UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE CIENCIAS

DISEÑO DE UN SISTEMA DE RECUPERACION DE INFORMACION DOCUMENTARIA PARA MICROCOMPUTADORA

T E S I S

Que para Obtener el Título de:

A C T U A R I O

P r e s e n t a:

MA. GUADALUPE IZQUIERDO DYRZO

México, D. F. 1988

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

I N D I C E

INTRODUCCION

La época actual se caracteriza por la gran cantidad de informacion que existe sobre cualquier área del conocimiento humano, razon por la cual, debe realizarse un gran trabajo de recopilacion y manejo de información cada vez que se desea hacer un analisis o investigacion en algun tema particular.

La evolucion tecnologica ha propiciado un auge en el estudio y desarrollo de sistemas de informacion apoyados por computadora, que presentan alternativas para el tratamiento de ésta y reducen el tiempo de búsqueda y análisis.

Dentro de estos sistemas de información se han venido distinguiendo los sistemas de recuperación de informacion documentaria, en base a su amplia aplicación, los modelos de representacion de la misma, los mecanismos de recuperación y las amplias ventajas que ofrecen para la localización e identificación de la información requerida en un momento dado.

Debido al avance en el desarrollo de las microcomputadoras y al importante crecimiento de aplicaciones en base a ese recurso de computo, se ha visto la conveniencia de diseñar e instrumentar un sistema de recuperacion de información documentaria para este tipo de equipo, cuyo propósito es el de dar respuesta a las necesidades de manejo de información documentaria.

En el presente trabajo se describe un modelo de sistema de recuperación de información documentaria de aplicacion en microcomputadoras, asi como el desarrollo del modulo de altas para dicho sistema.

Para lograr el objetivo del trabajo, la tesis se divide en cinco capitulos, a saber:

En el primer capitulo se dá una introduccion al concepto de sistemas de informacion, haciendo hincapie en las diferencias que existen entre datos e informacion.

En el segundo capitulo se tratan los conceptos y detalles a considerar para el desarrollo y operación de un sistema de recuperación de información documentaria, destacando los problemas que se presentan en el manejo del lenguaje natural.

En el capitulo tercero se presenta el modelo general del sistema, asi como los detalles técnicos para su desarrollo.

En el cuarto capitulo se detallan las herramientas a utilizar para el desarrollo e implementación del sistema.

Introducidos los aspectos fundamentales, en el capitulo quinto se describe el subsistema de altas. Finalmente, se presentan las conclusiones del trabajo.

Es necesario aclarar que a lo largo del trabajo, se omitió en lo posible, el uso de vocablos extranjeros, sin embargo, se hara uso de algunos debido a que no es fácil encontrar sinonimos que engloben su significado; tal es el caso de la palabra "abstract".

CAPITULO I

SISTEMAS DE INFORMACION

I.1- EL ENFOQUE DE SISTEMAS

Antes de introducirse a sistemas de informacion es recomendable tener una idea precisa de lo que es un sistema.

La palabra sistema se encuentra en diversas campos, asi podemos mencionar:

- El sistema nervioso,
- El sistema capitalista,
- El sistema operativo de una computadora,
- El sistema métrico decimal,
- etc.

¿Pero qué relacion existe entre el sistema nervioso y el sistema métrico decimal? Evidentemente en cuanto a objetivo ninguna, sin embargo analizando su estructura podemos decir que ambos contienen una serie de elementos, los cuales al combinarse alcanzan un objetivo especifico; de esta forma el sistema nervioso envia estimulos al cerebro y el sistema métrico logra dar un parámetro para la medición de longitudes, distancias, volumenes y superficies.

Consultando diferentes fuentes encontramos las siguientes definiciones de "sistema":

"Conjunto de cosas que ordenadamente relacionadas entre si constituyen a determinado objeto" [Q-1]

"Conjunto de reglas o principios sobre una materia enlazados entre si" [S-2]

En Biologia: "Conjunto de organos que intervienen en alguna de las principales funciones vegetativas"[S-3]

Es fácil observar que estas definiciones coinciden en que un sistema es: un conjunto cuyos elementos están definidos para lograr un proposito especifico. Es decir, es un conjunto cuyos componentes interactúan de tal forma que a través de tareas especificas logran conformar u obtener el (los) objetivo(s) para el cual fue diseñado.

En computación, Samuelson [JB-4] dá la siguiente definicion de sistema:

"Colección organizada de hombres, máquinas y material requerido para realizar un propósito especifico ligados todos por lazos comunicativos."

Podemos ver que esta definición se apega a las mencionadas anteriormente.

Tambien podemos tener el caso de un sistema cuyos componentes esten constituidos de varios elementos, los cuales a su vez constituyen un sistema inmerso en el sistema mayor, esto es lo que se conoce como subsistema Asi pues podemos decir que un sistema puede estar constituido a su vez por uno o varios subsistemas.

Teniendo ya un panorama de lo que es un sistema, podremos comprender plenamente lo que se entiende por un sistema de informacion.

Podemos decir que un Sistema de Información es la combinación de recursos por medio de los cuales y a través del procesamiento de datos, se logra obtener un objetivo especifico, o como John G. Bursch [JB-5] lo define:

"Conjunto sistematico y formal de componentes, capaz de realizar operaciones de procesamiento de datos."

O la definición de K. Samuelson [JB-6]:

"Colección de recursos y bases computacionales principales, los cuales resultan en la colección, recuperación, comunicación y uso de datos con el propósito de un manejo efectivo."

Sin embargo, ¿Que es un dato?, ¿Que es informacion?, ¿Que relacion existe entre ambos conceptos? En qué difieren uno de otro?. En la siguiente sección se amplian estos conceptos.

I.2- DATOS E INFORMACION

Es muy generalizado el hecho de utilizar los conceptos de DATOS e INFORMACION como sinonimos, sin embargo, esta sinonimia es aparente ya que en realidad denotan conceptos diferentes.

Por esta razón es necesario distinguir ambos conceptos y tener una idea precisa y clara de lo que cada uno de ellos significa, de esta forma podremos reafirmar el concepto de "Sistemas de Información" ya tratado anteriormente.

Con este fin a continuación se citan algunas definiciones de datos e información.

DATOS

"Son hechos aislados y en bruto los cuales situados en un concepto significativo mediante una o varias operaciones de procesamiento, permiten obtener deducciones relacionadas con la evaluación e identificacion de personas, eventos y objetos." [JB-7]

'Son hechos meramente individuales los cuales deben ser combinados o 'procesados' de alguna manera para darles significado." [D-8]

"Hechos, el renglon material de la informacion." [SA-9]

INFORMACION

"Es el significado que los humanos asignan a los datos por medio de convenciones usadas en sus representaciones." [SA-10]

"Procesamiento de datos que puede proporcionar un conocimiento o bién el entendimiento de ciertos factores. "[JB-11]

"Conocimientos adquiridos por el receptor mediante datos procesados." [JB-12]

Analizando y comparando las definiciones podemos decir que un dato es un hecho individual o aislado, el cual por si solo no tiene significado, y la información es el conocimiento que obtenemos a través del procesamiento de datos.

Asi mismo podemos decir que informacion significa: "un aumento de conocimientos para el receptor mediante la coordinacion apropiada de los elementos de los datos con las variables en problema". [JB-13]

Podemos concluir que la funcion de la informacion es aumentar los conocimientos del receptor (que llamaremos usuario), mientras que los datos por si mismos no proporcionan significado alguno.

I.3-RECUPERACION DE DATOS Y RECUPERACION DE INFORMACION

Como ya se menciono en la sección anterior, las palabras "datos" e "informacion" no son sinonimos, por lo que es de esperarse que los términos "recuperación de datos" y "recuperacion de información" no sean equivalentes.

Para caracterizar a los terminos "recuperación de datos" y "recuperación de información" nos basaremos en el criterio de C. J. van Rijsvergen [VR-14], para lo cual utilizaremos la siguiente tabla comparativa que nos da 7 puntos que diferencian a cada uno de estos conceptos.

	RECUPERACIÓN DE DATOS	RECUPERACIÓN DE INFORMACIÓN
Correspondencia	Exacta	*Parcial.la mejor correspondencia
Inferencia	Deductiva	Inductiva
Modelo	Determinístico	Probabilístico
Lenguaje Interrogativo	Artificial	Natural
Especificación de la pregunta	Completa	Incompleta
Términos requeridos	Correspondientes	Relevantes
Respuesta al error	Sensitiva	Insensitiva

* En nuestro contexto, "la mejor correspondencia" es la que presenta mayor número de similitudes y/o puntos en común con la requerida.

a) En recuperación de datos (RD) debe localizarse el dato exacto, mientras que, en recuperacion de información (RI) lo que se busca es encontrar aquella información que corresponda parcial o totalmente a la interrogativa. De dicha información se selecciona la que mejor concuerda con la buscada.

Por ejemplo, si se está trabajando con información documental y se busca la palabra 'niñez' y encontramos la palabra 'infancia' ésta debe de ser recuperada ya que ambas son sinonimos.

b) La inferencia utilizada (inferir es sacar una consecuencia) en RD es de tipo deductiva es decir se procede de lo universal a lo particular, en cambio en RI se va de lo particular a lo general, "de las partes al todo", por lo que la inferencia que se utiliza es de tipo inductiva.

c) Del punto b) y debido a las propiedades de inferencia utilizada en cada caso, se tiene que en RI se maneja un modelo probabilistico ya que las relaciones se establecieron con cierto grado de certeza y por lo tanto la confianza en la inferencia es variable. En cambio en RD se tiene un modelo deterministico (aqui no se hace uso de las probabilidades).

d) El lenguaje interrogativo en RD generalmente es de tipo artificial, es decir, un lenguaje con sintaxis y vocabulario restringido, a diferencia del lenguaje utilizado en RI que es de tipo natural. En el siguiente capítulo ahondaremos más en estos términos.

e) En RD se tiene que para hacer una pregunta se necesita una especificación completa y detallada de lo que se desea, a diferencia de la que se necesita en RI cuya pregunta puede

ser incompleta, esta distinción se debe básicamente a que, en RI buscamos la información relevante y no una concordancia exacta. Esta caracterización es consecuencia del punto a).

f) Dado que en RD se desea una concordancia exacta, se tiene que la recuperación de datos es más sensible al error, ya que si se comete un error de concordancia no se recuperará el elemento deseado, lo que implica que fracasó la recuperación; en cambio, en RI pequeños errores en la concordancia por lo general no afectan significativamente la recuperación.

Una vez detalladas las diferencias fundamentales que implica la recuperación de datos y la recuperación de información es posible introducirse al tema de sistemas de recuperación de información.

CAPITULO II

SISTEMAS DE RECUPERACION DE INFORMACION DOCUMENTARIA

II.1- CONSIDERACIONES Y MODELO BASICO

Desde el año de 1940 ha venido creciendo la atención prestada a los problemas de almacenamiento y recuperación de informacion, esto es natural ya que cada vez se vuelven más necesarios los sistemas de recuperación de información.

Por ejemplo, consideremos una biblioteca con un acervo grande, que requiere llevar un control del material bibliográfico lo que hace necesaria su catalogación, lo cual no es una tarea muy agradable. De esta forma si queremos consultar determinado libro primero debemos revisar los catálogos para conocer si la biblioteca cuenta o no con el titulo buscado.

Lo que se desearia es el poder ser ayudados por una computadora a realizar estas actividades, de esta forma si una persona desea consultar un libro, sólo le proporcionaria el titulo a la máquina y ésta determinaria si se cuenta con el volumen o no, sin necesidad de buscarlo en los catálogos.

Intuitivamente esto es lo que se conoce como un Sistema Recuperador de Información (SRI). Teniendo una idea de lo que es un SRI podemos dar la siguiente definición [L-1]:

"Un sistema recuperador de información no cambia el conocimiento del usuario sobre la materia objeto de su pregunta, solamente le informa sobre la existencia, no existencia y localización de los documentos relacionados con la pregunta."

Un Sistema Recuperador de Información básico puede verse como un conjunto de 4 elementos: (ver figura II.1)

- Entrada
- Procesador
- Banco de Información
- Salida

Donde cada elemento tiene una función especifica y a su vez está constituído complejamente por otros componentes. A continuación damos una breve explicación de la función de cada uno de estos elementos.

a) ENTRADA

La entrada es el elemento que se encarga de tomar las preguntas, las cuales son el motivo de la interrogación.

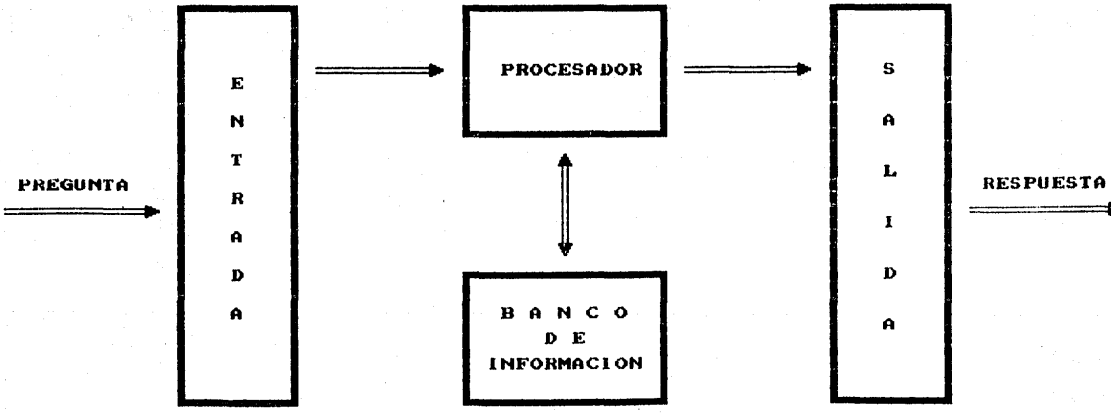

Fig. II.1

MODELO BASICO DE UN SISTEMA RECUPERADOR DE INFORMACION

b) PROCESADOR

El procesador es la parte del sistema que se encarga de realizar la recuperación, que puede incluir la estructuración de la información de una forma apropiada como puede ser el clasificarla. Así mismo es la parte encargada de realizar la estrategia de recuperación como respuesta a una pregunta planteada.

c) BANCO DE INFORMACION

El banco de información constituye "el saber" es decir, el conjunto de conocimientos o información que están contenidos en la computadora y que son suceptibles de recuperación.

d) SALIDA

La salida, es el elemento por el cual se emite la información que se obtiene como respuesta a la interrogación dada.

Se ha visto que un SRI necesita de un banco de información, en el caso de un sistema recuperador de información documentaria (SRID) el banco está constituido por la información contenida en los documentos, entonces surgen de manera natural la siguientes interrogativas: ¿Cómo proporcionar la pregunta, motivo de nuestro interés, al sistema?, ¿Cómo manejar los textos?

II.2- REPRESENTACION DE LA INFORMACION DOCUMENTARIA

Podemos partir del hecho de que en el banco de información no necesariamente se almacena el texto completo en el lenguaje natural en el que fue escrito, sino que en su lugar tiene un documento representativo; dicha representación debe ser tal que cuando sea relevante a una pregunta, permita que el documento sea recuperado en respuesta de la misma.

De esta forma el problema se traduce en diseñar un formato adecuado para proporcionar los documentos a la máquina para posteriormente poder realizar su recuperación. En este formato se debe contemplar el tipo de información que contiene el documento, asi como la forma en que será sustraida.

Para llevar a cabo dicha caracterización del documento, existen tres técnicas identificadas por: Indexación, Full-text y Abstract. A continuación se describe cada una.

INDEXACION

Indexación es la elaboración de una lista de palabras clave o descriptores, los cuales son utilizados para obtener la información contenida en la ficha de análisis, asi como para la elaboración de dicha ficha. La elaboración de esta lista de palabras clave puede realizarse en dos formas:

1- Llevar a cabo una valuación de la información contenida en los documentos fuente para así determinar aquellas que son representativas.

2- Crear la lista de palabras de forma independiente al documento.

Así mismo se parte de la suposición de que si dos palabras contienen la misma raíz deben referirse al mismo concepto y deben de ser indexadas de la misma forma; sin embargo de esta manera se produce una sobresimplificación ya que palabras como NEURONA y NEURITO contienen la misma raíz y sin embargo necesitan ser distinguidas ya que representan conceptos diferentes.

De lo anterior se tiene que un lenguaje indexado es el lenguaje utilizado para describir documentos y preguntas; cuyos elementos son los términos indexados los cuales como ya mencionamos son derivados del texto a describir.

Así mismo se tiene que un lenguaje indexado puede ser precoordinado o postcoordinado.

Se dice que es PRECOORDINADO si los términos son coordinados a la hora de la indexación, en cambio es POSTCOORDINADO si los términos son coordinados a la hora de la búsqueda.

Esto es, si es precoordinado, al indexar una combinación lógica de cualesquiera términos indexados, esta combinación se utiliza como etiqueta para identificar los documentos, en cambio si es postcoordinado el indexar los mismos términos se traduce en combinar los documentos recuperados con cada uno de lo términos indexados por separado, de forma tal que se tenga como resultado la combinación lógica requerida.

Veamos como funcionaría el sistema, por ejemplo, si se tiene un banco de información legislativo y se desea conocer lo relacionado con la expedición de licencias de manejo, planteando la pregunta "expedición de licencias". Un sistema postcoordinado lo que hará es buscar los documentos donde se localice la palabra 'expedición' así como los documentos donde se encuentre la palabra 'licencias', tomando como relevantes aquellos documentos donde se localicen ambos términos; en cambio si es precoordinado, lo que hará es buscar los documentos donde se encuentren las palabras 'expedición' y 'licencias' seguidas una de otra.

TEXTO INTEGRAL (Full-Text)

Como su nombre lo indica, esta técnica consiste en almacenar el texto completo en la computadora (e. i. se toman como palabras clave todas y cada una de las palabras contenidas en el documento), de esta forma se puede accesar el documento por cualquiera de las palabras que contenga.

En la aplicación de este método, debe tomarse en cuenta que se requerirá de una cantidad grande de memoria en la computadora, lo cual puede implicar un alto costo, por lo que esto puede ser considerado como desventaja si la aplicación se tratara de implementar en una microcomputadora o en un equipo con poca memoria; pero puede considerarse viable el método si se cuenta con una computadora de gran capacidad destinada para el uso exclusivo del sistema recuperador de información. Otro de los aspectos importantes a considerar es que en la aplicación de este método se pasan por alto conceptos implícitos, los cuales pueden ser motivo de recuperación.

ABSTRACT

Como una alternativa para la representación de la información documentaria se tiene el "abstract"(1), el cual consiste en obtener a partir del documento fuente las ideas principales que contiene y omitir lo que no es relevante. Así se tiene un estracto de lo más importante del documento, sin ser por ello la información meramente escueta.

La información así obtenida se organiza de forma lógica para lograr su recuperación.

---

1 "ABSTRACT" Vocablo inglés cuyo significado es: sumario, compendio. Para nuestros fines utilizaremos este vocablo denotando la técnica descrita en esta sección.

El utilizar esta técnica nos ofrece muchas ventajas ya que al poder utilizar un lenguaje natural, es posible extenderse en conceptos o ideas que son importantes pero que se tratan vagamente o que se encuentran implícitos dentro del documento; de la misma forma permite suprimir información considerada como no relevante y de esta manera optimizar memoria.

Los sistemas que utilizan el método de indexación, por consistir en una lista de descriptores, se dice que utilizan un lenguaje cerrado; en cambio los que utilizan el abstract son sistemas de lenguaje abierto ya que al contrario de la indexación, tiene un conjunto libre de palabras a utilizar.

En el sistema, propósito de nuestro trabajo, se utiliza la técnica de abstract para la representación de los documentos, lo cual nos lleva a un sistema con lenguaje abierto basado en nuestro lenguaje natural.

II.3- CONSIDERACIONES EN LA RECUPERACION DE INFORMACION DOCUMENTARIA

Supóngase que se tiene una colección de documentos los cuales pueden ser recuperados como respuesta a determinadas preguntas. Para que el sistema determine cuales son los

documentos relevantes,, éste debe "leer" todos y cada uno de los documentos; para que esto se logre la computadora tiene que emular el proceso de lectura humana.

Este proceso de lectura es complicado ya que implica extraer información, la cual está sujeta al tema y al contexto en que se está hablando, es decir, hay que tomar en cuenta la sintaxis y la semántica del lenguaje, así, como la información que sobre el tema se necesite para poder comprender el documento, y a partir de esta información decidir si el documento es relevante o no.

Sin embargo, aquí se presenta un problema ya que la mente humana establece que documentos son relevantes y cuales no lo son respecto a cierta pregunta, a diferencia de la computadora que no puede realizar estas acciones, por lo que se necesita un modelo en el cual puedan tomarse decisiones de este tipo.

Tómese en cuenta también que la mente humana es capaz de manejar conceptos imprecisos mientras que la computadora no está capacitada para realizarlo.

Reafirmando, la máquina tiene que simular el proceso de lectura y el proceso para determinar que documentos son importantes y cuales no, entonces es necesario tomar en

cuenta que el sistema debe realizar la simulación del proceso de lectura y que esto implica "un conocimiento acerca del mundo descrito por las oraciones." [ER-2]

El objetivo de una estrategia de recuperación automática, es el de recuperar todos los documentos relevantes y a la vez recuperar el mínimo número de documentos no relevantes.

II.4- PROBLEMAS EN EL MANEJO DEL LENGUAJE NATURAL

Si un SRI adopta el lenguaje natural es necesario analizar los problemas que se presentan como consecuencia de las características del lenguaje, así como de las diferencias existentes entre el pensamiento humano y el funcionamiento de una computadora.

Para analizar estos problemas (los cuales llamaremos lingüísticos) los dividiremos en gramaticales y semánticos. Los gramaticales abarcan cuestiones de tipo sintáctico (la función de las palabras en las oraciones) y morfológico. Los problemas de tipo semántico son precisamente los concernientes al significado de las palabras.

a)Gramaticales

Un problema muy común es el hecho de poder expresar una idea de varias formas utilizando palabras que morfológicamente son distintas pero que semánticamente denotan lo mismo, por ejemplo las siguientes tres oraciones aunque diferentes, denotan la misma idea:

Se ha nombrado un nuevo gobernador.
Se nombró nuevo gobernador.
Fue nombrado nuevo gobernador.

De esta forma el problema que se presenta es que el sistema deber recuperar frases que semánticamente son iguales pero que no lo son morfológicamente.

b) Semánticos

La semántica es el estudio del significado de las palabras y sus variaciones. Dentro de ésta se nos presentan los siguientes conceptos:

-Sinonimia
-Analogía
-Polisemia

A continuación se dá una breve explicación de que consiste cada uno de ellos.

i)Sinonimia (circunstancia de ser sinónimos dos o más vocablos)

Es la existencia de dos o más palabras con el mismo contenido semántico, por ejemplo:

-Educado
-Instruido
-Culto
-Correcto

ii)Analogia (similitud)

Es la posibilidad de referirse a un mismo concepto con palabras cuyo contenido semántico es similar, ejemplo:

-Nombramiento del gobernador
-Designación del gobernador
-Elección del gobernador

II)Polisemia (multiplicidad de acepciones en una misma voz o término)

Se dice que una palabra es polisémica cuando expresa más de un significado, por ejemplo:

ARCO
- de un circulo
- de un puente
- como arma para flechas
- de cerdas para tocar el violin
- hueso de forma arqueada

INDICE
- lista de capitulos de una obra
- dedo de la mano
- matemático (relación entre dos cantidades)
- manecilla de reloj

Una vez detallados los conceptos lingüísticos resultan evidentes los problemas que debe resolver un SRI ya que si por ejemplo se está interesado en el tema "niñez y juventud" y se tiene "infancia y juventud" el sistema debe ser capaz de recuperar esta información.

II.5- INSTRUMENTOS LINGUISTICOS

Para resolver los problemas que se presentan en un SRI por el uso del lenguaje natural, se definen dos instrumentos lingüísticos: Léxico y Thesaurus.

a) Léxico

Es el instrumento lingüistico que nos permite resolver los problemas de sintáxis, sinonimia y algunos problemas de polisemia, en base a la organización de palabras contenidas en los documentos al seguir ciertos criterios morfológico-semanticos.

El léxico puede estructurarse en nociones. Una NOCION es la agrupación de palabras que tienen la misma raiz; cada noción puede admitir a su vez varias subnociones; en cada subnoción se almacenan las palabras que se consideran semanticamente equivalentes; los accidentes gramaticales de género y número pueden ser ubicadas en la misma subnoción.

Por ejemplo, obsérvese el siguiente esquema de noción dividida en sus respectivas subnociones.

- Contrato
  contratos
- Contraten
  contratar
  contratado
- Contratista
  contratistas

De esta forma pueden ser resueltos los problemas de sinonimia y polisemia ya que al agrupar las palabras bajo cierto criterio, permite aislarlas de tal forma que no provoquen sobrerecuperación, pero que al mismo tiempo sea posible reagruparlas en la familia a la que pertenecen originalmente.

b) Thesaurus

Es el instrumento lingüistico que nos permite resolver problemas de analogia y antonimia considerando las palabras contenidas en el léxico y agrupándolas de tal forma que, en un mismo grupo se localicen aquellas que tienen el mismo sentido semántico. Vgr. sinónimos.

El thesaurus permite al usuario conocer tanto el contenido semántico de las palabras por las cuales se interroga, como el de los términos análogos que podrán utilizarse para precisar la pregunta.

CAPITULO III

DISEÑO DE UN SISTEMA RECUPERADOR DE INFORMACION DOCUMENTARIA

III.1- MODELO GENERAL

Para conceptualizar el sistema recuperador de información documentaria (SRID) que es de nuestro interés, y partiendo de lo mencionado en capitulos anteriores, veamos que se requiere determinar.

1.- Diseñar un formato en base a la representación documentaria que se ha establecido aplicar y por el cual se deben proporcionar los documentos al sistema. Asi como determinar los criterios para la codificación de los documentos.

2.- Los procedimientos para el análisis de información y captación de las fichas.

3.- Un proceso que analice y almacene las fichas válidas que posteriormente sean recuperables.

4.- Un proceso que cumpla nuestro modelo general de SRI (entrada, procesador, banco de información, salida).

En base a lo anterior nuestro modelo de SRID se puede definir con los siguientes componentes:

1.- Fichas de Información.

2.- Un subsistema que realice el filtrado de las fichas para determinar cuales son aceptadas o cuales no, y que almacene las fichas en el banco de información. A este subsistema lo denominaremos Subsistema de Altas.

3.- Una Base de Datos constituida por: el Banco de Información (véase modelo general de un SRI) y las estructuras auxiliares que faciliten el proceso de recuperación, que a su vez incluyan los esquemas para los instrumentos lingüisticos.

4.- Un susbsistema que corresponda al objetivo fundamental del sistema, el cual denominaremos Subsistema de Consulta y que es por el cual se reciben las interrogaciones, se procesa y se producen los resultados correspondientes. Vease la figura III.1 que muestra esquemáticamente el modelo del SRID.

A continuación se explican las caracteristicas de cada uno de los componentes.

III.2- FICHA DE INFORMACION

Recordemos que la técnica que se consideró aplicar en el sistema es la de abstract, por lo tanto, en base a ello describiremos la ficha.

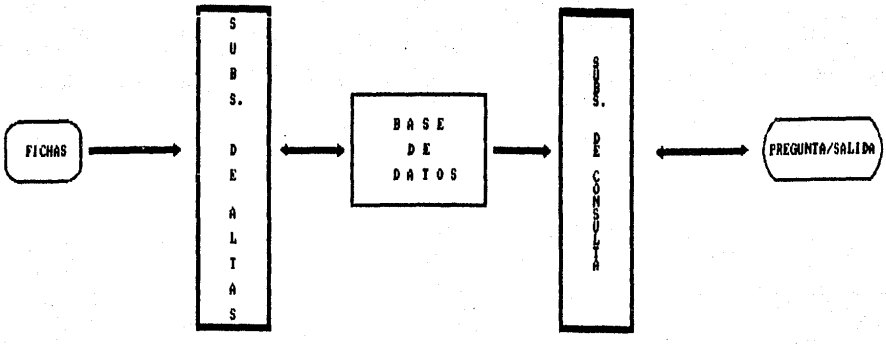

Fig. III.1

MODELO DE UN SISTEMA RECUPERADOR DE INFORMACION DOCUMENTARIA

El diseño del documento o ficha, se divide en tres zonas:

- Zona de claves
- Zona de resumen
- Zona de abstract

ZONA DE CLAVES:

Es la parte de la ficha en la que se puede tener información referente al origen o identificación externa del documento. Esta información puede ser: el tipo de documento, de donde proviene la información (libro, revista, periódico, etc.), fecha de publicación del documento, etc. Esta zona puede ser variable de acuerdo a la aplicacion que se lleve a cabo.

ZONA DE RESUMEN:

Esta parte se reserva para un extracto de las ideas fundamentales que contiene el documento, de modo tal que en la fase de recuperación, al leer esta zona, se tenga una idea clara del contenido global de la ficha.

ZONA DE ABSTRACT:

Es el cuerpo principal de la ficha y cuyo contenido será base de la recuperación.

El generar el abstract de las fichas nos lleva a situaciones que hay que controlar como es tamaño y cantidad de ideas incluidas, para lo cual es necesario estructurar el

contenido; esta estructuración puede formalizarse en similitud a como normalmente se elaboran textos, que es en frases, párrafos, etc.

Para efectos del tratamiento de la información por computadora es necesario distinguir la estructura de la ficha; debemos delimitar el inicio y término de la misma, de la frase, del párrafo, etc. Para ello se prevee el uso de una serie de caracteres que recibirán el nombre de "caracteres especiales", cada uno de ellos tendrá una aplicación especifica. A continuación se tiene un cuadro con cada uno de estos caracteres así como su significado.

CARACTER	SIGNIFICADO
*/	- inicio de abstract
/*	- fin de abstract
/./	- fin de párrafo
/	- fin de oración o frase

CARACTERES ESPECIALES

Obsérvese el siguiente abstract [UJ-1]:

/CONSTITUCION POLITICA ESTATAL, ARTICULO 3, REFORMA /MUNICIPIO DE JALAPA, CAMBIO DE DENOMINACION, AHORA JALAPA DE MENDEZ/

Se puede ver que en el ejemplo el abstract consta de dos oraciones, un párrafo y 16 palabras.

Como se mencionó, es necesario tener criterios para determinar cuando el documento se apega a la representación establecida o no, e.d. que la ficha esté construida correctamente con el fin de que las fichas que cumplan con este requisito, pasen a formar parte del "SABER" del BANCO DE INFORMACION, y de esta forma las fichas queden disponibles para efectos de recuperación.

III.3- SUBSISTEMA DE ALTAS

El Subsistema de Altas es la parte que se encarga de procesar el conjunto de fichas que se desea integrar al banco de información validando y almacenando la información, además, se encarga de detectar, analizar y registrar toda la información necesaria para facilitar la recuperación.

Para realizar el proceso se ha previsto utilizar unas técnicas especiales con el fin de detectar cada una de las palabras contenidas en el abstract de la ficha y su posición dentro de la misma, produciendo como resultado la aceptación o rechazo del documento, entendiéndose como rechazo de un documento cuando no cumpla con los criterios establecidos para ser almacenado en el Banco de Información.

III.4 LA BASE DE DATOS

La Base de Datos debe contener la información susceptible de recuperar y todos los elementos necesarios para facilitar dicha recuperación. La base de datos se divide en cuatro estructuras, a saber:

- Contenido de las fichas
- Descriptores de las fichas
- Descriptores de palabra
- Instrumentos lingüísticos

CONTENIDO DE LAS FICHAS

En esta estructura, que en lo sucesivo denominaremos **FICHAS**, se almacena la información contenida en los documentos que se hayan capturado y aceptado a través del Subsistema de Altas, que en escencia es nuestro Banco de Informacion a recuperar.

DESCRIPTORES DE FICHAS

Esta estructura contiene un conjunto de descriptores de cada una de las fichas en nuestro Banco de Información, en donde cada descriptor está constituido por la identificación de la ficha y su posición inicial y final en la estructura FICHAS; para mayor precisión a continuación se describen con mayor detalle estos aspectos.

IDENTIFICACIÓN DE FICHA: Es un número que se asigna a cada uno de los documentos para distinguirlos en forma única en el Banco de Información. La forma de asignación del número se sugiere secuencial en la elaboración de las fichas.

POSICION INICIAL: Es el dato que nos precisa el numero de registro en donde inicia el contenido de cada ficha.

POSICION FINAL: Es el dato que nos indica el número de registro al término del contenido de la ficha,

De esta manera si se tiene que:

- Identificacion de ficha = 5
- Posición Inicial = 12
- Posicion Final = 15

Se interpretará de la siguiente forma:

1- Se refiere a la ficha identificada con el número 5.

2- El inicio de la ficha se encuentra en el registro número 12.

3- El término de la ficha se localiza en el registro numero 15.

A estos descriptores de la ficha les llamaremos **FORMA GAMMA** del documento, y debe de existir un único descriptor por cada ficha. En base a estos descriptores se puede lograr la localización precisa de cada uno de los documentos contenidos en FICHAS. El conjunto de estos descriptores lo denominaremos **GAMMA**.

DESCRIPTORES DE PALABRAS

Esta estructura contiene información que permite determinar y localizar las palabras que conforman el abstract de las fichas, identificando la frase y párrafo donde se encuentran.

La información que contienen estos descriptores es la siguiente:

- Palabra
- Número de palabra dentro de la frase
- Número de frase dentro del párrafo
- Número de párrafo dentro de la ficha
- Número de ficha a la que pertenece la palabra. Este número debe coincidir con el respectivo de la forma GAMMA correspondiente a esta ficha.

Por ejemplo si se tiene el siguiente descriptor:

- Palabra : REGIMEN
- Número de palabra : 12
- Número de frase donde se encuentra: 1
- Número de párrafo :2
- Número de ficha :3

Se interpreta de la siguiente forma:

- la palabra que se detectó fue régimen.
- Es la doceava palabra de la primera frase del 2o. párrafo dentro de la ficha número 3.

De esta forma se debe tener un descriptor por cada una de las palabras contenidas en la ficha. Una vez reconocidas las palabras de estos descriptores, se transfiere la informacion a la estructura que almacenará los descriptores de palabra. Esta estructura la denominaremos **FORMAS LAMBDA**, su contenido será el del descriptor auxiliar exceptuando la PALABRA.

La estructura se define de tal forma que queden unidos en conjuntos de descriptores de cada palabra cuyo propósito será simplificar la recuperación.

INSTRUMENTOS LINGÜISTICOS

Para nuestro modelo solo se ha considerado lo correspondiente al lexico, dado que este instrumento es determinante para un modelo básico de SRID, a esta estructura la denominaremos **LEXICO**.

Como se habia indicado, el léxico debe resolver los problemas de sintaxis y sinonimia, por lo tanto en este instrumento lingüistico debe organizarse el conjunto de palabras en las estructuras de noción y subnoción con el fin de auxiliar la fase de recuperación, además la detección de existencia o no de palabras cuando se trate de almacenar al banco de información un nuevo conjunto de fichas.

Para organizar las estructuras de nocion y subnocion cada una de las palabras contenidas cuentan con una serie de referencias. La primera nos lleva a la localizacion en LAMBDA de las fichas aceptadas; la segunda establece las

relaciones entre subnociones; la tercera, la relación entre nociones, por último, se tiene una cuarta referencia para detectar la palabra que encabeza cada subnoción.

De esta forma, una porcion de esta estructura se vería así:

Académico	1	2	3	1
Académicos	2	1	3	0
Académica	3	4	1	1
Académicas	4	3	1	0
México	10	0	21	1
Mexicano	11	0	22	1
Mexicana	12	0	20	1

Permitiendo así que, al accesar la palabra "académico", ésta nos lleve a académicos, académica y académicas recuperando los documentos que contengan dichas palabras. Cabe destacar que si existen varios grupos de descriptores de palabra, el direccionamiento de LEXICO hacia esta estructura, se refiere al último conjunto de descriptores utilizados.

De esta forma podemos visualizar la Base de Datos en la figura III.2.

III.5- SUBSISTEMA DE CONSULTA

El subsistema o módulo de consulta es la parte del sistema, que se encarga de recibir las preguntas de los usuarios y emitir las respuestas, e.i. determinar cuales son los documentos de interés para el usuario e informarselo.

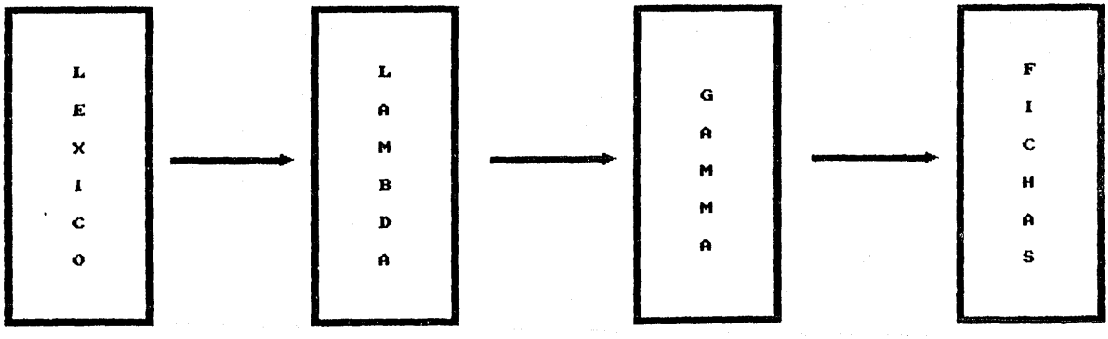

Fig. III.2

ESTRUCTURA DE LA BASE DE DATOS

EL subsistema de consulta al recibir una interrogacion separa las palabras en relevantes y no relevantes; son consideradas no relevantes cuando no modifican el sentido de la oracion, por ejemplo los articulos y las preposiciones.

Las palabras relevantes son las que se tomarán en cuenta para la recuperación. Estas palabras son buscadas en el léxico; una vez localizadas con sus referencias respectivas de la estructura LEXICO, se hace el acceso en la estructura LAMBDA, obteniéndose los descriptores de palabra correspondientes.

Una vez que se cuenta con cada uno de estos descriptores, se selecciona aquél subconjunto que cumpla con los requerimientos de acuerdo a la interrogativa. Con esto se garantiza que el objeto de la interrogante se encuentra en la ficha o fichas cuyo número de referencia está dado por su descriptor.

Ya que se tiene(n) la(s) ficha(s) de posible interés a la interrogante, se tomarán sus descriptores de ficha para determinar su posición en FICHAS y de esta forma poder emitir la información contenida en cada una ellas.

CAPITULO IV

ASPECTOS TECNICOS

Una vez desglosado el modelo general, es necesario conocer las herramientas a utilizar para el desarrollo del subsistema de altas.

IV.1- EQUIPO

El sistema que nos ocupa es para implementarse en microcomputadores. Por lo cual para determinar el equipo en que se desarrolle y opere el sistema, se tomó en cuenta lo siguiente:

- Caracteristicas de equipo
- Ventajas del sistema operativo
- Transportabilidad del sistema operativo
- Disponibilidad de equipo

En base a esto, se optó por utilizar una computadora de tipo personal, ya que este tipo de microcomputadora se apega a los requerimientos y tipo de implementaciones que deseamos.

Especificamente, se eligio una microcomputadora con sistema operativo MS-DOS debido a la alta comercialización asi como la sencillez de operación del mismo y su implementación en diferentes equipos.[M-1]

De acuerdo a los recursos disponibles el desarrollo del sistema se definió realizarlo en una microcomputadora Printaform cuyos componentes son: pantalla, teclado y CPU con capacidad de 512Kb de memoria y disco duro de 20 Mb.

IV.2- SOFTWARE

Para determinar el lenguaje, se tomaron en cuenta las caracteristicas del software con que se contaba.

En base a esto, se llegó a la conclusión de utilizar alguno de los siguientes lenguajes: C o Turbo-Pascal.

Para determinar cual de los dos lenguajes utilizar, se tomó en cuenta:

- Portabilidad del Lenguaje
- Rapidez de ejecución
- Facilidad de operación
- Caracteristicas necesarias para el manejo de información documental.

En base a estos criterios se elaboró un pequeño programa que generara una lista de números aleatorios y los mandara escribir a un archivo externo; después de este proceso, se accesaba el archivo en el n-ésimo registro y se mostraba.

El resultado de esta prueba fue favorable a Turbo-Pascal debido a su mayor eficiencia en la ejecución de dicho programa.

Por otra parte Turbo es sencillo de operar, así mismo cuenta con variables de tipo CHAR (variables de un solo caracter), que nos facilitan la implementación del sistema. De la misma forma, permite el uso de subprogramas externos y otras herramientas que agilizan la ejecución y permiten mayor aficiencia en el sistema.[M-2]

IV.3- AUTOMATAS

Como herramienta auxiliar se consideró utilizar los autómatas, en particular, una máquina de Mealy.

"Cuando decimos AUTOMATA nos referimos a un modelo matemático, cuyas propiedades y comportamiento podemos simular con un programa de computadora." [VE-1]

En particular, los autómatas, son muy útiles para el diseño de compiladores ya que son sencillos de programar y a su vez son eficientes.

Principalmente se tienen tres tipos de autómatas: aceptadores, generadores y transductores.

ACEPTADORES

Como su nombre lo indica su función es la de aceptar o rechazar cadenas de símbolos. Básicamente es una máquina a la que se le alimenta una sucesión de símbolos. Estos símbolos pertenecen a un conjunto finito de caracteres llamado ALFABETO DE ENTRADA.

Para esto se supone que la máquina (M) se encuentra en un estado fijo al principio del proceso. Este estado es llamado ESTADO INICIAL.

GENERADORES

En este tipo de autómata también se asume que se está en un estado inicial, de modo que cuando el autómata empieza a funcionar, regresa símbolos de un alfabeto finito llamado ALFABETO DE SALIDA. Decimos que el lenguaje generado por M, al que denominaremos L(M), es el conjunto de todas las sucesiones que puede producir.

TRANSDUCTORES

A este tipo de máquina al alimentarle una sucesión del alfabeto de entrada, produce otra sucesión a modo de respuesta, donde cada símbolo de esta sucesión es elemento de un alfabeto de salida.

Una vez que sabemos que existen básicamente tres aplicaciones de los autómatas, daremos la definición.

Un autómata finito M consiste de un conjunto de estados K y un conjunto de transiciones de estado a estado que ocurren cuando el autómata se le alimentan símbolos de un alfabeto.

AUTOMATAS FINITOS

Un autómata finito es una quinteta (K,Σ,δ,q_o,F) donde:

K es un conjunto finito y no vacío de estados.

Σ es un alfabeto finito de entrada

δ es una función llamada de transición.

$$\delta : K \times \Sigma \longrightarrow K$$

q_o está en K y es el estado inicial

$F \in K$ es el conjunto de estados finales

Al alimentarle un símbolo a la máquina se produce en ésta un cambio de estado que depende exclusivamente del estado en que se encuentra y del símbolo proporcionado. A este cambio de estado se le llama TRANSICION.

De esta forma se tiene que si al terminar de alimentar al autómata una sucesión, la última transición fue a un estado $q \in F$ decimos que el autómata acepta la sucesión, de lo contrario decimos que la rechaza.

Para especificar la función de transición se cuenta con tres métodos, el primero es dar explícitamente el conjunto de reglas de transición. Por ejemplo:

Supóngase que se desea construir un autómata para que dada una sucesión de caracteres, determine si contiene o no un número par de ceros y unos. De modo que si esto sucede, la sucesión sea aceptada o rechazada. De esta forma tendríamos los siguientes elementos:

$M = (K, \Sigma, \delta, q_0, F)$

$\Sigma = \{0,1\}$ ya que las cadenas solo contendrán ceros y unos.

$K = \{q_0, q_1, q_2, q_3\}$ el conjunto de estados

$F = \{q_0\}$

donde:

$$(q_0, 0) = q_2$$
$$(q_1, 1) = q_3$$
$$(q_2, 0) = q_0$$
$$(q_3, 1) = q_1$$
$$(q_0, 0) = q_1$$
$$(q_1, 1) = q_0$$
$$(q_2, 0) = q_3$$
$$(q_3, 1) = q_2$$

Otra forma de especificarla es mediante un diagrama de transición. Para elaborar un diagrama de transición es necesario remarcar de alguna forma el estado inicial y dibujar un nodo de la gráfica en cada estado en K. De esta forma el diagrama de transición para el autómata anterior sería el siguiente:

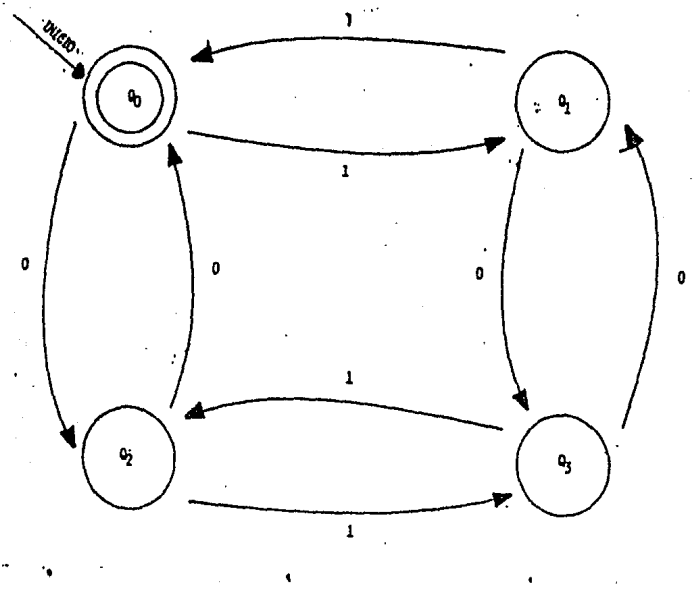

Fig. IV.1
Diagrama de transición de un autómata que determina si la sucesión contiene un número par de ceros y unos.

La tercera y última forma de representar la función es por medio de una tabla de transición, constituída por la matriz del alfabeto contra los estados y cada entrada de la matriz es el estado de transición.

La tabla de transición para el autómata anterior es la siguiente:

K \ Σ	0	1
q_0	q_2	q_1
q_1	q_3	q_0
q_2	q_0	q_3
q_3	q_1	q_2

TABLA DE TRANSICION

AUTOMATAS FINITOS NO DETERMINISTICOS

Este tipo de autómata es una modificación a la definición de autómata descrito anteriormente. El cambio es el siguiente: estando en cierto estado y al leer un simbolo del alfabeto de entrada, el autómata puede transferirse no sólo a un estado sucesor, sino que puede transferirse a más de un estado.

A continuación se tiene el diagrama de un autómata no determinístico que acepta el conjunto de oraciones que contengan dos ceros o dos unos consecutivos.

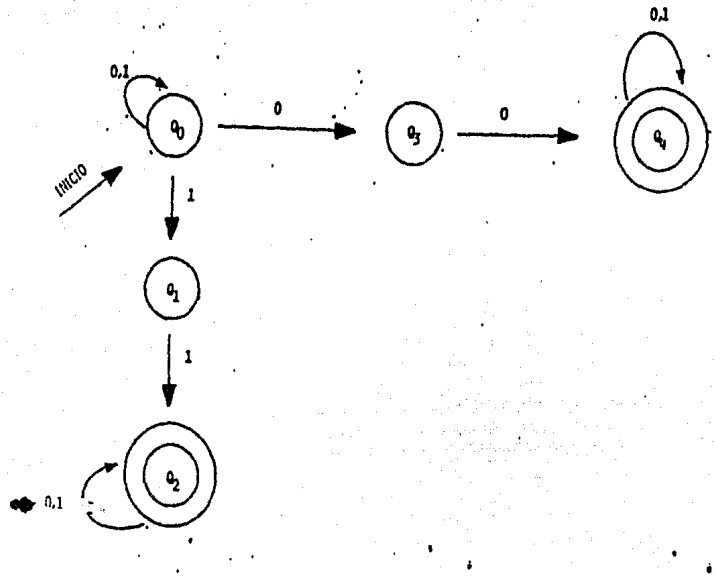

Fig. IV.2
Diagrama de transición de un autómata aceptador de cadenas con dos unos o dos ceros consecutivos.

Se puede observar que al leer el simbolo "0" la máquina puede elegir entre transferirse al estado q_0 o quedarse en el estado q_3 . Pero en lugar de "elegir", se transfiere a ambos, el siguiente caracter que se le introduce está tanto en el estado q_0 como en el estado q_3.

Formalmente un autómata finito no determinístico es una quinteta $M = (K,\Sigma,\delta,q_0,F)$ donde:

K,Σ,δ,q_0 y F tienen el mismo significado que en el autómata finito

AUTOMATAS DE MEALY

Un autómata o máquina de Mealy es aquél en el que se tiene una respuesta por cada transición, donde esta respuesta depende tanto del estado en que se encuentre la máquina como del símbolo que se está leyendo. Esto es, podemos ver la salida como una función, salida = $f(Q,\Sigma)$. De esta forma se tiene que una máquina de Mealy es una sexteta $M = (K,\Sigma,O,\delta,\lambda,q_o)$ donde:

K es el conjunto de estados

Σ es el alfabeto de entrada

O es el alfabeto de salida

δ es la función de transición

λ es la función de respuesta

q_o es el estado inicial

A continuación tenemos una máquina de Mealy que determina, en qué rango cae cualquier número módulo tres.

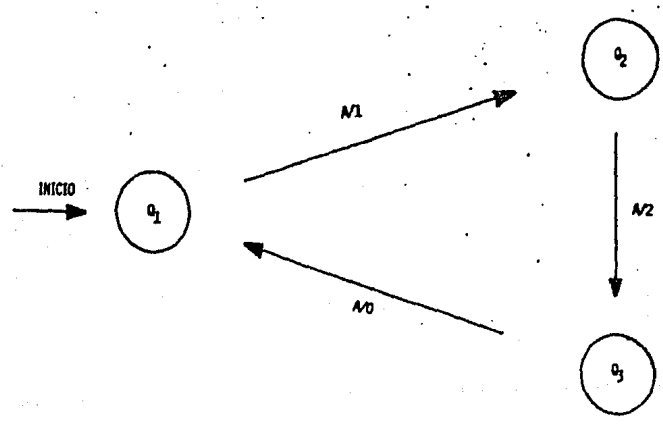

Fig. IV.3

Máquina de Mealy que determina en qué rango cae cualquir número módulo tres

CAPITULO V

EL SUBSISTEMA DE ALTAS DE UN SISTEMA RECUPERADOR DE INFORMACION DOCUMENTARIA

V.1 GENERALIDADES

El subsistema de altas es la parte encargada de analizar el conjunto de fichas que se desea integrar al Banco de Información determinando si la representación del documento se apega a los criterios establecidos, así como detectar, analizar y almacenar la información para que posteriormente pueda ser recuperable.

De acuerdo a lo anterior el subsistema de altas debe ser capaz de detectar el inicio y término de cada ficha, cada una de las palabras contenidas, emitir un diagnóstico y registrar en la base de datos la información correcta.

Para realizar estas operaciones el subsistema de altas se ha diseñado en base a una máquina de Mealy a través del lenguaje Turbo-Pascal y en una microcomputadora personal PC.

El sistema diseñado puede tener diversidad de aplicaciones, sin embargo, el desarrollo del sistema que aquí se describe está enfocado a la aplicación en información legislativa, por lo tanto, a continuación se describe el formato de la ficha específico para esta aplicación.

Como fue expresado en la sección "Representacion del Documento o Ficha", ésta consta de tres zonas. Para la aplicación, la zona de claves está constituida por los siguientes datos:

- Referencia de la ficha.
- Procedencia geográfica [pg].
- Tipo de documento [td]..
- Nombre de la publicacion [np].
- Fecha de publicacion, dada en el formato ddmmyy.
- Campo extraordinario [ce].
- Claves especiales: de significado propio para los usuarios del sistema.
- Fecha de captura del documento.

Donde cada uno de los datos anteriores tienen la siguiente longitud:

DATO	LONGITUD
- Referencia de la ficha.	4
- Procedencia geográfica [pg].	2
- Tipo de documento [td].	2
- Nombre de la publicación [np].	2
- Fecha de publicación, dada en el formato ddmmyy.	6
- Campo extraordinario [ce].	1
- Claves especiales: de significado propio para los usuarios del sistema.	9
- fecha de captura del documento.	2
TOTAL	28

En el anexo A se describen los posibles valores de las claves [U-1].

V.2 EL FORMATO DE LA FICHA

Las zonas de resumen y abstract son consideradas como fue descrito en la misma sección antes indicada. Para el proceso de captación de las fichas el formato es determiando por renglones de 80 caracteres y dado que las fichas pueden ser de longitud variable se tiene que la ficha comprende varios renglones.

En base al formato de la ficha y su captura, nuestra estructura de archivo, fuente de información para el subsistema de altas, está constituida por registros de tipo arreglo con longitud de 82 caracteres. De este arreglo, solo se utilizan los primeros 80 campos, debido a que los lugares 81 y 82 son ocupados por un CARRIAGE RETURN (CR) y un LINE FEED (LF) que la máquina coloca al transmitir los datos.

De esta forma tenemos que una parte del archivo que contiene las fichas se veria asi:

```
7910a831ALDO121179S          QUE CONSISTE LA LEY SIMPSON RODINO */ LA LEY
SIMPSON RODINO ES UNA EXPRESION MAS DEL CARACTER CONFLICTUAL QUE SUELEN
ASUMIR LAS RELACIONES BILATERALES ENTRE MEXICO Y ESTADOS UNIDOS; DICHA LEY
PROHIBE A LOS EMPLEADORES CONTRATAR A EXTRANJEROS INDOCUMENTADOS, LO CUAL INCRE
MENTARA LA CIRCULACION DE DOCUMENTOS FALSOS /*
```

Al concluir la captura de una ficha, el comienzo de la siguiente, debe iniciar el siguiente registro para que se preserve que los primeros 28 caracteres se localicen en el primer registro, por lo tanto el archivo deberia quedar de la siguiente forma:

```
79104831ALD01211795          QUE CONSISTE LA LEY SIMPSON RODINO */ LA LEY
SIMPSON RODINO ES UNA EXPRESION MAS DEL CARACTER CONFLICTUAL QUE SUELEN
ASUMIR LAS RELACIONES BILATERALES ENTRE MEXICO Y ESTADOS UNIDOS; DICHA LEY
PROHIBE A LOS EMPLEADORES CONTRATAR A EXTRANJEROS INDOCUMENTADOS, LO CUAL INCRE
MENTARA LA CIRCULACION DE DOCUMENTOS FALSOS /*
79104831ALOD121184           UTILIDAD ECONOMICA DE LAS CARTAS MAREOGRAFICAS
*/ LA INVESTIGACION MAREOGRAFICA EN MEXICO, INICIADA EN 1946, SE CONCENTRA
EN EL INSTITUTO DE GEOLOGIA DE LA UNAM, DONDE SE ELABORAN CARTAS MAREOGRAFICAS
CON APLICACION EN DIVERSOS CAMPOS DE LA ECONOMIA NACIONAL/*
```

V.3 AUTOMATAS COMO ACEPTORES

Ya que se tienen capturadas las fichas, se procede a analizarlas para determinar si se darán de alta o no.

Para dar de alta una ficha se debe pasar por 4 etapas:

a) Formateo del archivo que contiene la información

b) Verificación de la estructura de la ficha.

c) Verificación de la existencia de las palabras en el léxico.

d) Agregación de la ficha al banco de información

Para llevar a cabo la aceptación de la ficha nos auxiliaremos de una máquina de Mealy guiada por sintaxis, la cual, como ya se vió, permite que la respuesta dependa tanto

del estado en que se encuentra como del símbolo que se está leyendo.

V.3.1 FORMATEO DE LA INFORMACION

Antes de iniciar el análisis para la aceptación o rechazo de la ficha, se procede a transferir la información contenida en el archivo de captura a una representación más compacta, eliminando los espacios superfluos del archivo que contiene las fichas; de la misma forma, los CR y LF generados en la captura son reemplazados cada uno por una "@".

Así, si en el archivo de captura se tiene lo siguiente:

```
87015505RSP0091286          RESOLUCION DE LA COMISION REGIONAL DE LOS
SALARIOS MINIMOS DE LA ZONA 5-D CORRESPONDIENTE A LA COMARCA LAGUNERA.
10 PUNTOS RESOLUTIVOS. PP.1-3. 14-XI-86. ¹/RESOLUCION DE
 LA COMISION REGIONAL DE LOS SALARIOS MINIMOS DE LA ZONA 5
 "D" CORRESPONDIENTE A LA COMARCA LAGUNERA, SALARIOS MINIMOS,
FIJACION/./COAHUILA/SALARIOS MINIMOS/¹
87015605RSP0051286          RESOLUCION DE LA COMISION REGIONAL DE LOS
SALARIOS MINIMOS DE LA ZONA 5-A CORRESPONDIENTE A COAHUILA, SALTILLO.
 3 PUNTOS RESOLUTIVOS. PP.1-3. 10-XI-86. ²/RESOLUCION DE LA
 COMISION REGIONAL DE LOS SALARIOS MINIMOS DE LA ZONA 5 "A"
 CORRESPONDIENTE A COAHUILA, SALTILLO, SALARIOS MINIMOS, FIJACION/./
 COAHUILA/SALARIOS MINIMOS/²
```

Después de las modificaciones el archivo contendrá la siguiente informacion:

```
87015505RSP0091286          RESOLUCION DE LA COMISION REGIONAL DE LOS@@SALARIOS M
INIMOS DE LA ZONA 5-D CORRESPONDIENTE A LA COMARCA LAGUNERA.@@10 PUNTOS RESOLUTI
VOS. PP.1-3. 14-XI-86. ¹/RESOLUCION DE@@ LA COMISION REGIONAL DE LOS SALARIOS MI
NIMOS DE LA ZONA 5@@ "D" CORRESPONDIENTE A LA COMARCA LAGUNERA, SALARIOS MINIMOS
,@@FIJACION/./COAHUILA/SALARIOS MINIMOS/¹@@87015605RSP0051286          RESOLUCION
 DE LA COMISION REGIONAL DE LOS@@SALARIOS MINIMOS DE LA ZONA 5-A CORRESPONDIENTE
 A COAHUILA, SALTILLO.@@ 3 PUNTOS RESOLUTIVOS. PP.1-3. 10-XI-86. ²/RESOLUCION DE
 LA@@ COMISION REGIONAL DE LOS SALARIOS MINIMOS DE LA ZONA 5 "A"@@ CORRESPONDIEN
TE A COAHUILA, SALTILLO, SALARIOS MINIMOS, FIJACION/./@@ COAHUILA/SALARIOS MINIM
OS/²@@
```

Una vez terminada esta tarea se prosigue con el análisis del documento.

V.3.2 VERIFICACION DE LA ESTRUCTURA DE LA FICHA

En la verificacion de la estructura se valida que la ficha haya sido elaborada correctamente, es decir, que no se haya cometido ningún error en la representación del documento.

Asi mismo se van generando y detectando las palabras contenidas en el abstract de la ficha y se generan sus respectivos descriptores. Para determinar si la estructura es correcta o no se deben tomar en cuenta:

1) Los valores de las claves

Para este proceso se verifica que las claves contenidas en el documento sean válidas, es decir, que se apeguen a las claves propias requeridas por el sistema.

Si alguna de estas claves es errónea, (no concuerda con las claves preestablecidas), la ficha es rechazada automáticamente, en caso contrario, se prosigue con el análisis.

2) La estructura global de la ficha

En esta etapa se verifica que la estructura de la ficha sea la adecuada, es decir, que se puedan detectar el inicio y fin del abstract, los párrafos de la ficha, etc. Para lograr este objetivo se utiliza un automata de Mealy.

El alfabeto que acepta el autómata es el siguiente:

- A-Z,"-","#"
- 0-9
- caracteres de puntuación:;,:,",(,) etc
- caracteres especiales: .,+/,@," "

Para referenciar los caracteres utilizaremos la siguiente notación:

α	-	A-Z,"-","#" y de 0-9
ó	-	caracteres de puntuación
*	-	"*"
/	-	"/"
.	-	"." -
T_1	-	"("
T_2	-	")"
b	-	caracter en blanco
@	-	"@"
F	-	fin de archivo
Z	-	cualquier otro caracter no especificado anteriormente

Los paréntesis, nos sirven para ampliar o aclarar el significado de la oración, sin embargo, la información contenida dentro de los paréntesis será ignorada para efectos de recuperación. La máquina de Mealy consta de 13 estados, 14 posibles respuestas y 10 posibles acciones. La respuesta, es el mensaje que manda y las acciones es lo que deberá realizarse. Cabe destacar que no se puede pasar a 2) sin haber tenido éxito en 1). El autómata a utilizar tiene la matriz de transición y el diagrama respectivo que se describen a continuación.

q	α	ó	*	/	.	(	)	b	F	Z	e
0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	2	0	0	0	1	1	1	1
2	3	5	2	2	2	4	2	2	2	2	2
3	3	7	3	8	7	4	3	6	3	3	3
4	11	11	11	11	11	4	12	4	4	4	4
5	3	5	5	8	5	4	5	5	5	5	5
6	3	7	7	8	7	4	6	6	6	6	6
7	3	7	7	8	7	4	7	7	7	7	7
8	3	5	10	2	9	4	8	8	8	8	8
9	3	9	10	2	9	4	9	9	9	9	9
10	10	10	10	10	10	10	10	10	10	10	10
11	11	11	12	11	11	11	11	12	11	11	11
12	6	6	6	8	6	4	12	6	12	12	12

CUADRO V.1

MATRIZ DE TRANSICION DE LA MAQUINA DE MEALY

En caso de que se detecte algún error en la estructura de la ficha, ésta se rechaza, sin embargo, el análisis continuará hasta que se detecten 5 errores o la ficha termine, lo que ocurra primero.

La matriz de respuesta de la máquina de Mealy es la siguiente:

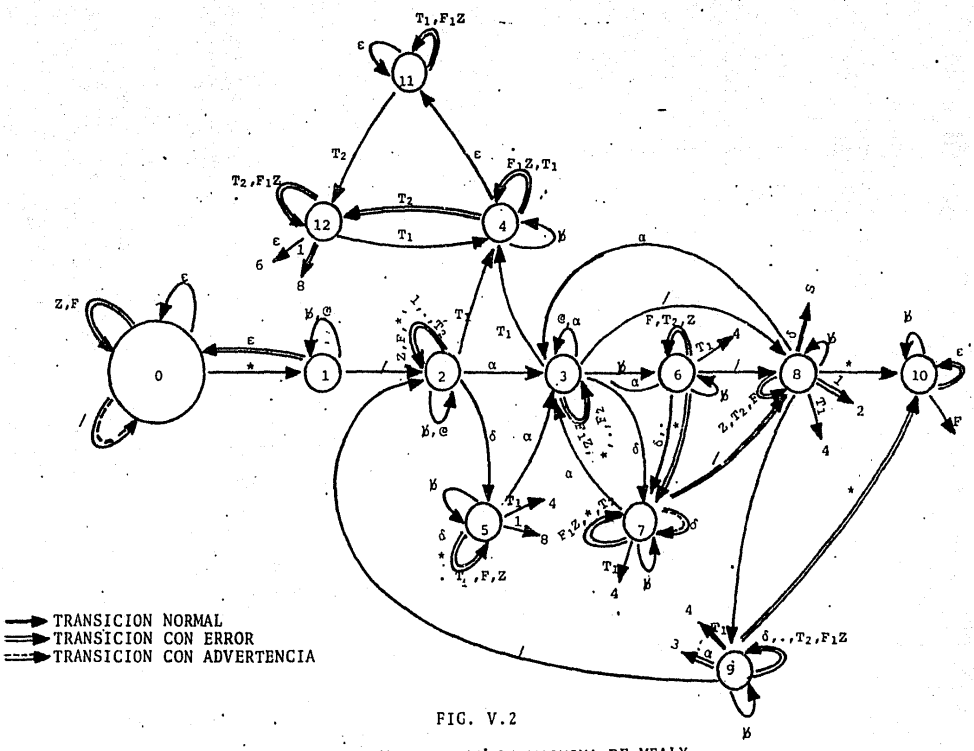

FIG. V.2

DIAGRAMA DE TRANSICION DE LA MAQUINA DE MEALY

q	α	ó	*	/	.	(	)	b	F	Z	e
0	0	0	0	1	0	0	0	0	13	14	0
1	2	2	2	0	2	2	2	0	13	14	0
2	0	3	5	3	5	0	4	0	13	14	0
3	0	0	5	0	0	0	4	0	13	14	0
4	0	0	0	0	0	6	7	0	13	14	0
5	0	8	5	0	5	0	4	0	13	14	0
6	0	0	5	0	0	0	4	0	13	14	0
7	0	8	5	9	5	0	4	0	13	14	0
8	0	3	0	10	0	0	4	0	13	14	0
9	11	11	12	0	5	11	4	0	13	14	0
10	0	0	0	0	0	0	0	0	0	14	0
11	0	0	0	0	0	6	0	0	13	14	0
12	0	0	0	0	0	0	7	0	13	14	0

CUADRO V.3

MATRIZ DE RESPUESTA DE LA MAQUINA DE MEALY

Los posibles errores o respuestas que se pueden detectar son los siguientes:

1.- SE ENCONTRÓ UNA DIAGONAL EN EL RESUMEN.

2.- MARCA DE INICIO DE ABSTRACT INCOMPLETA.

Este error sucede cuando se tiene un "*" seguido de cualquier otro caracter menos de diagonal.

3.- DIAGONAL SEGUIDA DE UN CARACTER DE PUNTUACION.

En el texto se encontró una diagonal y enseguida un caracter de puntuación diferente al punto.

4.- SE ENCONTRÓ UN PARENTESIS DERECHO SIN PRECEDERLE UN PARENTESIS IZQUIERDO.

En el texto aparece una frase del tipo ABROGACION DE LA LEY).

5.- CARACTER DE CONTROL INVALIDO.

Se encontró "*", "." cuando se esperaba cualquier otro caracter.

6.- PARENTESIS IZQUIERDO DUPLICADO.

Se encontró algo de la siguiente forma: PALABRA(PALABRA(PALABRA), o PALABRA((.

7.- PARENTESIS DERECHO INESPERADO.

Se detectaron parentesis sin que mediara texto entre ellos, o se tuvo algo de la forma (PALABRA)).

8.- CARACTER DE PUNTUACION INESPERADO.

Se encontró un caracter de puntuación cuando se esperaba uno de otro tipo.

9.- TERMINÓ LA FRASE CON UN CARACTER DE PUNTUACION.

Se encontró algo de la forma PALABRA,/

10.-FRASE NULA.

Se encontraron dos diagonales sin que mediara texto entre ellas.

11.-MARCA DE FIN DE PARRAFO INCOMPLETA.

Se tiene "/." seguido de cualquier otro caracter menos de "/" .

12.-SE ENCONTRO '*' DESPUES DE PUNTO.

Se leyó: "CARACTER.*" .

13.-FIN DE ARCHIVO INESPERADO.

Se encontró el fin de archivo antes de detectar el fin de ficha

14.-CARACTER NO PERMITIDO.

Se encontró un caracter diferente a los aceptados por el autómata.

En caso de que suceda algún error se manda un mensaje con el número y tipo del error, así como la linea donde se detectó. Es en esta fase, donde se van detectando las palabras contenidas en la zona de abstract de la ficha, asi como en qué párrafo y frase se encuentran, escribiendo a un archivo temporal: "PALABRAS" el correspondiente descriptor temporal de palabra. Este archivo contiene registros de tipo RECORD estructurados de la siguiente forma:

```
Registro_pal : record
                 NW,
                 NF,
                 NPRR,
                 NF: INTEGER;
                 PAL : ARRAY[0..24] OF CHAR;
```

Donde:

NW es el numero de palabra dentro de la frase

NPR es el numero de frase dentro del párrafo

NPRR es el número de párrafo dentro de la ficha

NF es el identificador de ficha

PAL almacena la palabra detectada.

Para Obtener esta información, se tiene la siguiente tabla de acciones:

q	α	ó	*	/	.	(	)	b	F	Z	☻
0	0	0	0	0	0	0	0	0	9	1	0
1	0	0	0	2	0	0	0	1	9	1	0
2	3	0	0	0	0	0	0	1	9	1	0
3	0	4	4	8	4	4	4	4	9	1	0
4	0	0	0	0	0	0	0	0	9	1	0
5	3	0	0	0	0	0	0	0	9	1	0
6	3	0	0	5	0	0	0	1	9	1	0
7	3	0	0	5	0	0	0	0	9	1	0
8	3	0	6	0	0	0	0	1	9	1	0
9	3	0	6	7	0	0	0	1	9	1	0
10	0	0	0	0	0	0	0	1	10	1	0
11	0	0	0	0	0	0	0	0	9	1	0
12	3	0	0	5	0	0	0	0	9	1	0

CUADRO V.4

MATRIZ DE ACCIONES DE LA MAQUINA DE MEALY

Donde cada una de las acciones representa un procedimiento especifico . Las posibles acciones son las siguientes:

0) Emitir el caracter leido
1) Acción nula
2) Inicio de abstract
3) Inicia palabra
4) Termina palabra
5) Termina frase
6) Termina ficha
7) Inicia párrafo
8) Termina frase y termina palabra
9) Terminación inesperada de archivo
10) Terminación de archivo

En caso de tener éxito en esta fase se transmite la ficha al archivo externo, en caso contrario se continúa con el análisis de la siguiente ficha, provocando la pérdida de los descriptores de la ficha anterior.

Una vez que la ficha fue aceptada en esta etapa continuamos con:

V.3.3 VERIFICACION DE LAS PALABRAS EN EL LEXICO

Recordemos que estamos auxiliados de un léxico. Esta parte del sistema toma la información guardada en el archivo PALABRAS, creado en la fase anterior; una vez separadas las palabras en relevantes e irrelevantes, determina si cada una de las primeras existen o no en la estructura LEXICO.

En caso de que alguna(s) palabras(s) no este(n) en el lexico esta(s) palabra(s) se incluye(n) en un archivo de palabras desconocidas. De esta forma, si se desea, en una fase posterior podran darse de alta en LEXICO.

El archivo de palabras desconocidas tiene la misma estructura que el archivo PALABRAS descrito en la sección anterior. Por el contrario si la palabra existe en el léxico, se recupera su descriptor, mandándolo escribir a un archivo temporal.

Este archivo temporal contiene registros del tipo Registro_pal y además el apuntador del registro del léxico correspondiente a la palabra buscada, el cual indica en qué registro se encuentran los descriptores (dentro del archivo LAMBDA) de la palabra en cuestión. El hecho de que exista al menos una palabra desconocida causa el rechazo de la ficha, pero si todas las palabras existen en el léxico, se procede a la agregación de la ficha en el banco de información.

V.3.4 AGREGACIÓN DE LAS FICHAS A LA BASE DE DATOS

Una vez que la ficha ha sido aceptada, ésta debe ser introducida a la base de datos de manera adecuada para poder ser recuperada posteriormente, para esto es necesario que sepamos cuantos registros tiene y que almacenemos en

algún lugar las palabras que contiene el documento, así como los descriptores generados en el proceso previo, de esta forma, la información a agregar será:

1- Los descriptores y

2- La ficha.

V.4 -ESTRUCTURA DE LA BASE DE DATOS

En la base de datos están los documentos que fueron aceptados en la fase descrita en la sección anterior.

La información que debemos de rescatar es la siguiente:

a) texto completo de la ficha

b) formas GAMMA del documento

c) formas LAMBDA del documento

Como mencionamos en el capítulo anterior, esta información será almacenada en tres estructuras diferentes, una para cada elemento.

a) Texto completo de la ficha

El banco de información debe contener la ficha aceptada en forma integra por lo que el documento completo se anexa al archivo que contiene todo el conjunto de las fichas que han sido aceptadas.

El archivo que contiene los documentos aceptados (FICHAS) es un archivo cuyos registros son de tipo ARRAY[0..81] of char.

b) Formas Gamma del documento

Las formas gamma, como ya vimos, son los descriptores de la ficha, dichos descriptores son calculados durante el proceso de agregación. De esta forma es sencillo localizar determinada ficha y leer exactamente su contenido sin el riesgo de leer registros de más o registros de menos. El archivo que contiene estos descriptores (GAMMA), tiene registros del tipo:

```
Reg_tarjeta : record
              Número_ficha,
              pini,
              pfin : integer.
              end;
```

Donde:

Número_ficha	es el identificador de ficha
pini	es la posición inicial
pfin	es la posición final.

c) Formas Lambda

La estructura LAMBDA, como ya se mencionó, contiene bloques que permiten almacenar hasta 50 descriptores de una palabra; cada uno de estos descriptores contiene la siguiente información:

```
Desc_palabra : record
                 nw,
                 nf,
                 nprr,
                 nf : integer;
```

Donde:

nw	número de palabra dentro de la frase
nf	número de frase donde se encontró la palabra
nprr	número de párrafo dentro de la ficha
nf	número de ficha a la que pertenece la palabra.

De esta forma tenemos toda la información necesaria del documento que estamos dando de alta.

V.5- ALGORITMOS DE AGREGACION

Para la inserción del documento aceptado realizamos el siguiente procedimiento:

a) Para insertar el texto completo de la ficha, simplemente nos posicionamos al final del archivo correspondiente y en esta posición anexamos integramente el documento contenido en el archivo externo descrito en la sección V.2.2.

b) Para las formas gamma, una vez generada la información necesaria, ésta se anexa a la estructura GAMMA.

c) Para las formas lambda. Una vez que por medio de la estructura LEXICO conocemos la localidad donde se encuentran los descriptores de dicha palabra, nos posicionamos respectivamente en LAMBDA y procedemos a agregar los descriptores requeridos.

Para realizar la agregación, se hará en orden creciente de acuerdo al identificador de ficha al que pertenecen las palabra, esto es, si por ejemplo tenemos 4 descriptores de la palabra SOCIAL, y éstos pertenecen a las fichas 1,4,2,5, estos deberán ser dados de alta de modo tal que primero se quede el descriptor correspondiente a la ficha 1, enseguida el de la 2a., después el de la 4a. y por último el de la 5a. ficha.

Pero ¿qué sucede si tengo saturado un bloque de descriptores y necesito intercalar otro? Previendo esta situación, al agregar los descriptores, solo se utilizarán los primeros 47, dejando los tres restantes para cuando se presente la situación anterior.

De esta forma, si tenemos que intercalar algún descriptor, lo que procede es realizar un corrimiento hacia los últimos 3 campos y así realizar la intercalación.

Por otra parte, si nuestro bloque no está saturado, lo que procede es realizar un corrimiento hacia el siguiente elemento del grupo.

Pero, ¿qué sucede si los último tres campos ya han sido utilizados y nuevamente es necesario hacer una intercalación? En este caso lo que procede es realizar un corrimiento hacia los campos de reserva vacios más próximos. Asi, tenemos los descriptores ordenados en orden creciente de acuerdo al número de ficha al que pertenecen las palabras, agilizando de esta forma la búsqueda.

Para ejemplificar, véase la base de datos en la fig. V.5 .

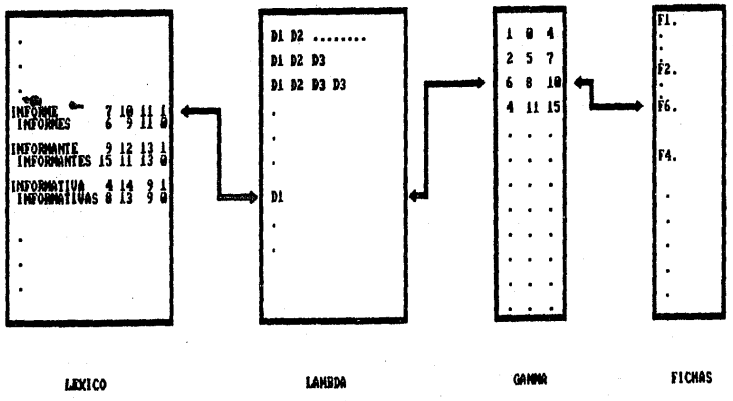

Fig. V.5

ESTRUCTURA DE LA BASE DE DATOS

CONCLUSIONES

En el desarrollo de un modelo de recuperación de información documentaria para microcomputadora se plantean muchas tareas, el sistema aqui presentado se encarga de realizar algunas de ellas.

El sistema lleva a cabo las tareas de filtrado y validación de la información sobre los documentos que son capturados para que posteriormente, formen parte del acervo, además prepara los elementos necesarios para la fase de recuperación.

Dicho sistema puede ser modificado de tal forma que se agilice su funcionamiento como por ejemplo, en que permita capturas en línea y cuente con un thesaurus; por otra parte, se cuenta con las características siguientes:

TRANSPORTABILIDAD, ya que fue desarrollado en un lenguaje de alto nivel que se encuentra disponible en cualquier microcomputadora moderna. Debido al auge que han tomado las computadoras personales, el implementar el sistema en un equipo de este tipo tiene la ventaja de ponerlo al alcance de un mayor número de personas.

SENCILLEZ DE OPERACION, lo cual redunda en que cualquier persona sin necesidad de una preparación o experiencia en sistemas basados en computadoras, lo pueda usar.

VERSATILIDAD, ya que mediante unos pequeños cambios es factible alterar la longitud de la zona de campos fijos o claves permitiendo asi adecuarlo a otras necesidades del usuario, por ejemplo para llevar control del acervo de una biblioteca, de expedientes médicos, etc., permitiendo asi su uso a un número mayor de aplicaciones.

Con el trabajo aqui presentado se ha dado el primer paso para el modelo general de un SRID, quedando pendiente el sistema de consulta.

ANEXO A

CATALOGO DE CLAVES

PROCEDENCIA GEOGRAFICA: a cada estado le corresponde un número del 01 al 32, el cual se le asigna una vez que los estados han sido ordenados alfabéticamente, asi a Aguascalientes le corresponde el 01 y a Zacatecas el 32.

La única excepción es el distrito federal al cual le corresponde el número 00.

NOMBRE DE LA PUBLICACION:

Diario Oficial	DO
Periódico Oficial	PO
Boletin Oficial	BO
Gaceta de Gobierno	GG
Gaceta Oficial	GO

CAMPO EXTRAORDINARIO:

Alcance	A
Número Extraordinario	E
Suplemento	S
Primera Sección	P
Segunda Sección	G
Tercera Sección	T
Cuarta Sección	C
Quinta Sección	Q
Sexta sección y siguientes	X
Anexo	N

TIPOS DE DOCUMENTOS:

Aclaración	AL
Actas	AT
Acuerdo	AC
Autorización	AU
Aviso	AV
Bando de Política	BP
Bases	BA
Circular	CI
Código	CD
Consecion	CS
Constitución	CO
Contrato-Ley	CL
Convenio	CV
Convocatoria	DC
Decreto	DE
Decreto-Ley	DL
Disposición General	DG
Estatutos	ES
Fe de Erratas	FE
Informe	IN
Instructivo	IS
Ley	LE
Ley Orgánica	LO
Ley Reglamentaria	LR
Lista	LI
Manual	MA
Norma Oficial	NO
Oficio	OF
Oficio Circular	OC
Ordenanza	OR
Plan	PL
Presupuesto	PS
Programa	PO
Prontuario	PR
Regla	RL
Regla General	RG
Reglamento	RE
Regulación	RA
Resolución	RS
Resúmen	RU
Tabla	TB
Tarifa	TA
Tratado	TI

BIBLIOGRAFIA

[Q-1]- Diccionario Enciclopédico Quillet
Tomo VIII.

[S-2]- Gran diccionario enciclopédico ilustrado de
Selecciones del Rider's Digest
Tomo VII.

[S-3]- Idem.

[JB-4]-Sistemas de información:teoria y práctica
Bursch, John G. et al.

[JB-5]-Information Systems and Networks
Op. Cit.

[JB-6]-Sistemas de Información: teoria y práctica
Op. Cit.

[JB-7]-Sistemas de Información: teoria y práctica
Op. Cit.

[D-8]- Information Procesing Systems
Davis, William S.

[SA-9]-Informática, presente y futuro
Sanders.

[SA-10]-Information Procesing Systems
Op. cit.

[JB-11]-Sistemas de Información: teoria y práctica
Op. Cit.

[JB-12]-Information Systems and Networks
Op. Cit.

[JB-13]-Sistemas de Información: teoria y practica
Op. Cit.

[VR-14]-Information Retrieval
C. J. van Rijsvergen.

[L-1]- Lancaster F. W.
Information Retrieval Systems:Characteristics,
Testing and evaluation

[ER-2]- Elaine Rich.
Artificial Intelligence.

[M-1]- Para mayor referencia remitirse al manual:
"Microsoft MS-DOS, Operating System
User's guide"
Microsoft Corporation.

[M-2]- Remitirse al manual:
"Turbo-Pascal versión 3.0.

[VE-1]- Elisa Viso Gurovich
Autómatas y lenguajes formales
Facultad de Ciencias.

[U-1]- El sistema UNAM-JURE un banco de datos
legislativo
Instituto de Investigaciones Juridicas
Dirección General de Servicios de Cómputo para
la Administración.