.

Por ejemplo: Sería interesante localizar información sobre la guía dinamita, entonces debemos consultar un diccionario que -- nos dé la relación palabra-código. Un seguimiento de este diccionario sería:

Español	Código
Dinámico	C - 8
Dinamita	G - 1

Dinamita isómera	P - 93
Dinastía	E - 5
Dioptría	S - 3
Diócesis	A - 42

Refiriéndonos al diccionario sobre el seguimiento de la letra "G"

Código	Español
G - 1	Dinamita
G - 2	Nitroglicerina
G - 3	Pólvora negra
G - 4	Pólvora que no produce humo.
G - 5	R D X.

En este ejemplo la guía dinamita nos puede dar una lista de todos los explosivos que tenemos en el archivo.

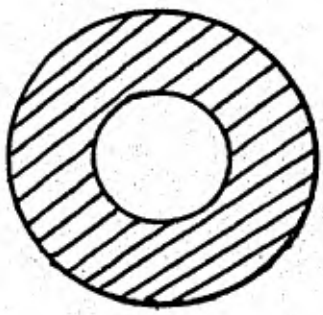
El diccionario es particularmente útil cuando términos nuevos van a ser añadidos al sistema, entonces si se encontrara un nuevo término y su asignación de código fuera requerida, hacer referencia a la sección correspondiente del código permitiría determinar si existen sinónimos del nuevo término o no los hay; en caso de que los hubiera se le asignaría al mismo código del sinónimo. Si una guía no es localizada en el diccionario, entonces las siguientes alternativas son posibles.

1. Localizar otra guía que sea sinónimo.
2. Asumir que la idea particular no aparece en el diccionario y por lo tanto no ha sido encontrada en el archivo donde se buscaría.
3. Asumir que un error se ha cometido en el almacenamiento de la guía en el diccionario.

C. ESTRATEGIAS DE BUSQUEDA

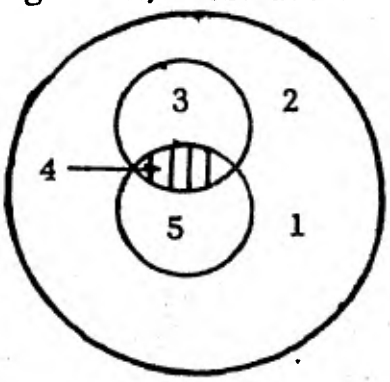
Si una de las guías de una pregunta ha sido transformada en lenguaje ó código de un sistema particular de búsqueda, entonces es necesario seleccionar - una estrategia de búsqueda, la cual explotaría mejor el contenido de un archivo particular de registros en respuesta a una pregunta.

Antes de considerar formalmente los tipos específicos de estrategias de - búsqueda, sería bueno examinar qué es lo que intentaremos llevar a cabo. Consideremos el diagrama:



donde el círculo de afuera representa todos los materiales fuente incluidos en un sistema de recuperación de información, y el círculo de adentro representa aquella información fuente que se desea recuperar, entonces, la ejecución de un sistema perfecto requerirá que toda la información que se recupere (parte no sombreada) sea de interés, y que toda la información no recuperada (parte sombreada) no sea de interés

Consideremos el caso general, mostrado en el siguiente diagrama:



donde el círculo de afuera representa el total del material fuente del sistema.

Círculo 1. Representa la información que fué recuperada.

Círculo 2. Representa la información que es de interés.

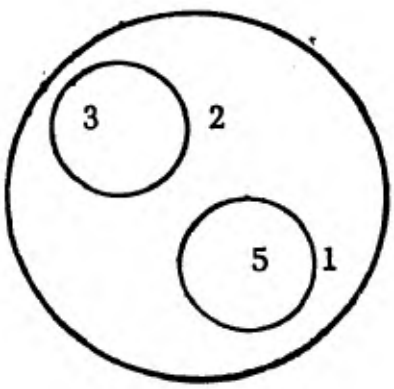
Área 3. Representa la información de interés no identificada por el sistema de recuperación.

Area 4. (Sombreada). Representa la información recuperada por el sistema, que es de interés.

Area 5. Representa la información recuperada que no es de interés.

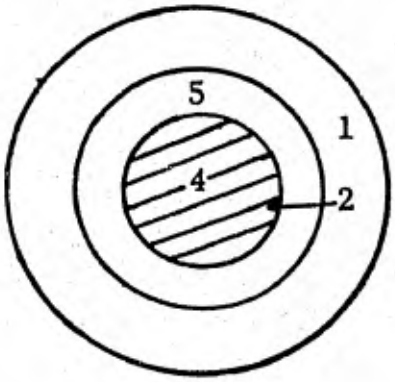
El truco es regular el sistema de tal forma que las áreas 3 y 5 sean lo - más pequeñas posibles. Si ambas no se pueden minimizar simultáneamente, la naturaleza de la búsqueda sugerirá cuál criterio tendrá prioridad.

Veamos algunos casos particulares. Consideremos el siguiente diagrama.



Como en el caso general planteado anteriormente, el círculo 1 representa la información que fué recuperada, el círculo 2 representa la información que es de interés, el área 3 representa la información de interés no identificada y el área 5 representa la información recuperada que no es de interés. En este ejemplo no existe el área 4 (sombreada) que representa - la información recuperada que es de interés. Este es el peor caso de un sistema de recuperación de información, donde la información de interés si existía, pero la búsqueda no la localizó.

Veamos otro ejemplo:



Tomando la misma notación que en el caso general, vemos que toda la información de interés fue recuperada, por lo tanto en este ejemplo el área 3 no existe, pero también fue recuperada información que no interesaba. Este tipo de resultado puede dar al usuario la confianza de que nada falta, pero si la cantidad de material recuperado es muy grande y debe examinarse para descubrir sólo algo de interés, entonces el resultado no será satisfactorio.

Una mejor forma de regular el sistema de información es una elección propia de la estrategia de búsqueda.

En la siguiente discusión asumiremos que las letras del alfabeto A, B, C... y Z, son guías seleccionadas de una pregunta, buscadas en un archivo.

Para propósitos de esta parte de la discusión, no importará si las guías son expresadas como palabras o códigos.

1. Búsqueda de aspecto único (o guía única). En este caso nosotros estamos interesados en identificar todos los registros en un archivo que tienen una guía única en común, por ejemplo todos los registros que tienen la guía A. Como ejemplo hagamos la siguiente pregunta:

   Deseo tener todos los registros del archivo que tratan de alunizaje.

   En esta pregunta la única guía es el tema alunizajes (A).

   Las búsquedas de aspectos únicos son características de sistemas convencionales indexados. Dado que estos registros son generalmente listados en órden alfabético, el acceso es logrado especificando un tema ó autor - único y buscándolo en orden alfabético.

2. Estrategia de la suma lógica. En este caso estamos especificando dos ó más guías como puntos de referencia en una búsqueda y estamos interesados en identificar todos los registros en un archivo que están caracterizados por alguna de esas guías.

   La estrategia de búsqueda puede ser representada simbólicamente como:

   $$A + B + C + D... + Z$$

   lo cual significa que el registro es deseado se tiene guía A ó guía B ó -- guía C o guía D.... ò guía Z.

   Un ejemplo sería preguntar:

   Deseo tener todos los registros del archivo cuyos autores son: John R. - Meredith o John S. Meredith o John R. Merredith ó John S. Merredith.

   En esta pregunta las cuatro guías son diferentes alternativas del nombre del autor.

John R. Meredith (A)

John S. Meredith (B)

John R. Merredith (C)

John S. Merredith (D)

La estrategia de búsqueda sería representada simbólicamente como:

A + B + C + D.

Hay otras formas de uso para este tipo de estrategia, por ejemplo; podemos dar como guías varias palabras que son sinónimos.

Algunas veces la suma lógica es usada cuando el sistema de recuperación nos da aspectos genéricos como puntos de referencia para la búsqueda.

Por ejemplo:

Podemos estar interesados en buscar la guía "perro", pero cómo haremos la pregunta. Habrá personas que generalicen en "animales," o bien otras - que directamente especifiquen en "perro".

Si solamente se usa la guía "perro", entonces la persona interesada en la guía "animales" requerirá enumerar los animales específicos que le sean de interés.

La expresión resultante será: Dame todos los documentos sobre perros o gatos o caballos o etc.

La expresión simbólica de la estrategia de suma lógica será:

A + B + C + ...

3. Estrategia de producto lógico. En este caso estamos interesados en identificar todos los registros que tienen dos o más guías en común: por ejemplo: todos los registros que tienen las guías A y B.

La estrategia de búsqueda puede ser identificada simbólicamente como:

A X B

lo cual represente que el registro deseado es sólamente si y sólo si presenta ambas A y B.

Un ejemplo de esta estrategia sería:

Deseo todos los registros del archivo que tratan de alunizajes por los - americanos en 1970. En este caso las guías son:

Alunizajes	(A)
Americanos	(B)
1970	(C)

La estrategia requiere que un sólo registro contenga las tres guías antes de ser considerado de interés. La representación simbólica de la estrategia de búsqueda es:

A X B X C

En los sistemas convencionales indexados la estrategia de búsqueda de -- guía única a menudo nos da una respuesta muy grande, conteniendo muchas referencias que son únicamente periféricas a las de interés. A fin de restringir las respuestas se efectúa una búsqueda visual la cual permite una limitación efectiva de referencias identificadas.

4. Estrategia de producto lógico de sumas lógicas. En este caso deseamos identificar aquellos registros en un archivo que tienen uno ó unas de varios conjuntos de guías en común. Por ejemplo: todos los registros que tienen las guías A ó B y C ó D. Esta estrategia de búsqueda puede representarse como:

(A + B) X (C + D).

Lo cual significa que un registro es deseado si alguna de la siguiente combinación de guías se presenta:

1. A y C
2. A y D
3. B y C
4. B y D
5. A, B y C
6. A, B y D
7. B, C y D
8. A, C y D
9. A, B, C, y D

Un ejemplo sería esta pregunta:

Deseo tener todos los registros de un archivo cuyo autor es John R. Meredith (A) ó John S. Meredith (B) y que fue publicado en 1960 (C) o 1961 (D).

(A+B) X (C+D).

En esta pregunta sólo las combinaciones 1 a 4 tendrían sentido. Sin embargo es concebible que las combinaciones 5 a 9 puedan representar criterios válidos para búsquedas exitosas si algunas de las siguientes dos condiciones se cumple:

1. Al menos un registro en el archivo tiene dos autores llamados John R. Meredith y John S. Meredith respectivamente.
2. Al menos un registro representa dos volúmenes: el primero publicado en 1960 y el segundo publicado en 1961.

Otro ejemplo de este tipo de estrategia sería:

Deseo tener todas las pinturas de un archivo que son óleos (A) o acuarelas (B) y que representan animales (C) tales como caballos (D) o vacas (E) y montañas (F) o árboles (G).

La representación simbólica de esta estrategia de busqueda es

(A+B) X (C + D + E) X (F + G).

Aquí, una pintura sería considerada pertinente a la pregunta si alguna de las siguientes combinaciones de guías se presentan:

A y C y F

A y D y F

A y E y F

A y C y G

.

.

.

A y B y C y D y E y F y G

Por supuesto que la interpretación sugerida por la representación simbólica no es la única que puede ser válida para la pregunta. Otra interpretación podría ser:

A + B X (C + D + E) X (F + G)

La estrategia del producto lógico de sumas lógicas es posiblemente la más útil visto desde un punto de vista práctico. El poder de resolución de esta estrategia está basado en su rápida habilidad de identificar material de interés requiriendo que dos ó más aspectos se presenten en el material fuente. Sin embargo, la estrategia es efectiva solamente si hay:

a. Certeza de que los aspectos de interés han sido usados con consistencia real por el analista de materiales fuente.

b. No hay incertidumbre en lo que concierne al buscador de información viendo dos aspectos que realmente son de interés.

Por supuesto en la práctica hay muy a menudo incertidumbre sobre ambos puntos.

5. Estrategia de diferencia lógica. Aquí nosotros deseamos identificar registros en el archivo que están caracterizados por la presencia de una o más guías y la ausencia de una o más guías por ejemplo: todos los registros que tienen la guía A pero la guía B está ausente. Esta estrategia de búsqueda se representa como:

A - B

Un ejemplo de pregunta que ilustrará este tipo de estrategia es,
Dame todos los registros sobre el uso de aspirinas (A) exceptuando sobre dosis (B).

Aunque esta estrategia puede ser útil, es también bastante peligroso usarla en muchas situaciones.

Supongamos que en el ejemplo, alguno de los materiales fuentes incluidos en el sistema tiene información sobre dosis normales de aspirina y sobredosis de fenobarbital. También supongamos que en la entrada para: aspirina, dosis normal, fenobarbital y sobredosis, fue sin indicación de cómo las dosis normales y sobredosis estaban ligadas a aspsrina y fenobarbital respectivamente. La estrategia de búsqueda de la diferencia lógica,

A (aspirina) - B (sobredosis).

causará que el documento discutido no sea seleccionado como un documento apropiado aunque si lo es.

6. Estrategia de secuencia. Nuevamente deseamos identificar dos ó más guías pero la guía debe ser encontrada en el registro en una secuencia exacta, por ejemplo: las guías A y B deben estar presentes, pero A debe siempre prece-

der a B. Esto puede ser representado como:

A X B

Un ejemplo de esta estrategia de búsqueda en un texto es buscar las oraciones que contienen las guías persianas y venecianas en este órden, cualquier oración que contenga estas guías pero en el órden opuesto tendrá un significado diferente y por lo tanto no se aceptará como referencia.

7. Estrategia de búsquedas entre vallas. En este caso deseamos identificar dos ó más guías en un registro, pero las guías deben encontrarse dentro de una subunidad específica del registro. Esto puede ser representado - como:

Valla (p. e. punto) X (A X B) X Valla (p. e. punto)

Un ejemplo de esta estrategia es:

Seleccionar cualquier registro que de información de un libro cuyo autor sea John R. Meredith (A) y que haya sido publicado en 1960 (B). Esta información debe aparecer en una única oración del registro.

La razón por la cual se debe hacer una especificación es que el registro -- puede tener información sobre un libro de John R. Meredith publicado en - 1961 y sobre otro libro de John S. Meredith publicado en 1960. Si la especificación de la valla no hubiera sido hecha, el registro hubiera sido seleccionado basado en ambas guías, y mandar la identificación de un registro que no es de interés como respuesta a la pregunta.

La estrategia de búsqueda entre vallas puede ser útil para explotar los encadenamientos que puedan haber sido hechos durante la operación de análisis.

8. Estrategia de mayor que y menor que. En este caso deseamos identificar registros que contienen, usualmente datos numéricos, los cuales están entre ciertos límites específicos.

   Todos los registros con información publicada entre 1960 y 1970.

   La estrategia de búsqueda sería para años de publicación mayores que -- 1959 pero menores que 1971.

   > 1959

   < 1971

9. Otras estrategias. Cuando se usa una estrategia de suma lógica sola, o como parte de una estrategia más compleja, la pregunta puede llegar a ser como si cada una de las alternativas son de igual interés o utilidad. Por ejemplo:

   Cuando estoy interesado en platería china (A), aceptaré materiales chinos como tazas (B), tazones (C),jarrones (D), dulceros (E), o utensilios (F). Si tuviera que pesar mi interés, le daría un valor de 10 a la platería (A) dado que es lo que más me interesa, y le daría un valor de 1 a cada uno de los otros elementos(B a F) . Sería necesario dar elementos de peso a la estrategia.

   Al algoritmo de peso podría ser calculado de una manera sofisticada si los pesos hubieran sido asignados durante el análisis de los documentos fuente, como una indicación de importancia del tema en un documento dado. Algunos esquemas de peso no eliminan documentos de una respuesta a una pregunta, pero más o menos los ordenan en base al valor de peso.

EL PROBLEMA Y LA SOLUCION

ANALISIS Y DISEÑO DEL PROGRAMA

a) Antecedentes.

El INFONAVIT es una empresa descentralizada cuya función básica es manejar un fondo de ahorro de los trabajadores para la construcción de vivienda. Esto se lleva a cabo de la siguiente manera:

Todas las empresas que se encuentran bajo el régimen INFONAVIT, esto es empresas que no pertenecen al gobierno y que no tienen prestación de vivienda para sus trabajadores superior al 5% de su salario, están obligados a aportar bimestralmente esta cantidad con el fin de ir generando dicho fondo de ahorro, y será a partir de éste -- que se irán construyendo las viviendas y estas serán otorgadas a los trabajadores mediante un crédito, el cual el trabajador pagará por - medio de la empresa.

Vemos entonces que para el INFONAVIT es indispensable un catálogo básico de empresas, con el fin de tener:

1. Los datos generales de identificación de la empresa.
2. El pago bimestral actualizado que las empresas hacen en favor de cada uno de sus trabajadores.
3. El monto bimestral de los recargos sobre pagos retardados.
4. La cuenta de abonos por concepto de créditos pagados.
5. El monto anual de aportaciones que la empresa hizo en favor de sus trabajadores en base a su declaración anual.
6. El monto anual de aportaciones declarado por la empresa.

7. El número de trabajadores que la empresa ha tenido por año.

Por lo tanto, este archivo contendrá dos grandes bloques de información.

1. Identificación de la empresa.

Serán los datos generales de la misma:

Registro Federal de Causantes, determinante de sucursal, - número de expediente INFONAVIT (número único que da el INFONAVIT a cada empresa), nombre, denominación ó ra-- zón social de la empresa, domicilio, colonia, población, municipio, entidad federativa, número de registro del IMSS, giro, número de sucursales, central obrera, etc.

2. Información contable de la empresa.

Pagos bimestrales de 1972 a la fecha.

Recargos bimestrales de 1972 a la fecha.

Abonos bimestrales de 1972 a la fecha.

Abonos anuales de 1972 a la fecha

Aportación anual de trabajadores de 1972 a la fecha

Número de trabajadores anual de 1972 a la fecha

b) Estructura.

El archivo tiene la siguiente estructura:

El número de expediente INFONAVIT se forma de la siguiente manera:

Clave de estado	2 dígitos
Número secuencial por estado	6 dígitos
Verificador	1 dígito

Para cada estado se estimó un número máximo de empresas, éste número será el número secuencial máximo por estado, por lo tanto, no todos los registros del archivo están utilizados. Además el algoritmo usado - para el cálculo del verificador da un valor entre 0 y 10, dado que dicho subcampo sólamente es de un caracter, aquellos números de expediente con caracter verificador igual a 10 no existen en el archivo.

Catálogo Básico de Empresas.

Registros de Control.

clave Edo.	Núm. secuencial				
01	000001 D	C1	C2	. . .	Cn
01	000002 D	C1	C2	. . .	Cn
.					
.					
.					
01	M1 D	C1	C2	. . .	Cn
02	000001 D	C1	C2	. . .	Cn
02	000002 D	C1	C2	. . .	Cn
.					
.					
.					
02	M2 D	C1	C2	. . .	Cn
.					
.					
.					
32	000001D	C1	C2	. . .	Cn
32	000002 D	C1	C2	. . .	Cn
.					
.					
.					
32	M32 D	C1	C2	. . .	Cn

donde M1, M2, . . . M32 es el número secuencial máximo por estado, D es el dígito de control de cada uno de los números de expediente, y C1, C2, . . ., Cn son los diferentes campos de cada registro.

Vemos que en base a esta estructura el acceso por número de expediente no resulta muy complejo, ya que para poder accesar un registro debemos saber los números secuenciales máximos por estado, así como cuántos huecos existen por estado, o sea aquellos números cuyo verificador fue igual a 10, de esta manera tendremos la dirección de la empresa deseada.

c) Uso del archivo.

Para poder tener esta información actualizada necesitamos de varios subsistemas.

La identificación general de la empresa se capta a partir del registro empresarial, que ésta llena cuando se dá de alta, o bien al captar información de ella: pagos, trabajadores, etc., por lo tanto, necesitamos de un subsistema que dé altas, bajas y cambios al catálogo.

Así mismo, la información contable va a ser captada por diferentes subsistemas cuya finalidad será validar la información que la empresa declara o paga para asi poder tener actualizado el archivo.

Al captar y validar la información contable de cada una de las empresas nos enfrentamos a un grave problema: muchas de las formas que recibimos no tienen el número de expediente de la empresa, solamente contamos con su registro federal de causantes y nombre o razón social; es posible que la empresa ya esté dada de alta y no nos haya declarado su número de expediente, ó también es posible que no esté dada de alta en el archivo y no hayamos recibido su registro empresarial.

Para resolver este problema contamos con una rutina de acceso al catálogo básico de empresas por registro federal de causantes, con esto accesaríamos el registro y de él tomaríamos el número de expediente de la empresa, sin embargo, existen dos inconvenientes: primero, el registro federal de causantes no tiene un dígito verificador, por lo tanto en caso de que se digitara mal no accesaríamos el registro deseado o bien accesaríamos otro con lo cual estaríamos incurriendo en un grave error, ya que la información de una empresa la estaríamos transfiriendo a otra: segundo, puede ser que el registro sea correctamente digitado, pero nos enfrentaríamos al problema de que puede ser múltiple, o sea que existan varias empresas con el mismo registro federal de causantes, en cuyo caso no podríamos saber cual es su número de expediente. Por otra parte, obteniendo estadísticas del archivo, sabemos que 28,000 registros federales de causantes son múltiples, o sea existen al menos 2 iguales, por lo que al utilizar la rutina no siempre tendremos éxito.

Para resolver este problema lo que comunmente se hace es buscar por nombre de la empresa en un listado ordenado por dicho campo, con el fin de obtener su número de expediente.

En base a lo planteado anteriormente se vió la necesidad de desarrollar un subsistema de búsqueda por nombre, además este subsistema formaría parte del consultor general del catálogo, antes del desarrollo de esta tesis sólo se podía consultar por registro federal de causantes y número de expediente INFONAVIT, así también serviría para depurar el catálogo ya --

que sabemos que en él existen empresas que estan dadas de alta más de una vez y de esta manera sabremos cuántas y cuáles son.

Por otra parte esta rutina también se podrá utilizar para efectos estadísticos, ya que por medio de una parte del nombre podremos saber su actividad; por ejemplo: dando únicamente la palabra BANCO, sabremos cuántos y cuáles bancos tenemos registrados.

Por último, podremos desplegar todos los datos de la empresa consultada, ya sean los de identificación o bien los datos contables.

d) Diseño de la consulta.

Teniendo como objetivo primario lo expresado en los párrafos anteriores, se tenía que pensar en un algoritmo que fuera óptimo en cuanto a la utilización de recursos de máquina, ya que por una parte nos enfrentamos al problema del volumen de información: El archivo tiene actualmente 400,000 registros dados de alta; y por otra parte el algoritmo tiene que ser rápido dada la gran cantidad de proceso que hay actualmente, así también el tamaño de sus tablas de acceso debe ser pequeño debido a las limitaciones de recursos en disco.

Veamos algunas ideas del acceso secuencial con índice,* es el que usaremos en nuestros algoritmos.

Al momento de crear un archivo secuencial con índice todos los registros deben ir clasificados por su llave de acceso, con el fin de ir generando dos tablas. La primera es la tabla fina que contiene la primera ó última llave de cada bloque del archivo, esta tabla nos dará el acce-

*Traducción al español de la palabra index-sequential en inglés.

so a cada uno de los registros; la segunda tabla es la tabla gruesa y -- contiene la primera ó última llave de cada bloque de la tabla fina con lo que lograremos su acceso. La estructura de ambas tablas es explicada a detalle más adelante donde se explica el algoritmo propuesto.

Este es entonces un archivo que podemos accesar de dos formas diferentes: leyéndolo secuencialmente, o bien dada una llave, accesar el registro que la contenga, esto último se hace mediante búsquedas binarias sobre la tabla gruesa, tabla fina y el mismo archivo. También para poder dar altas al archivo existe un área de sobreflujo, o sea dado que no es posible insertar un nuevo registro en el lugar que le corresponde de acuerdo al orden de la llave, el nuevo registro se grabará en esa área, marcando en el registro precedente el apuntador correspondiente, para de ésta forma poder accesarlo. En caso de dar de baja registros, éstos solamente serán marcados y en caso de contener apuntadores, reapuntados.

Formas de acceso.

Proposiciones:

1.- La primera proposición fue la siguiente: tomando el nombre de la empresa, a partir de este se generaría la parte alfabética del registro federal de causantes, y tomando las rutinas de acceso por medio de esta llave se buscaría el primer registro con esta parte alfabética y a partir de ese leyendo secuencialmente el directorio se consultaría el catálogo de empresas con el propósito de comparar el nombre propuesto -

contra el del archivo hasta saber si existe ó no el registro, para que en caso de existir desplegar su información.

Desventajas: Este algoritmo realmente no sería una búsqueda por -- nombre ya que para poder accesar el registro dependeríamos de un registro federal de causantes incompleto (sin la parte numérica) y -- que no sabemos si en realidad esa parte corresponde al RFC, por otra parte, si tomamos únicamente la parte alfabética del RFC (cuatro primeros caracteres) de los 400,000 registros federales de causantes del archivo, solamente 30,000 son únicos, ó sea no estan duplicados ó triplicados, etc.

Además, el algoritmo sería muy lento, ya que nos veríamos obligados a efectuar un rastreo secuencial sobre el directorio para después por cada registro del directorio hacer un acceso al catálogo de empresas, y esto es debido a que el catálogo no está ordenado por registro federal de causantes.

2.- La segunda proposición es similar a la primera: Igualmente a partir del nombre de la empresa, generaríamos la parte alfabética del registro federal de causantes y usando las rutinas de acceso por medio del RFC llegaríamos a un directorio el cual contiene registros federales de causantes y el nombre de la empresa, evitándonos así la lectura -- del catálogo, lo cual vimos que entorpecía mucho el algoritmo.

Desventajas: Como en la proposición número 1, realmente no hablaríamos de una búsqueda por nombre y estaríamos trabajando con un -- RFC incompleto y que posiblemente no corresponde. Por otra parte -

además del rastreo secuencial que estaríamos obligados a hacer en el directorio de RFC y nombre, nos enfrentaríamos al problema del espacio; este directorio sería muy voluminoso, veamos cuántos caracteres ocuparía en disco.

Número de registros del catálogo de empresas	400,000
Número de caracteres del registro federal de causantes	10
Número de caracteres del nombre	40
Número de caracteres de la llave del catálogo	6

Entonces el número de caracteres necesarios para generar nuestro directorio sería:

400,000 X (10 + 40 + 6) = 22,400.000

Vemos entonces que sería muy costoso tener un directorio de 22 millones de caracteres para una consulta por nombre que no sería segura.

3.- En base a las dos proposiciones anteriores vemos que para tener una verdadera consulta por nombre debemos trabajar con este campo, entonces nos vemos en la necesidad de generar un directorio para uso del mismo.

La generación de este directorio consiste en ir leyendo secuencialmente el catálogo de empresas para ordenarlo por nombre e ir generando un directorio que contenga todos los nombres del catálogo ordenados, así como su llave física para acceso al archivo. Para poder accesar el directorio de nombres también vamos generando tablas, por medio

de la tabla fina accesaremos el bloque del directorio donde se encuentra el registro deseado, y la tabla gruesa nos dará el bloque de la tabla fina.

Este algoritmo realmente ya es una búsqueda por nombre, pero nuevamente nos volvemos a enfrentar al problema del espacio.

Veamos cuántos caracteres tendría nuestro directorio

Número de registros del catálogo de empresas	400,000
Número de caracteres del nombre	40
Número de caracteres de la llave del catálogo	6

400,000 X (40+ 6) = 18,400,000

Dado que necesitamos 18 millones de caracteres para almacenar el directorio, lo cual representa más de medio paquete * (considerando los paquetes que tenemos en nuestra computadora), el algoritmo deja de ser óptimo dadas las limitaciones de espacio en disco.

Vemos que el problema que nos queda a resolver es el de espacio en disco, ya que considerando que el catálogo utiliza 1,800,000 segmentos (1 segmento es igual a 30 palabras), el directorio utilizaría - - - 80,000 segmentos que equivale al 4.4% del tamaño del archivo, por lo que podemos pensar en un directorio más pequeño, lo que nos lleva a la solución considerada como la mejor y que es el algoritmo propuesto.

* Traducción al español de la palabra 'pack' en inglés.

ALGORITMO PROPUESTO:

El algoritmo propuesto se ha adaptado para recuperar los nombres en el Catálogo Básico de Empresas, generado en el INFONAVIT, y consiste en lo siguiente:

Con respecto al texto que el programa tiene que recuperar, el análisis fué bastante simple ya que en realidad no se trata de ningún texto, sino de una cuerda que es el nombre de las empresas, la que tiene características bien definidas entre las cuales encontramos:

1. Su longitud es constante: 40 caracteres en código BCL.
2. Casi todas las palabras son significativas, por lo que únicamente fué necesario hacer una lista de exclusión, que contendría artículos, preposiciones, conjunciones y leyendas como S.A. y C.V., tal y como se plantea en el indexamiento de permutación, pero no con el fin de generar un reporte sino un directorio, por lo que no fue necesario contar con un grupo de analistas de información para seleccionar las palabras significativas; entonces esta selección no fue subjetiva.

Con el fin de resolver el problema del espacio en disco, a cada una de las palabras del nombre se les aplicó una técnica de compresión que consiste en:

El primer caracter es el pivote y éste siempre prevalece, se continúa analizando la palabra omitiendo las vocales y dejando únicamente tres consonantes más, sin que éstas se repitan, de esta manera hacemos la compresión de la palabra.

Ejemplo:

ABARROTES ABRT

Por otra parte obteniendo estadísticas del archivo, vimos que el 79.5% de los nombres son de tres palabras significativas o menos y los que tienen más de tres palabras, en muchos casos son el segundo nombre de una persona física, por lo tanto el sistema de recuperación únicamente trabajaría con las tres primeras palabras significativas, y no tomando en cuenta las de la lista de exclusión.

Al ir analizando y comprimiendo todos los nombres del archivo vamos generando un directorio empacado: sabemos que por cada nombre de empresa vamos a almacenar un máximo de 3 palabras, y por cada palabra un máximo de 4 caracteres, ahora bien, en 5 bits nosotros podemos almacenar $2^5 - 1$, o sea un valor entre 0 y 31, si asociamos un valor numérico a cada letra del alfabeto, cada caracter lo podemos almacenar en 5 bits, siendo 4 caracteres por palabra y 3 palabras por nombre, por lo tanto, necesitamos 20 bits por palabra y 60 bits por el nombre completo.

En la computadora en la que trabajamos, cada palabra tiene 48 bits, por lo tanto necesitaremos de dos palabras, en la primera almacenaremos las 2 primeras palabras del nombre, en la segunda almacenaremos la tercera palabra del nombre y en los bits restantes de la palabra guardaremos la dirección física del archivo donde se localizó dicho texto.

Una vez que hemos terminado de generar el directorio del archivo, éste lo ordenamos de manor a mayor, tomando como llave el nombre comprimido; en nuestro procedimiento de salida de la clasificación, vamos generando 2 tablas que nos daran un fácil acceso a nuestro directorio, para generar estas tablas utilizamos el siguiente criterio:

Sabemos que estamos trabajando sobre un archivo de 400,000 registros, por lo cual debemos encontrar un número múltiplo de 30, tal que nos - permita guardar para cada uno de nuestros bloques de nuestro directorio el rango de los nombres comprimidos que contiene, buscamos un número múltiplo de 30, debido a que este es en palabras el tamaño de un segmento en disco. Este primer directorio es la tabla fina. También deberemos generar la tabla gruesa que contendrá para cada bloque de nuestra tabla fina, el rango de nombres comprimidos que contiene cada bloque de dicha tabla.

Proposiciones:

Consideramos un número el cual será el número de registros por bloque de nuestro directorio de nombres comprimidos, de esta manera - tendremos 400,000 /n bloques para nuestro directorio, además este - número n deberá ser múltiplo de un segmento.

Por lo tanto nuestra tabla fina tendrá las siguientes características:

Por cada bloque del directorio contendrá 2 palabras con la siguiente - información: en la primera palabra y mitad de la segunda guardará el nombre comprimido contenido en el bloque, y la otra mitad contendrá el número físico del bloque que contiene dicha información, entonces el tamaño de nuestra tabla fina será igual al tamaño de bloques que contenga el directorio. Ahora bien, démosle a la tabla fina el mismo factor de bloque que le dimos al directorio o sea n, por lo tanto cada bloque de la tabla fina contendrá n al cuadrado posibles nombres comprimidos del directorio.

Por último, generemos la tabla gruesa de la misma manera que la ta-

bla fina, siendo el tamaño de ésta el equivalente de un bloque o sea n - registros, conteniendo ésta n al cubo, posibles nombres comprimidos del directorio, por lo que sabiendo que nuestro archivo tiene 400,000 registros, entonces para encontrar el mejor factor de bloque:

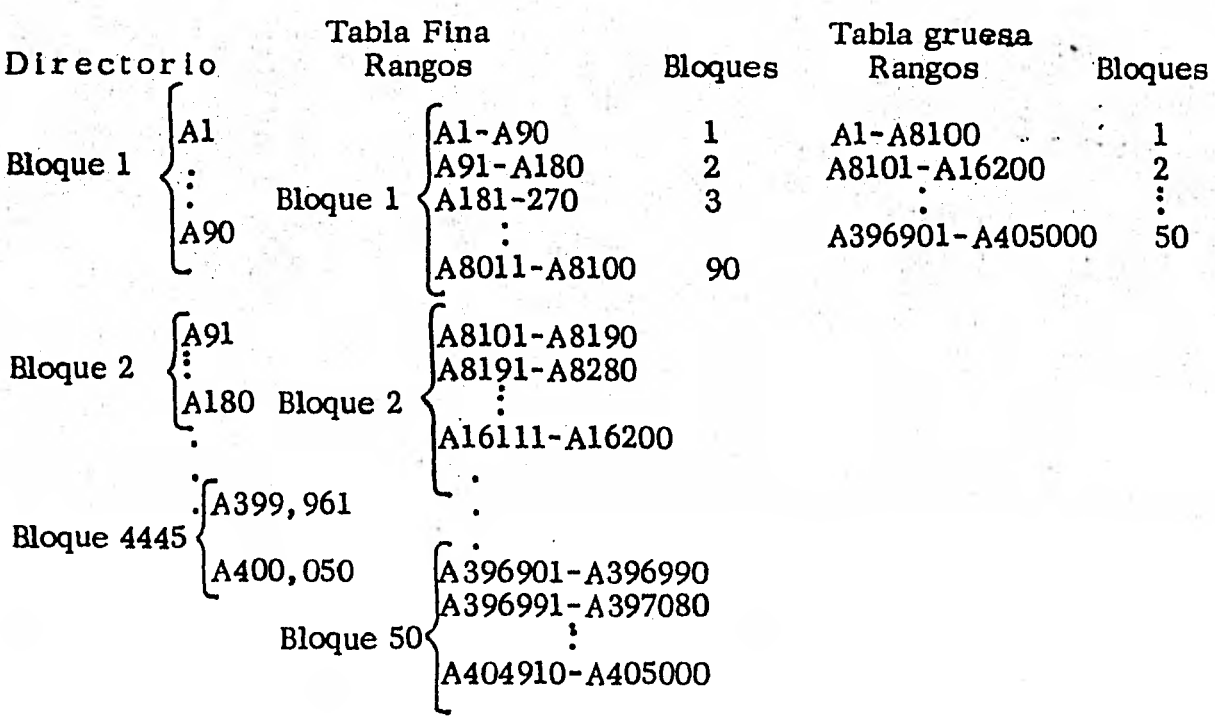
$$n = \sqrt[3]{400,000}$$

$$n = 70.58$$

Este no puede ser un factor de bloque dado que si bien nos permitiría almacenar la información del archivo en tablas, no es múltiplo de 30. Entonces el número múltiplo de 30 más próximo es 90.

$$90^3 = 729000$$

Este número además de ser múltiplo de 30, da un amplio rango de crecimiento al archivo.



De esta manera quedarán conformadas la tabla fina y la tabla gruesa. Una vez que hayamos generado nuestras tablas, el archivo quedará listo para efectuar consultas y recuperar por nombre; dado un nombre se le aplica la técnica de compresión, una vez teniendo la configuración - binaria, se efectuará una búsqueda binaria sobre la tabla gruesa para que de esta manera sepamos en qué bloque de la tabla fina se encuentra ubicado dicho nombre, nuevamente efectuando una búsqueda binaria sobre el bloque de la tabla fina sabremos bloque del directorio se encuentra el nombre deseado, al final efectuaremos la búsqueda del - nombre dado en un bloque del directorio y sabremos si éste se encuentra o bien si existe alguno parecido y así tendremos la dirección física del registro para leerlo y desplegar los datos de la empresa.

Ahora bien, analizando el tipo de recuperación que el sistema da, podemos encontrar las siguientes características:

1. Debido al algoritmo de compresión de las palabras, muchas de ellas siendo diferentes, pudieron haber tenido la misma configuración de letras al comprimirse.
2. La comparación que el sistema efectúa en la búsqueda, siempre la hace sobre los nombres comprimidos y nunca sobre los nombres originales.
3. El tipo de búsqueda que se hace sobre los directorios siempre da una respuesta, esto es de que en caso de que las cuerdas comprimidas no sean iguales, el recuperador devuelve la que más se le parece.

Entonces, basándonos en lo visto en la sección de estrategias de búsqueda (página 72), podemos concluír lo siguiente:

El recuperador no es perfecto, ya que siempre nos devolverá información, y sobre ésta habrá que seleccionar la que sea de interés, en caso de no existir información de todas maneras nos enviará lo más parecido que hay en el archivo. Sin embargo, el recuperador nos garantiza que de existir información de interés siempre nos la dará.

También el directorio nos puede dar mucha ayuda para procesos de depuración del archivo:

Rastreándolo secuencialmente podremos desplegar fácilmente empresas con el mismo nombre, o bien empresas con nombres muy similares que en un momento dado pueden ser las mismas y estar duplicadas en el archivo.

Así mismo, la generación de nuestro directorio va a ser de muchísima utilidad para efectuar cruces por medio del nombre contra catálogos de otras dependencias, como podrían ser la Secretaría de Hacienda y Crédito Público o bien el Instituto Mexicano del Seguro Social, y de esta forma las empresas existentes en esos catálogos y que nosotros no las tengamos registradas darlas de alta con el fin de enriquecer nuestro catálogo.

Por último nos enfrentamos al problema de la actualización al catálogo o sea las altas, bajas y cambios que se dan cotidianamente al archivo; esto quiere decir que después de haber generado nuestro directorio, tabla fina y tabla gruesa, y se hagan actualizaciones al catálogo nuestros archivos para acceso van a quedar desactualizados.

Por lo tanto debemos preveer un subsistema de actualización al directorio.

Consideraciones.

Tenemos 3 tipos de movimientos: altas, bajas, y cambios.
Las bajas no generan ningún problema al directorio ya que de hecho no se da la baja física al registro sino que sólamente se le pone una marca debido a que cuando una empresa cierre o se dé de baja oficialmente ante la Secretaría de Hacienda y Crédito Público, debemos conservar la información de ella, dado que podemos tener pagos registrados en favor de sus trabajadores. De esta forma al recuperar el registro, se desplegará su estado ó sea si está vigente o dado de baja.
De los cambios únicamente nos van a interesar aquellos que sean cambios de nombre, para lo cual debemos grabar en otro archivo dos transacciones para el directorio, una de baja que será el nombre antiguo y una de alta que será con el nuevo nombre. En este caso es importante hacer notar que el nombre antiguo se perderá, debido a que el cambio se produce por errores ortográficos, o bien cambio de razón social de la empresa. El formato de estas transacciones será:

CAMPO	POSICIONES	
Movimiento	1	"A" ó "B"
Nombre	40	
Dirección	6	

Por último, con respecto a las altas, generaremos una transacción de alta a nuestro directorio la cual tendrá el mismo formato que las transacciones de cambios.

Estas transacciones al directorio pasarán por un proceso de comprensión del nombre, idéntico al de la generación del directorio y clasificación - por nombre comprimido, con el fin de que una vez que tengamos las transacciones con los nombres comprimidos y clasificados, hagamos un proceso de intercalación sobre el directorio a fin de insertar los nuevos -- nombres y dar de baja a los que se les haya efectuado cambio de nombre.

Dado que estaremos modificando al directorio, al momento de ir generando el nuevo por medio de la intercalación, deberemos regenerar nuestras tablas fina y gruesa.

E J E M P L O S

CONSULTA POR EMPRESA

R.F.C

NOMBRE O

EXPEDIENTE QUE DESEE CONSULTAR

*BANCO NACIONA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
001	310045037	BNF440215	BANCO NAC DE FOMENTO COOPERATIVO SA CV	BD
002	090065999	BNA730207	BANCO NACIONAL AGROPECUARIO S A	CI
003	090066006	BNA650331	BANCO NACIONAL AGROPECUARIO. S.A.	BD
004	120052938	BNA730523	BANCO NACIONAL AGROPECUARIO SA	CC

006	090066022	BNC370608	BANCO NACIONAL DE COMERCIO EXTERIOR S A	CC
007	090066014	BNC470812	BANCO NACIONAL CINEMATOGRAFICO, S.A.	CI
008	080011063	BNC520308	BANCO NACIONAL DE CRED EJIDAL S A	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
009	080011071	BNO520308	BANCO NACIONAL DE CRED EJIDAL SA DE CV	CI
010	160084512	BNC520306	BANCO NACIONAL DE CRED RURAL SA P IND VE	AA
011	250101947	BNC750709	BANCO NACIONAL CRED RURAL SA FID CRED AR	CC
012	010002936	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL S A DE	BD
013	010039252	BNC561030	BANCO NACIONAL DE CREDITO AGRICOLA S A	BD
014	030011132	BNC561030	BANCO NACIONAL DE CREDITO AGRICOLA SA	BD
015	040001547	BCE520306	BANCO NACIONAL DE CREDITO EJIDAL	CC
016	040016900	BNC561030	BANCO NACIONAL DE CREDITO AGRICOLA SA.	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
017	040016986	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA CV	BD
018	040017796	BNC520306	BANCO NACIONAL DE CREDITO EJIDA	BD
019	040019284	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA DE C	CI
020	040023559	BNC611030	BANCO NACIONAL DE CREDITO AGRICOLA SA	AA

021	070003602	BNC560103	BANCO NACIONAL DE CREDITO AGRICOLA	CI
022	070003610	BNC730301	BANCO NACIONAL DE CREDITO AGRICOLA S A	CI
023	070003629	BNC740424	BANCO NACIONAL DE CREDITO AGRICOLA S A	A
024	070003637	ENC561030	BANCO NACIONAL DE CREDITO AGRICOLA S A	A

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
025	080011098	BNC561030	BANCO NACIONAL DE CREDITO AGRICOLA SA	BD
026	080011101	BNC520308	BANCO NACIONAL DE CREDITO EJIDAL S A	BD
027	080128971	BNC561030	BANCO NACIONAL DE CREDITO AGRICOLA SA	BD
028	090066030	BNC740401	BANCO NACIONAL DE CREDITO EJIDAL S A DE	CI
029	090066049	BNA740515	BANCO NACIONAL DE CREDITO RURAL S A FIDE	BD
030	090066057	BNC750709	BANCO NACIONAL DE CREDITO RURAL SA	CC
031	090911350	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA CV	BD
032	091110688	BNC730207	BANCO NACIONAL DE CREDITO RURAL SA	CI

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
033	100003532	BCM561030	BANCO NACIONAL DE CREDITO AGRICOLA	CI
034	100003540	BNC610630	BANCO NACIONAL DE CREDITO AGRICOLAS S.A	BD
035	100045626	BNC561030	BANCO NACIONAL DE CREDITO AGRICOLA SA	CC

037	110008677	BNC740709	BANCO NACIONAL DE CREDITO EJIDAL SA	CI
038	110094670	BNC760601	BANCO NACIONAL DE CREDITO RURAL	CC
039	120003023	BNG260304	BANCO NACIONAL DE CREDITO AGRICOLA S	A
040	120003031	BNC151030	BANCO NACIONAL DE CREDITO AGRICOLA S A	A

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
041	120003058	BNC520206	BANCO NACIONAL DE CREDITO EJIDAL	A
042	140014268	BCN520306	BANCO NACIONAL DE CREDITO EJIDAL	CC
043	140249451	BNC750709	BANCO NACIONAL DE CREDITO RURAL SA	CI
044	140252630	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA	BD
045	140277366	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA DE	BD
046	160064864	BNC750709	BANCO NACIONAL DE CREDITO RURAL S A	CC
047	160067723	BNC520306	BANCO NACIONAL DE CREDITO RURAL S A P I	BD
048	160073871	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA CV	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
049	160076072	BNC520306	BANCO NACIONAL DE CREDITO RURAL	BD
050	170002535	BNC750414	BANCO NACIONAL DE CREDITO EJIDAL	A
051	170027376	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA CV	BD
052	200002503	BNG561030	BANCO NACIONAL DE CREDITO AGRICOLA S A	CI
053	200002511	BNC250306	BANCO NACIONAL DE CREDITO EJIDAL SA DE C	CI

054	200038850	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA	A
055	210005971	BNS561030	BANCO NACIONAL DE CREDITO AGRICOLA SA	A
056	230000754	BNC740614	BANCO NACIONAL DE CREDITO EJIDAL	A

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
057	240075978	BNC730319	BANCO NACIONAL DE CREDITO RURAL SA	BD
058	250083884	BNC561030	BANCO NACIONAL DE CREDITO AGRICOLA SA	BD
059	250093219	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA CV	BD
060	250093413	BNC520306	BANCO NACIONAL CREDITO EJIDAL SA CV	BD
061	260009180	BNC760219	BANCO NACIONAL DE CREDITO AGRICOLA S A	A
062	260009199	BNC260304	BANCO NACIONAL DE CREDITO AGRICOLA S A	A
063	260148865	BNC750907	BANCO NACIONAL DE CREDITO RURAL SA	CC
064	280097271	BNC520306	BANCO NACIONAL DE CREDITO RURAL SA	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
065	280104820	BNC750709	BANCO NACIONAL DE CREDITO RURAL SA	CC
066	280105312	BNC730627	BANCO NACIONAL DE CREDITO RURAL SA	CC
067	280106882	BNC520306	BANCO NACIONAL DE CREDITO EJIDAL SA	BD
068	300011024	BNCA561030	BANCO NACIONAL DE CREDITO AGRICOLA	A

070	020007140	BNF611002	BANCO NACIONAL DE FOMENTO COOPERATIVO SA	BD
071	020120745	BNF440215	BANCO NACIONAL DE FOMENTO COOP SA DE CV	BD
072	070040966	BNF440215	BANCO NACIONAL DE FOMENTO COOPERATIVO SA	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
073	200039172	BNF440215	BANCO NACIONAL DE FOMENTO COOPERATIVO SA	BD
074	091268028	BNH240426	BANCO NACIONAL DE HUNGRIA BUDAPEST HUNGR	AA
075	230000762	BNM740711	BANCO NACIONAL MONTE DE PIEDAD	A
076	090066073	BNM040515	BANCO NACIONAL DE MEXICO	BD
077	090066081	BNM840515	BANCO NACIONAL DE MEXICO S A	CC

078	090928709	BNM840515	BANCO NACIONAL DE MEXICO SA	BD
079	120035332	BNM840515	BANCO NACIONAL DE MEXICO S A	BD
080	260130664	BNM840515	BANCO NACIONAL DE MEXICO SA	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
081	260193348	BNM840515	BANCO NACIONAL DE MEXICO SA	BD
062	270026525	BNM040515	BANCO NACIONAL DE MEXICO	BD
083	040001555	BNO740705	BANCO NACIONAL DE OBRAS	A
084	090066103	BNO670315	BANCO NACIONAL DE OBRAS Y SERVICIOS PUBL	CI

085	090901711	BNO670515	BANCO NACIONAL DE OBRAS Y SERV PUB SA	BD
086	090907434	BNO670515	BANCO NACIONAL DE OBRAS Y SERV PUBL	BD
087	120003066	BMO750818	BANCO NACIONAL DE OBRAS Y SERVICIOS PU	A
088	140193561	BNO670515	BANCO NACIONAL DE OBRAS Y S P S A	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
089	140241744	BNO710129	BANCO NACIONAL DE OBRAS Y SERV PUB SA	BD
090	140251197	BNO710129	BANCO NACIONAL DE OBRAS Y SERV PUBS SA	BD
091	040024482	BNP800125	BANCO NACIONAL PESQUERO Y PORTUARIO SA	AA
092	090066065	BNP800101	BANCO NACIONAL PESQUERO Y PORTUARIO	CI
093	260184896	BNP440215	BANCO NACIONAL PESQUERO Y PORTUARIO SA	AA
094	060001240	BNT581118	BANCO NACIONAL DE TRANSPORTES S A	A
095	010039473	BNU740926	BANCO NACIONAL URBANO SA	BD
096	020086784	BNU740926	BANCO NACIONAL URBANO SA	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
097	020088795	BNU740926	BANCO NACIONAL URBANO S A	BD
098	050113933	BNU740926	BANCO NACIONAL URBANO S A	BD
099	090780310	BNU740826	BANCO NACIONAL URBANO S A	CI

101	140171436	BNU740926	BANCO NACIONAL URBANO S A	BD
102	140251928	BNU740926	BANCO NACIONAL URBANO SA	BD
103	150087896	BNU740926	BANCO NACIONAL URBANO S A	BD
104	190159987	BNU740926	BANCO NACIONAL URDANO SA	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
105	210070536	BNU740926	BANCO NACIONAL URBANO SA	BD
106	230009905	BNU740926	BANCO NACIONAL URBANO S A	BD
107	230009956	BNU740926	BANCO NACIONAL URBANO S A	BD
108	230010482	BNU740926	BANCO NACIONAL URBANO SA	BD
109	240005120	BNU740926	BANCO NACIONAL URBANO S A	BD
110	250093286	BNU740926	BANCO NACIONAL URBANO SA	BD
111	260118990	BNU740926	BANCO NACIONAL URBANO S A	BD
112	260132500	BNU740926	BANCO NACIONAL URBANO SA	BD

110

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
113	270030212	BNU740926	BANCO NACIONAL URBANO SA	BD
114	280096429	BNU740926	BANCO NACIONAL URBANO S A	BD
115	300127537	BNU740926	BANCO NACIONAL URBANO SA	BD
116	310049806	BNU740926	BANCO NACIONAL URBANO S A	BD
117	310062578	BNU781130	BANCO NACIONAL URBANO SA	BD

118 320020797 BNU740926 BANCO NACIONAL URBANO SA BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

*ABARROTES DEL CENTRO S

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
001	200031333	ACE770122	ABARROTERA DEL CENTRO S A DE C V	CI
002	090919041	ACC770214	ABARROTES CONTRY CLUB SA	CI

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

*ORTIZ MEZA JOSE DE JESUS

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
001	240066820	OIMJ081020	ORTIZ MAZO JUAN JOS	CI

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

*INSTITUTO DEL FONDO NACIONAL PARA LA VIVIENDA DE LOS TRABAJADORES

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
001	250093685	IFN720501	INSTITUTO DEL FONDO NACIONAL DE LA VIV	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION (ENTER)

*MEXICANA DE AVIACION

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
001	091095301	MAV240820	CIA MEXICANA DE AVIACION S\	AA

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION (ENTER)

*SECRETARIA DE HACIENDA Y CREDITO PUBLICO

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
001	280076886	SAGO331117	SACRISTE GARCIA OSCAR	A

SI NO DESEA CONSULTAR ALGUNA
TRANSMITA LA INSTRUCCION FUERA

*FUNDACION ARTURO ROSENBLEUTH

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
001	091161614	FAR780725	FUNDACION ARTURO ROSENBLUETH PARA EL AVA	AA

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

*UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
001	180019384	UNA690821	UNIVERSIDAD NACIONAL AUTONOMA DE NAYARIT	CC

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

*NACIONAL HOTELERA

NUM	EXPEDIENTE	R.F.C.	N O M B R E	T.U.M
001	260130656	NHO571004	NACIONAL HOTELERA	BD
002	030018439	NHB770831	NACIONAL HOTELERA BAJA CALIFORNIA SA	CC
003	090882822	NHO571004	NACIONAL HOTELERA SA HOTEL EL PRESIDENTE	BD
004	090793773	NHO571004	NACIONAL HOTELERA KOALA AEROPUERTO	BD
005	090849027	NHO571004	NACIONAL HOTELERA PASSE PARTOUT	BD
006	090781600	NHO571004	NACIONAL HOTELERA REST MUSEO NAL ANTROP	BD
007	090849019	NHO571004	NACIONAL HOTELERA SA REST FOCOLARE	BD

SI NO DESEA CONSULTAR ALGUNA DE ESTAS EMPRESAS
TRANSMITA LA INSTRUCCION FUERA

*ESQUIVEL SANCHEZ RAUL

CONCLUSIONES

CONCLUSIONES.

Día a día las técnicas de computación se van mejorando, nuevos algoritmos y mecanismos se van implementando. Por ejemplo: Las técnicas de uso de bases de datos cada vez son mejores y ofrecen algoritmos para la recuperación de información ya programados y probados, Esto da a las personas que trabajan en sistemas mucha facilidad para el diseño, programación e implementación.

Definitivamente se puede seguir pensando y analizando esta problemática, es posible conseguir la bibliografía más moderna acerca de la recuperación de información y con esto lograr un diseño mejor y más rápido para la recuperación de nombres de empresas,que es el problema que resuelve esta tesis.

Sin embargo el algoritmo propuesto y desarrollado si cumplió con los objetivos que lo originaron, ya que además de lograr la recuperación de nombres como consulta por terminal, ha servido para hacer cruces con catálogos de otras dependencias y con esto se ha logrado una mayor depuración de la información del archivo.

Ahora bien, analizando la técnica usada, vemos que el algoritmo normalmente recupera más nombres de los deseados, esto se debe que al comprimir los nombres, debido a la omisión de vocales muchos que

. . . .

dan iguales. Por ejemplo:

LUGO	LG
LAGO	LG

Sin embargo, esto ha beneficiado más que perjudicado, ya que por una parte tenemos la certeza de que no nos va a faltar información, y por otra parte en muchos casos pueden ser los mismos nombres y tener errores ortográficos u omisiones.

BIBLIOGRAFIA

Allen Kent
Information Analysis and Retrieval
Becker and Hayes, New York
1966

C.J. Date
An Introduction to Data Base Systems
Addison-Wesley Publishing Company, Massachusetts
1977

Elias M. Award
Proceso de Datos en los Negocios
Editorial Diana, México
1968

Sammet Jean E.
Programming Languages
Pretice Hall, Englewood Cliffs
1969

Herman H. Goldstine
The Computer From Pascal to Von Neumann
Princeton University Press, Princeton New Jersey
1973

Saul Rosen
Electronic Computers: A Historical Survey
Computing Surveys
Vol. 1 No. 1
1969