UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Unidad Académica de los Ciclos Profesional y de Posgrado del C.C.H.

Sofía

Un Sistema de Recuperación de Información por indexación de triadas

T E S I S

PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACION

P R E S E N T A :

GERARDO VEGA HERNANDEZ

DIRECTOR DE TESIS: DR. PABLO BARRERA SANCHEZ

MEXICO, D. F. SEPTIEMBRE 1994

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

Resumen de Tesis

"Sofía un Sistema de Recuperación de Información por Indexación de triadas"

Actualmente, la mayoría de los Sistemas de Recuperación de Información (SRI) que realizan indexación automática utilizan las palabras de un documento para diferenciar éste del resto de la colección de documentos. Esto trae como consecuencia la necesidad de idear complejos algoritmos que sean capaces de "comprender" la morfología del idioma en que esté escrita la colección de documentos. El presente trabajo de tesis tiene como objetivo el mostrar que es posible generar por cada documento un cierto patrón semejante a un patrón de voz y utilizar este patrón para la caracterización del documento dentro de la colección. En otras palabras, se muestra que es posible el desarrollo de un SRI que realice la recuperación de documentos basado en la cuantificación del parecido de las palabras que conforman cada documento en la colección, con las palabras utilizadas para la búsqueda. De esta manera y, debido a que un SRI con estas características no utiliza directamente las palabras para la búsqueda, en muchos casos tal sistema será capaz de recuperar documentos correctamente aún cuando las palabras empleadas en la búsqueda estén mal escritas. Esto último es de gran importancia, pues de esta forma se está terminando con la fuerte dependencia que algunos de los SRI tienen de la morfológia del idioma en que estén escritos los documentos de la colección.

Otra ventaja de un SRI con estas características es el ahorro en espacio de almacenamiento que se logra, pues se ha comprobado que el archivo que resulta del proceso de indexar toda la colección de documentos, puede ser comprimido considerablemente, de tal manera que ocupe entre un 10% y 50% del espacio requerido por toda la colección de documentos. A este respecto se ha podido también comprobar que aún cuando el archivo de indexación se encuentre comprimido, esto no impacta fuertemente, en forma negativa, en el tiempo de respuesta al momento de las búsquedas. Esto último, se hace la aclaración, es logrado gracias a los algoritmos tanto de indexación como de búsqueda, que son descritos en el trabajo de tesis.

El trabajo de tesis está organizado en dos capítulos. El primer capítulo está dedicado a los antecedentes de los SRI, en donde se darán las definiciones, así como las bases de los SRI. El segundo capítulo está dedicado a exponer las ideas y algoritmos en que se sustenta el desarrollo de un SRI por indexación de triadas.

Es importante hacer resaltar que el presente trabajo es totalmente original pues no está basado en ningún artículo extranjero o nacional. Este trabajo es el resultado de una plática informal entre el **Dr. Victor Guerra O.**[1], el **Dr. Enrique Daltabuit G.**[2] y el **Mat. Gerardo Vega H.**. En dicha plática se discutió la posibilidad de que el proceso de indexación de un SRI, en lugar de emplear a las palabras para la creación de las representaciones de los documentos utilizara fragmentos contiguos de tres letras (triadas) que estuvieran contenidos en cada una de estas palabras. Después de esta plática siguió un año de investigación en el que se logró establecer la metodología para el desarrollo de un Sistema de Recuperación de Información que hemos calificado "SRI por indexación de triadas".

A continuación se presenta una lista de referencias bibliográficas de los documentos en que está basado principalmente el primer capítulo de esta tesis. Las referencias para el segundo capítulo no se presentan, como ya se hizo mención buena parte de este trabajo es original.

[1] **Information Retrieval.**
Second Edition.
C.J. van Rijsbergen.
University of Cambridge.
Butterworth & Co (Publishers) Ltd., 1979.
London.

[2] **Introduction to modern information retrieval.**
Gerald Salton.

[3] **Coding and information theory.**
R. W. Hamming & Englewood Cliffs, N. J.
Prentice - Hall, 1980.

Vo. Bo.

Mat. Gerardo Vega Hernández

Dr. Pablo Barrera Sánchez
Profesor de tiempo completo
de la Facultad de Ciencias
de la UNAM.

México D.F. a 29 de Agosto de 1994.

1 Director General de la Dirección General de Servicios de Cómputo Académico de la UNAM.
2 Director de la Dirección de Cómputo para la Investigación de la Dirección General de Servicios de Cómputo Académico de la UNAM.

Sofía

Un Sistema de Recuperación de Información por indexación de triadas

A mi familia con
todo mi cariño:
Sara, **Sofi** y **Arte**

Agradezco profundamente al **Dr. Pablo Barrera S.** su apoyo y consejos, los que he sentido que vienen más de un amigo que de un profesor o director de tesis.

Agradezco en lo general a la **DGSCA** y muy en especial a sus directores por su apoyo, el cual fué fundamental para el desarrollo de este trabajo.

Agradezco muy particularmente al **Dr. Enrique Daltabuit G.** sus siempre acertados comentarios y observaciones, así como también su decidido apoyo para la realización de este trabajo.

Indice

Introducción

Actualmente, la mayoría de los Sistemas de Recuperación de Información (SRI) que realizan indexación automática utilizan las palabras de un documento para diferenciar éste del resto de la colección de documentos. Esto trae como consecuencia la necesidad de idear complejos algoritmos que sean capaces de "comprender" la morfología del idioma en que esté escrita la colección de documentos. El presente trabajo de investigación tiene como objetivo el mostrar que es posible generar por cada documento un cierto patrón semejante a un patrón de voz y utilizar este patrón para la caracterización del documento dentro de la colección. En otras palabras, se mostrará que es posible el desarrollo de un SRI que realice la recuperación de documentos basado en la cuantificación del parecido de las palabras que conforman cada documento en la colección, con las palabras utilizadas para la búsqueda. De esta manera y, debido a que un SRI con estas características no utiliza directamente las palabras para la búsqueda, en muchos casos tal sistema será capaz de recuperar documentos correctamente aún cuando las palabras empleadas en la búsqueda estén mal escritas. Esto último es de gran importancia, pues de esta forma se está terminando con la fuerte dependencia que los SRI tienen de la morfológia del idioma en que estén escritos los documentos de la colección.

El presente trabajo de tesis está organizado en dos capítulos. El primer capítulo está dedicado a los antecedentes de los SRI, en donde se darán las definiciones, así como las bases de los SRI. El segundo capítulo está dedicado a exponer las ideas y algoritmos en que se sustenta el desarrollo de un SRI por indexación de triadas.

Es importante hacer resaltar que el presente trabajo es totalmente original pues no está basado en ningún artículo extranjero o nacional. Este trabajo es el resultado de una plática informal entre el **Dr. Víctor Guerra O.**[1], el **Dr. Enrique Daltabuit G.**[2] y el **Mat. Gerardo Vega H.**. En dicha plática se discutió la posibilidad de que el proceso de indexación de un SRI, en lugar de emplear a las palabras para la creación de las representaciones de los documentos utilizara fragmentos contiguos de tres letras (triadas) que estuvieran contenidos en cada una de estas palabras. Después de esta plática siguió un año de investigación en el que se logró establecer la metodología para el desarrollo de un Sistema de Recuperación de Información que hemos calificado "SRI por indexación de triadas".

1 Director General de la Dirección General de Servicios de Cómputo Académico de la UNAM.
2 Director de la Dirección de Cómputo para la Investigación de la Dirección General de Servicios de Cómputo Académico de la UNAM.

1. Capítulo I Antecedentes

Desde los inicios de la historia humana y con la invención de la escritura, el hombre fue creando documentos (tablillas, papiros, etc.), como una fuente de información y como un legado para las generaciones futuras, en los que transcribía los conocimientos, hechos e inquietudes de su época. En los primeros tiempos el total de documentos disponibles cambiaba relativamente poco. Sin embargo, durante el siglo XIX el monto de publicaciones científicas se fué duplicando cada 50 años. Más recientemente con el impresionante crecimiento de la ciencia y la tecnología, el promedio de crecimiento de documentos disponibles se ha incrementado considerablemente. En la actualidad, aparentemente no existe un límite en el promedio de crecimiento de los documentos que están disponibles. Por esta razón, en nuestros días ha sido necesario contar con mecanismos automáticos o semiautomáticos que ayuden con el manejo de estos vastos cúmulos de información.

Comunmente aquellos sistemas que tratan con la representación, almacenamiento, organización y acceso a la información disponible a través de una computadora, se conocen con el nombre de Sistemas de Recuperación de Información (SRI).

Para que los Sistemas de Recuperación de Información puedan manipular en forma indirecta a la colección de documentos es necesario crear una representación especial de esta colección. Se han ideado técnicas manuales y automáticas para obtener tal representación. Una de las maneras más comunes para obtener la representación de un documento es mediante una lista de términos o palabras clave extraídas a partir del documento, que describan lo mejor posible el contenido de dicho documento, además de que contribuyan a diferenciar éste del resto de la colección.

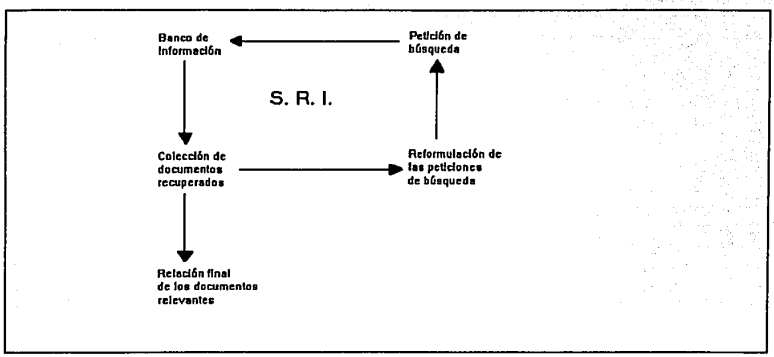
1.1. Los Sistemas de Recuperación de Información.

A fin de tener una idea más clara de lo que es un SRI, es útil distinguir éstos de los manejadores de base de datos. Aunque ambos sistemas implican la representación, almacenamiento, organización, mantenimiento y recuperación de la información disponible, tienen diferencias notables. En un manejador de base de datos la información es almacenada en **forma de registros** (los que en conjunto forman lo que comunmente se denomina como base de datos) los cuales a su vez se encuentran separados por diversos campos, por otro lado en un SRI la información almacenada está constituída por la **representación** de cada uno de los documentos o extractos de documentos (que forman lo que comunmente se conoce como el banco de información), la representación de un documento podría, por ejemplo, ser una lista de palabras extraídas del texto del documento original, consideradas como significativas para representar el contenido del documento. La información recuperada a través de un manejador de bases de datos consistirá de todos aquellos registros los cuales cumplan exactamente con una condición de búsqueda, mientras que en un SRI la condición de búsqueda puede cumplirse sólo parcialmente. Con respecto al tipo de inferencia usado para relacionar, los registros en el caso de un manejador de base de datos, y las representaciones de los documentos en el caso de un SRI, los manejadores de bases de datos utilizan un tipo de inferencia deductiva, es decir, por ejemplo, si el registro **A** está relacionado con el registro **B** (**A Rel. B**) y el registro **B** está relacionado con el registro **C** (**B Rel. C**), entonces el registro **A** está relacionado con el registro **C** (**A Rel. C**). En un SRI es más común utilizar un tipo de inferencia inductiva, donde las relaciones entre las representaciones de los documentos, son solamente establecidas con un cierto grado de certidumbre o incertidumbre, de aquí que nuestra confianza en la inferencia es variable. Lo anterior nos permite describir a un manejador de base de datos por medio de un modelo determinístico, mientras que en un SRI un modelo probabilístico sería más apropiado. También existen diferencias en el tipo de lenguaje que se utiliza para la formulación de las peticiones, ya que el manejador de bases de datos utiliza un lenguaje artificial, donde la sintaxis y el vocabulario están restringidos, mientras que los sistemas de recuperación de información pueden utilizar un lenguaje más amplio, como el lenguaje natural.

1.2. Componentes y funcionamiento básico de un SRI.

Cualquier Sistema de Recuperación de Información puede ser descrito como una colección de documentos por un lado, y como una serie de preguntas o peticiones relacionadas con la colección de documentos por el otro. Siendo así, es necesario contar con un mecanismo intermedio para determinar cuáles documentos del conjunto de la colección, cumplen con los requerimientos de las peticiones dadas. En la práctica la determinación de los documentos que son relevantes al conjunto de peticiones, no es efectuada en forma directa. Más precisamente cada uno de los documentos, así como también cada una de las peticiones, son mapeados a una representación especial, la cual se encontrará dentro de lo que se denomina **lenguaje de indexación**. Al mapeo de cada documento al lenguaje de indexación se le conoce como **proceso de indexación**, así como al mapeo de cada petición al lenguaje de indexación se le conoce como proceso de **formulación de petición de búsqueda**. Los procesos para determinar cuáles documentos de la colección son relevantes serán llevados a cabo utilizando las representaciones de las peticiones y de los documentos dentro del lenguaje de indexación. Un operador de relación será el encargado de llevar a cabo la función de recuperación mediante la cuantificación de la similaridad entre las representaciones dentro del lenguaje de indexación de los documentos y las peticiones, de esta forma determina cuáles de los documentos podrían ser relevantes a la petición de búsqueda y como consecuencia ser considerados para su consulta.

Es importante mencionar que la mayoria de los sistemas de recuperación de información permiten la utilización de la información obtenida, como resultado de una petición de búsqueda, para reformular las peticiones de búsqueda precedentes, que permitan mejorar la relación final de los documentos relevantes presentados por el sistema. Al proceso de reformulación de peticiones a partir de la información obtenida de peticiones precedentes se le conoce con el nombre de **retroalimentación** ("*feedback*").



Dentro de los trabajos efectuados en el área de recuperación de información podríamos distinguir tres líneas principales de investigación, que son: **análisis de contenido, estructuras de información** y **evaluación**. El primer término se relaciona con la descripción del contenido de los documentos en una forma adecuada para el procesamiento por la computadora (lenguajes de indexación). El segundo explota las relaciones entre los documentos para mejorar la eficiencia y efectividad de las estrategias de recuperación. Por otro lado, la evaluación da elementos para cuantificar la eficiencia y efectividad de la recuperación. A continuación se dará una explicación más detallada de estas tres líneas de investigación.

1.3. Análisis de contenido.

De todos los procesos requeridos en la recuperación de información el más crucial y problablemente el más difícil es el que implica el estabecimiento de un lenguaje de indexación, es decir, asignar, mediante el uso de reglas establecidas, términos apropiados e identificadores capaces de representar el contenido de la colección de documentos y las peticiones de búsqueda relacionadas con esta colección.

Cualquiera que sea el tipo de lenguaje de indexación utilizado, se asume que un documento deberá ser representado por una lista de elementos del lenguaje de indexación, por ejemplo un documento titulado "*Estudio sobre el tratamiento biológico de aguas residuales*", podría ser representado utilizando los términos o palabras clave, "**aguas residuales**","**tratamiento**" y "**biológico**".

El vocabulario que se emplea en el lenguaje de indexación, puede ser **controlado** o **no controlado**. Un vocabulario de indexación no controlado en principio puede incluir una variedad amplia de términos, en lenguaje natural, dándole así una mayor flexibilidad, sin embargo, podrían también presentarse casos de ambigüedad y error. Por lo tanto una restricción sobre el lenguaje de indexación con frecuencia favorece que los términos disponibles para la identificación del contenido sean rígidamente controlados, esto permite la eliminación de sinónimos ya que para cada clase de palabras sinónimas (*thesaurus*) se escoge un término único. Algunas veces una combinación de elementos de un lenguaje controlado y uno no controlado pueden combinarse para formar un lenguaje de indexación.

Los lenguajes de indexación también pueden ser descritos como **pre-coordinados** o **post-coordinados**. El primero indica que los términos son coordinados al momento de efectuarse la indexación y el último al momento de la búsqueda. Más específicamente cuando términos compuestos son utilizados para propósitos de indexación, consistentes en frases que posiblemente incluyan sustantivos, adjetivos, preposiciones y una variedad de indicadores de relación, el proceso es llamado pre-coordinado. Por ejemplo un documento titulado "*Estudio sobre el tratamiento biológico de aguas residuales*" puede ser indexado, de manera precoordinada, utilizando las siguientes frases "**tratamiento biológico**" y "**aguas residuales**". Por otro lado en proceso post-coordinado, el mismo documento sería identificado al momento de la búsqueda por combinación de los términos indexados individualmente; "**aguas**" y "**residuales**" y "**tratamiento**" y "**biológico**".

Otra distinción acerca de los lenguajes de indexación es la que se refiere a la amplitud o precisión con que éstos cubren los temas o áreas que trata la colección de documentos. Tomando en cuenta esto se tiene que entre los parámetros que presentan especial importancia a este respecto se encuentran la **exhaustividad** y la **especificidad** del lenguaje de indexación. Un lenguaje de indexación exhaustivo contiene términos que cubren todas las posibles áreas mencionadas en la colección de documentos; correspondientemente el resultado de una indexación exhaustiva implica que todas las áreas son apropiadamente reflejadas mediante los términos indexados asignados a los documentos. Por otro lado un lenguaje de indexación específico nunca cubre distintas áreas utilizando un simple término, más bien restringe la utilización de los términos empleados, usando para la representación de la colección de documentos, sólo aquellos términos que sean muy precisos.

Existen dos maneras de llevar a cabo la indexación, una en forma manual y la otra en forma automática. Ambas tienen como fin extraer del texto del documento fuente (este texto podría ser un extracto del propio documento o quizá el título del documento) un documento representativo formado por las palabras clave, mismas que deberán ser elementos del lenguaje de indexación. La principal diferencia que puede hacerse

entre estos dos tipos de indexación es que el análisis de contenido en la indexación manual es realizado como su nombre lo indica, en forma manual, teniendo por esto que ser realizado por personal capacitado. Por otro lado, cuando la asignación de identificadores de contenido es realizada con ayuda de equipo de cómputo, la operación es llamada indexación automática.

Actualmente la indexación manual es la regla antes que la excepción en los medios ambientes operacionales, debido esto principalmente a una costumbre más que a una actitud reacia al cambio. En este tipo de indexación existe una serie de herramientas que permiten al indexador controlar el proceso de indexación, como son listas de terminología, manuales de instrucción y principalmente hojas de trabajo estructuradas para registrar los productos de indexación. También suelen utilizarse notas que permitan definir el significado y la interpretación de cada uno de los términos indexados permitidos en un documento dado.

Hay evidencias que indican que los métodos de indexación automáticos simples son rápidos y económicos y permiten obtener resultados al menos equivalentes a los obtenidos cuando la indexación es efectuada manualmente.

A continuación se darán las generalidades involucradas en el proceso de indexación automática las cuales están basadas en las ideas de **Luhn** [1].

1.3.1. Indexación automática.

Luhn utiliza la frecuencia de aparición de las palabras en el texto de un documento para determinar cuáles de aquellas palabras son suficientemente significativas para representar o caracterizar el documento en la computadora, a estas palabras se les denomina como **palabras clave**. A partir de cada documento se obtendrá una lista de palabras clave que representarán al documento. Además la frecuencia de ocurrencia de las palabras en el texto puede ser utilizada para determinar el grado de significancia de cada palabra. A partir de ésto se obtiene un esquema de ponderación de las palabras clave en cada lista. **H.P. Luhn** [1], uno de los pioneros de la indexación automática establece que:

> **"La justificación de medir la significancia de una palabra usando su frecuencia está basada en el hecho de que un escritor normalmente repite ciertas palabras tanto como él lo requiera o de acuerdo como su argumento varíe y como él elabore un aspecto de un tema. Este significado de énfasis es tomado como un indicador de significancia..."**

De hecho **Luhn** observó que cuando se ordenaban las palabras contenidas en un texto en forma decreciente con respecto a su frecuencia de ocurrencia dentro del texto (las palabras de mayor frecuencia primero), y graficaba la frecuencia de ocurrencia de las palabras con respecto al lugar que ocupaban dentro de la lista ordenada por frecuencia (orden) , entonces se obtenía una curva hiperbólica, es decir, la frecuencia y el orden de la frecuencia de la palabras describían la siguiente relación:

FRECUENCIA X ORDEN ~ CONSTANTE

A continuación se da un ejemplo de este comportamiento, considerando las palabras de mayor frecuencia obtenidas a partir del texto del estatuto del personal académico de la UNAM.

(Número de palabras en el documento N = 10753)

Orden (O)	Término	Frecuencia	O*(F/10753)
1	el	1194	0.111
2	de	920	0.171
3	en	284	0.079
4	que	253	0.094
5	y	246	0.114
6	o	221	0.123
7	su y sus	171	0.111
8	artículo	151	0.112
9	del	140	0.117
10	se	138	0.128

Tabla 1 .Ilustración del comportamiento de la ley frecuencia-orden.

Dado que ni las palabras con alta frecuencia, ni las palabras con baja frecuencia son buenos identificadores de contenido, **Luhn** conjeturó que las palabras con "mayor poder de resolución" para representar el texto de los documentos eran precisamente aquellas palabras con frecuencia media.

Lo anterior puede ser entendido con mayor claridad si se observa la figura 1, en la cual se ha graficado la frecuencia de ocurrencia de las palabras contra su orden, que como ya se dijo da origen a una curva hiperbólica. Además se observan dos cotas, las cuales determinan el poder de resolución o grado de significancia de las palabras, mismas que deben encontrarse entre estos dos límites, ya que se considera que las palabras que se encuentren antes del primer límite o límite superior son comunes y por lo tanto no contribuyen significativamente a la representación del documento. Por otro lado, las palabras que se encuentran después del segundo límite o límite inferior, se consideran raras y por ende también poco significativas para la representación de

los documentos. Es importante aclarar que el establecimiento de los límites se lleva a cabo de manera un tanto arbitraria, ya que por lo general se determinan en base a prueba y error.

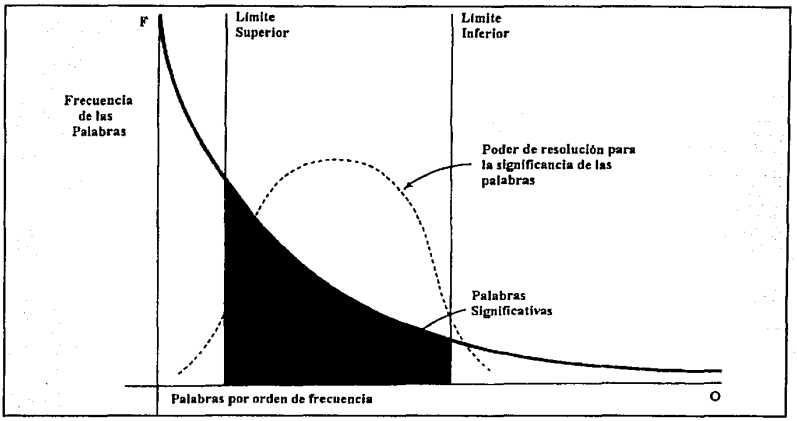


Fig. 1 Gráfica frecuecia & orden de ocurrencia de palabras, Se muestra además las cuotas que determinan las palabras significativas. Figura extraída de la referencia [1].

La representación adecuada de un documento para un sistema de recuperación automático podría estar dada simplemente como una lista de términos, donde cada término representa una clase de palabras las cuales suceden dentro del texto de la colección de documentos. Un documento será indexado por un término dado si una de las palabras que forman el texto del documento se encuentra como miembro de esta clase.

Para la obtención de la representación de un documento se pueden seguir los siguientes 3 pasos:

1.- Eliminación de palabras con alta frecuencia de aparición, es decir eliminación de las palabras nulas.

2.- Remoción de sufijos y prefijos de las palabras no nulas. El conjunto de raíces de palabras que resulte de este paso conformará la lista de términos que servirá como representació del documento para el SRI.

La eliminación de palabras con alta frecuencia de aparición podría efectuarse si se establece la cota superior en la figura 1. Una manera de ubicar esta cota podría hacerse utilizando una lista de palabras nulas (por ejemplo: la, el, los, de, etc.) y comparando ésta con el texto de los documentos, y así, el valor de la cota superior estaría dado por el orden por abajo de cuyo valor se encuentran los órdenes de las palabras nulas.

La remoción de sufijos y prefijos es un proceso más complicado. Una manera de realizar esta remoción es construyendo una lista completa de sufijos y prefijos y utilizar esta lista para obtener las raíces de las palabras vía la eliminación de sus sufijos y prefijos. Desafortunadamente la remoción de sufijos o prefijos sin tomar en cuenta algunas de las reglas preestablecidas, podría llevar a errores significativos, por ejemplo: en la palabra **PERMITAN** bien se podría remover el sufijo **AN**, sin embargo esto no sería conveniente en el caso de la palabra **PLAN**. Para evitar errores de remoción de sufijos o prefijos, se pueden idear reglas que permitan la remoción sólo si el contexto es adecuado. Otro problema relacionado con la obtención de las raíces de las palabras es que aún cuando en general la remoción de sufijos de palabras morfológicamente equivalentes se obtienen raíces iguales, hay excepciones, por ejemplo las palabras **CONVOQUEN** y **CONVOCAR**, que expresan un concepto similar, tienen raíces **CONVOQU** y **CONVOC**, repectivamente. De esta forma, para determinar que estas dos raíces se refieren a un mismo concepto se deberán identificar los finales de raíces **QU** y **C** como equinvalentes.

Un procedimiento alternativo para la remoción de sufijos es truncar las palabras de tal manera que solo se conserven las primeras **N** letras de cada palabra (**N** podría ser por ejemplo igual a 6). Este procedimiento además de ser simple y fácil de implementar tiene la ventaja de ser independiente del idioma alfabético en que esté escrita la colección de documentos. Sin embargo, este procedimiento no resuelve por un lado el problema de la remoción de prefijos y por otro lado la lista de términos (lenguaje de indexación) encontrados con este procedimiento no necesariamente es la más exacta para representar a cada documento dentro de la colección.

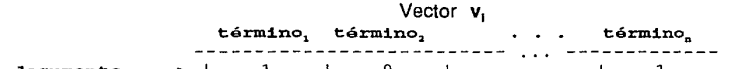
1.4. Estructuras de información.

Esta sección pretenderá hacer una revisión general acerca de las técnicas más comunes relacionadas con la organización de archivos, estrategias de búsqueda y metodologías para la recuperación de información.

Una vez que se han obtenido las representaciones de los documentos que forman el banco de información, éstos deberán almacenarse bajo ciertos criterios de organización: la organización de los datos se hará en forma lógica y física. La organización de tipo físico tiene que ver con cuestiones inherentes al medio físico que se usará para almacenar estos datos. Por otro lado, la organización lógica se refiere a las relaciones que deberán existir entre cada uno de los elementos que conforman el banco de información, independientemente de la manera en la cual estas relaciones puedan establecerse dentro de cualquier computadora. El establecimiento de las relaciones que deben existir entre los elementos del banco de información es de gran importancia, ya que de esto depende que dada una petición de búsqueda al Sistema de Recuperación de Información, se obtenga de éste una recuperación óptima. Por esta razón, en nuestro caso, es fundamental la organización lógica de los datos.

El término estructuras de información cubre específicamente la organización lógica de la información formada por las representaciones de los documentos, con el fin de que posteriormente se pueda efectuar su recuperación.

Como se mencionó en la sección anterior, un documento puede ser representado por una lista de elementos del lenguaje de indexación. Esta lista puede ser expresada por medio de un vector binario, cuya dimensión deberá ser igual al número de elementos del lenguaje de indexación. De esta manera, si se establece una biyección entre los elementos (términos) del lenguaje de indexación y las entradas de su vector binario, es posible relacionar a cada **documento**$_i$ de la colección con un vector $\mathbf{v}_i$ donde la entrada **k** del vector $\mathbf{v}_i$ será igual a uno, si aparece dentro del texto del **documento**$_i$ el término del lenguaje de indexación correspondiente a la entrada **k** del vector $\mathbf{v}_i$, y cero de lo contrario.



Forma en la cual se asocia a cada documento un vector binario (donde **n** es el número de elementos del lenguaje de indexación).

Utilizando cada uno de los vectores binarios asociados a cada documento de la colección es posible construir otro tipo de configuración la cual estaria dada por una matriz binaria en donde cada renglón repreresenta un documento que fué indexado con los términos correspondientes a las entradas cuyos valores sean iguales a uno.

Las diferencias entre estos dos tipos de configuración para la representación de documentos repercuten ampliamente en el proceso de la recuperación de información. Cuando uno desea recuperar documentos que incluyan algunos de los términos

utilizados para la indexación de la colección de documentos, bastará con acceder al archivo invertido (matriz binaria) el cual muestra el conjunto de documentos que responde a los términos contenidos en la petición de búsqueda. Dicho conjunto es fácilmente identificado ya que cada uno de los términos tiene asignado una serie de valores números, cero o uno, que identifican los términos con los cuales fueron indexados los documentos. Por otro lado, el acceso a un documento en particular en un archivo directo (vector binario) muestra de manera inmediata aquellos términos con los cuales tal documento fué indexado.

Sin embargo con cualquiera de estas dos configuraciones, en muchos casos, se obtienen resultados inferiores a los esperados al momento de la búsqueda, pues obviamente algunos términos son más importantes para la representación de un documento que otros.

Una manera de resolver este problema es asignar una ponderación a los términos con que serán indexados los documentos. Eso va más allá de la simple asignación de valores números cero o uno (presencia o ausencia de un término en un documento). Una de las formas más comunes de ponderar un término es siguiendo la idea de **H.P. Luhn** la cual toma en cuenta la frecuencia de aparición ($\mathbf{Frec}_{ik}$) del **término**$_k$ a lo largo de un **documento**$_i$ y el número de documentos dentro de la colección que contienen este término ($\mathbf{DocFrec}_k$). A partir de lo anterior se tiene que la ponderación del **término**$_k$ en el **documento**$_i$ podría estar dada por el cociente de $\mathbf{Frec}_{ik}$ y $\mathbf{DocFrec}_k$:

$$\mathbf{Ponderación}_{ik} = \mathbf{Frec}_{ik} / \mathbf{DocFrec}_k$$

De esta manera la representación de un documento, utilizando los valores de ponderación de los términos, se haría mediante un vector en cuyas entradas se encuentran los valores correspondientes a la ponderación de cada uno de los términos del lenguaje de indexación. Esto es, dado el documento $\mathbf{Doc}_i$, su representación estará dada por medio de la sucesión de valores $\mathbf{Term}_{i1}$, $\mathbf{Term}_{i2}$, ... , $\mathbf{Term}_{it}$. Donde $\mathbf{Term}_{ij}$ representa la ponderación o importancia del término **j** asignado al **documento**$_i$. De esta forma una colección de documentos puede ser representada como un arreglo o matriz cuyas entradas son las ponderaciones de los términos ($\mathbf{Term}_{ij}$), donde cada renglón de la matriz representa un documento y cada columna representa la ponderación o importancia asignada de un término específico a lo largo de la colección de documentos.

En forma análoga, una **petición**$_j$ al sistema puede ser identificada como un vector $\mathbf{PTerm}_{j1}$, $\mathbf{PTerm}_{j2}$, ..., $\mathbf{PTerm}_{jt}$, donde $\mathbf{PTerm}_{jk}$ representa la ponderación o importancia del **término**$_k$ asignado a la **peticion**$_j$.

Las representaciones de la colección de documentos al igual que de la petición, pueden ser consideradas como elementos de un espacio vectorial sobre los reales que llamaremos espacio de documentos. La dimensión del espacio de documentos estará dada por el número de términos utilizados en la indexación de la colección de documentos, es decir por la cardinalidad del lenguaje de indexación.

Para identificar aquellos documentos dentro de la colección que mejor respondan a una petición particular, **petición**$_i$, es necesario establecer una medida que sea capaz de evaluar la similitud entre cada una de las representaciones de la colección y la representación de la petición dada. Una **medida de similitud** frecuentemente usada es la medida del coseno, definida como:

$$Cos(a) = P \cdot D \; / \; ||P|| \; ||D|| \qquad (1)$$

Donde el producto "·" es el producto punto usual en un espacio vectorial de dimensión finita, // // norma del vector y **a** es el ángulo entre los vectores **P** (representación de una petición) y **D** (representación de un documento).

Cuando una medida numérica de similitud es usada para la colección de documentos y peticiones, no es necesario como en el caso de las matrices binarias vistas anteriormente, recuperar todos aquellos documentos que contengan exactamente los términos utilizados en la petición. En lugar de ello la recuperación de los documentos puede depender de un límite particular preestablecido para la medida de similitud, o de un número específico de documentos por ser recuperados. Asumiendo, por ejemplo, que se establece como límite 0.50 para la medida de similitud, todos aquellos documentos los cuales presenten valores, en relación a la expresión 1, mayores o iguales que 0.50 serán recuperados. Alternativamente se puede establecer un tope de **N** documentos por recuperar.

Las representaciones de los documentos recuperados pueden ser convenientemente presentadas al usuario en orden decreciente con respecto a sus valores de similitud con la correspondiente petición de búsqueda. Esto es de importancia especial en una situación de recuperación interactiva, pues nuevas y mejores formulaciones de peticiones de búsqueda pueden ser construidas a partir de la información obtenida de las representaciones de los documentos previamente recuperados (*feedback*).

La expresión 1 nos provee de una medida para evaluar la similitud entre un vector-documento y un vector-petición en el espacio de documentos. Así pues, si quisiéramos establecer la similitud entre grupos de documentos, esta misma medida resulta una herramienta válida para formar agrupamientos de documentos que exhiben un alto grado de similitud entre sí. Tomando en cuenta esto, dados dos vectores-documento **Doc**$_i$ y **Doc**$_j$ los cuales son las representaciones de dos

documentos, el valor de similitud entre estos dos documentos puede ser definida como:

$$Doc_i \cdot Doc_j \;/\; ||Doc_i|| \; ||Doc_j|| \qquad (2)$$

Así como en una biblioteca se establece una clasificación de los libros que comparten temas relacionados, con el fin de organizarlos de tal modo que se permita facilitar la tarea de localización de los libros que traten temas afines, dentro de la colección de documentos también puede procederse de la misma forma, efectuando agrupamiento de documentos que tienen un alto grado de similitud entre ellos, tomando en cuenta la expresión 2.

Una manera de generar los agrupamientos dentro de una colección podría efectuarse por medio de la comparación de pares de representaciones de documentos utilizando la expresión 2, y así aquellos documentos que exhiban un valor de similitud entre ellos mayor a algún límite establecido serán considerados como parte de un grupo. La figura 2 esquematiza el proceso de agrupamiento, donde cada circunferencia representa un grupo dado de documentos que exhibieron un alto grado de similitud. Note que existen intersecciones entre los grupos, esto simplemente representa la posibilidad de que existan documentos que puedan pertenecer a varios grupos a la vez.

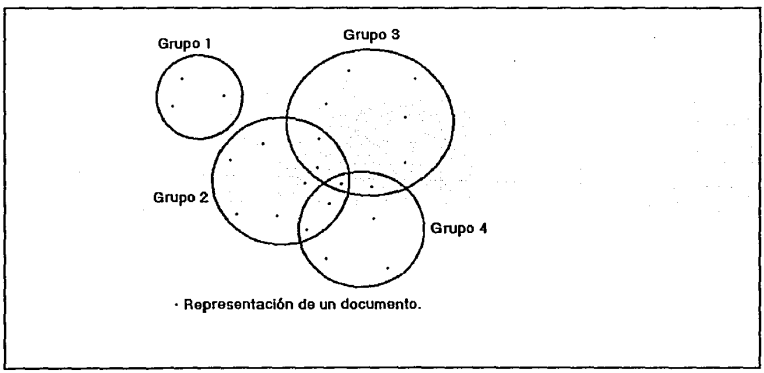


Fig. 2 Formación de agrupamientos de documentos que exhiben un alto grado de similitud.

Estos agrupamientos se efectúan con el fin de reducir el tiempo de búsqueda, de lo contrario, para la localización del conjunto de documentos relevantes a una petición dada deberá efectuarse una comparación de cada una de las representaciones de los documentos de la colección con la representación de la petición de la búsqueda y de esta forma se determinaría qué documentos son relevantes, lo que implicaría un mayor tiempo de búsqueda. Por otro lado, si se efectúan agrupamientos de documentos que exhiban un alto grado de similitud entre ellos, utilizando la expresión 2, y además se escoge por cada agrupamiento un vector promedio, al cual llamaremos **centroide**, que permita representar al grupo, entonces el tiempo de búsqueda de aquellos documentos relevantes a una petición dada, podría ser reducido vía la comparación ya no de cada una de las representaciones de los documentos sino únicamente comparando los centriodes de cada uno de los agrupamientos con la petición dada, y así aquellos centroides que reflejen una gran similitud con la representación de la petición serán escogidos para que ahora se proceda a comparar solamente los elementos de este grupo con la representación de la petición de búsqueda. De este modo se evita la comparación de aquellos elementos pertenecientes a los grupos cuyos centroide muestren poca similitud con la representación de la petición.

Una forma para determinar el centroide de cada grupo es mediante la determinación de un "vector promedio" de todos los elementos del grupo. Suponiendo que dado un conjunto de **M** vectores-documento que constituyan un cierto grupo **P**, el centroide correspondiente a este grupo **P** podría ser calculado como:

$$\mathbf{CENTROID_p = (CTerm_{p1}, CTerm_{p2}, \ldots, CTerm_{pt})}$$

Donde cada $CTerm_{pk}$ puede ser obtenido como el promedio de las ponderaciones del $término_k$ de todas las **M** representaciones de los documentos que componen el grupo **P**, esto es:

$$CTerm_{pk} = \frac{1}{M} \sum_{l=1}^{M} term_{lk}$$

La figura 3 muestra la localización del centroide de cada grupo en la figura anterior.

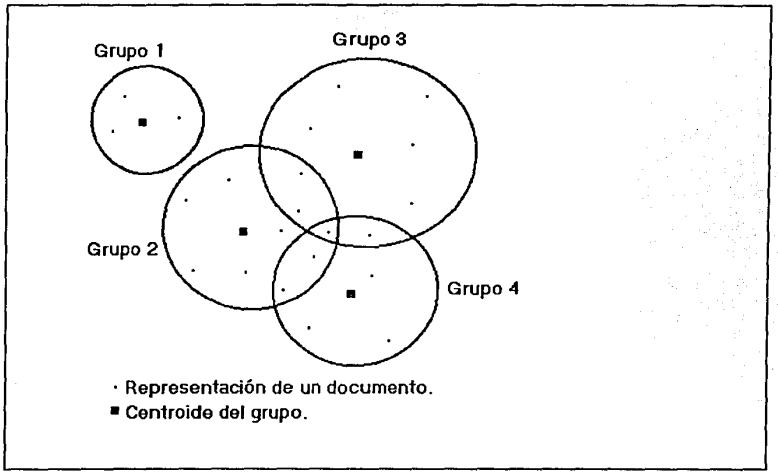


Fig. 3 Localización del centroide dentro de cada agrupamiento.

1.5. Evaluación.

Muchas de las investigaciones y gran parte del desarrollo en los Sistemas de Recuperación de Información se realizan con el objeto de mejorar la **efectividad** y la **eficiencia** de la recuperación. La eficiencia es medida usualmente en términos de los recursos del sistema de cómputo utilizados tales como la capacidad de almacenamiento y el tiempo de CPU. Siendo así es difícil medir la eficiencia sin tomar en cuenta el dispositivo de cómputo que se utiliza. Por otro lado, la efectividad de un Sistema de Recuperación de Información está determinada por la habilidad con que este sistema satisface las necesidades de información de sus usuarios. Actualmente la viabilidad de un Sistema de Recuperación de Información depende tanto de la calidad del sistema (efectividad), como del costo de operación del sistema (eficiencia).

Existen muchas razones para evaluar un Sistema de Recuperación de Información. Por ejemplo, esta evaluación podría ser de mucha utilidad en el caso en el que se quisiera comparar un sistema ya existente con otro sistema alternativo. También podría desearse determinar cómo el funcionamiento de un sistema cambia cuando algún o algunos componentes del sistema son cambiados; por ejemplo evaluar los cambios funcionales del sistema cuando el tipo de petición es alterado o cuando la especificidad y exhaustividad del lenguaje de indexación son cambiadas.

El funcionamiento de un Sistema de Recuperación de Información con frecuencia es medido en términos de la integridad de la recuperación (**recall**) y la **precisión** de la recuperación de la información. Más precisamente **recall** es una medida de la habilidad del sistema para recuperar los documentos útiles, mientras que la **precisión** es una medida de la habilidad del sistema para rechazar documentos no útiles.

Una de las maneras más comunes de efectuar el cálculo de **recall** y **precisión**, dada una petición de búsqueda, es mediante el cálculo del número total de documentos relevantes (relevantes a la petición de búsqueda) en una colección (**NDRel**), el número total de documentos recuperados (**NDRec**) y el número de documentos recuperados que son relevantes a la petición (**NDRR**). Tomando en cuenta lo anterior se suele definir a recall como la porción de la colección de documentos relevantes que fueron recuperados (el cociente de **NDRR** y **NDRel**) y por otro lado a la precisión como la porción de los documentos recuperados que son relevantes (el cociente de **NDRR** y **NDRec**)

$$Recall = \frac{NDRR}{NDRel} \qquad Precisión = \frac{NDRR}{NDRec}$$

Como puede notarse, para obtener el valor de recall se requiere conocer el número total de documentos relevantes en la colección con respecto a la petición de búsqueda. Cuando el tamaño de la colección de documentos es relativamente pequeño, con frecuencia es posible evaluar la relevancia para todos los documentos de la colección con respecto a la petición. Sin embargo, cuando el tamaño de la colección es grande esto no es posible. Así para obtener valores confiables de recall es necesario estimar el número total de documentos relevantes en la colección. Esto último podría ser llevado a cabo mediante técnicas de muestreo.

Se puede observar que tanto para el cálculo de **recall** como para el de **precisión**, es necesario determinar la relevancia o no relevancia de un documento con respecto a una petición dada. Esto nos conduce a definir cuándo un documento es relevante. A este respecto puede decirse que el concepto de relevancia es un tanto subjetivo ya que por lo general la relevancia de un documento dado depende de la utilidad del mismo para un usuario determinado. Un punto de vista más objetivo acerca de la

relevancia de un documento, toma en cuenta sólo a la petición dada y a un documento en particular, esto es:

> **Relevancia es la correspondencia en cuanto a contexto entre una petición y un documento, es decir es el grado en el cual el documento cubre el material que es apropiado para la petición.**

La efectividad de la recuperación de un Sistema de Recuperación de Información puede ser fácilmente medida utilizando la definición objetiva de la relevancia de un documento. Sin embargo, aún utilizando este punto de vista objetivo sobre la relevancia, se presentan dificultades al momento de cuantificar la relevancia de un documento con respecto a una petición, pues suelen presentarse casos de desacuerdo en cuanto a determinar los límites entre los diferentes grados de relevancia y la evaluación de la misma. Esto ha conducido a definir a la relevancia, por algunos, en términos probabilísticos. En este caso la relevancia suele definirse como una función de aquella probabilidad de similaridad entre el vocabulario de un documento y el de la petición de búsqueda que podría llevar a un usuario a aceptar un documento dado en respuesta a una petición particular.

Como ya se mencionó, para el cálculo de **recall** y **precisión** es importante, dada una petición de búsqueda, definir la relevancia o no relevancia de aquellos documentos en la colección, así como también la del conjunto de documentos que son recuperados. De esta forma, para cada petición de búsqueda se obtiene un par de valores correspondientes al recall y la precisión, que son utilizados como indicadores para evaluar la efectividad de la recuperación del sistema, dada la petición inicial de búsqueda. De este modo pueden ser comparados pares de valores de precisión y recall para dos peticiones de búsqueda **i** y **j**, y siempre que **recall**$_i$ <= **recall**$_j$ y **precisión**$_i$ <= **precisión**$_j$ entonces los resultados de la búsqueda con la petición **j** se considerarán superiores a los obtenidos con la petición de búsqueda **i**. Desafortunadamente los problemas aparecen cuando, por ejemplo, se obtiene **recall**$_i$ < **recall**$_j$ y **precisión**$_i$ > **precisión**$_j$ o viceversa, **recall**$_i$ > **recall**$_j$ y **precisión**$_i$ < **precisión**$_j$. En estos casos el usuario es el que determinará si su principal interés se encuentra en recuperar la mayor cantidad posible de documentos relacionados con su petición de búsqueda o si desea obtener sólo aquellos documentos que sean muy precisos con respecto a su petición, ya que en un sistema de recuperación convencional el **recall** se ve incrementado cuando el número de documentos recuperados se incrementa también y al mismo tiempo la **precisión** probablemente decrezca. Por lo tanto, cuando se desea obtener un valor de **recall** alto suelen utilizarse peticiones de búsqueda que involucren términos muy generales que permitan recuperar un buen número de documentos, mientras que cuando se desea obtener un valor de **precisión** alto suelen utilizarse peticiones de búsqueda que involucren términos muy específicos.

En la sección previa se mencionó que es posible obtener, por medio de un Sistema de Recuperación de Información, dada una petición de búsqueda, una lista ordenada con las referencias de los posibles documentos de interés. El orden dentro de la lista puede ser establecido presentando primero todas aquellas referencias de los documentos cuyas representaciones compartan al menos **k** (arbitrario) términos con la representación de la petición, en segundo lugar las referencias cuyas representaciones compartan exactamente **k-1** términos, y así sucesivamente podría continuarse hasta listar las referencias de documentos cuyas representaciones no compartan ningún término con la representación de la petición de búsqueda.

Otra manera de ordenar la lista de referencias es mediante el cálculo del coeficiente de similaridad que se mencionó en la sección previa. Este coeficiente refleja la similaridad entre cada una de las representaciones de los documentos en la colección y la de la petición de búsqueda. De esta manera la lista de referencias puede presentarse en orden decreciente conforme a los valores obtenidos con el coeficiente de similaridad entre las representaciones documento-petición.

Utilizando la lista ordenada de las referencias de los documentos, puede calcularse para cada orden **k** en la lista, pares de valores de recall y precisión tomando en cuenta, para este cálculo, sólo el conjunto de documentos cuyo orden en la lista sea a lo más **k**.

Como ejemplo considérese un sistema que contiene las representaciones de una colección de doscientos documentos, de la cual se sabe existen cinco documentos relevantes a una petición dada. La lista ordenada en forma decreciente de las representaciones de los documentos recuperados es mostrada en la figura 4. Donde además se han marcado (**X**) las referencias de aquellos documentos relevantes también se han calculado los valores de recall y precisión para cada orden en la lista. Por ejemplo para el orden 6 en la lista, se han recuperado 6 documentos de los cuales 4 son relevantes, por lo tanto, para este orden en la lista, se tienen los valores de recall y precisión de 4/5 ó 0.8 (NDRR = 4/NDRel = 5) y 4/6 ó 0.67 (NDRR = 4/NDRec = 6) respectivamente.

Resultados de Recall y Precisión después de la recuperación de n documentos			
n	# de documento	Recall	Precisión
1	588 x	0.2	1.0
2	589 x	0.4	1.0
3	576	0.4	0.67
4	590 x	0.6	0.75
5	986	0.6	0.60
6	592 x	0.8	0.67
7	984	0.8	0.57
8	988	0.8	0.50
9	578	0.8	0.44
10	985	0.8	0.40
11	103	0.8	0.36
12	591	0.8	0.33
13	772 x	1.0	0.38
14	990	1.0	0.36

Fig. 4 Ejemplo de la evaluación de una lista ordenada de documentos recuperados después de una petición de búsqueda. Con (X) se han marcado aquellos documentos que son relevantes a la petición.Figura extraída de la referencia [2].

Es posible hacer una representación gráfica de **recall** y **precisión**. En la figura 5 se ha elaborado una gráfica de recall y precisión utilizando los valores de la figura 4. Sin embargo esta gráfica no puede ser vista ni como una función de la precisión con respecto al recall, ni tampoco como una función de recall con respecto a la precisión, pues como se puede observar, por ejemplo, para el valor de **recall** = 0.4 corresponden dos valores de **precisión** (1.0 y 0.67). Esto último podría ser solucionado redefiniendo la gráfica de la figura 5 de tal modo que ésta se ajuste a una gráfica de tipo escalonada, como se muestra en la figura 6, en la cual se tiene un único valor de **precisión** para cada punto de **recall**.

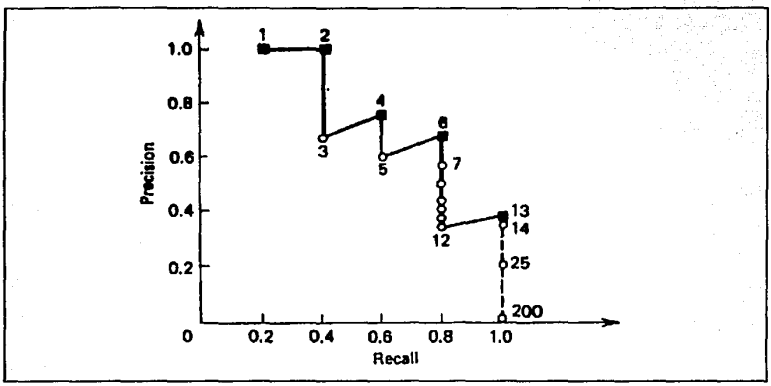


Fig. 5 Representación gráfica de los resultados de recall y precisión. Figura extraída de la referencia [2].

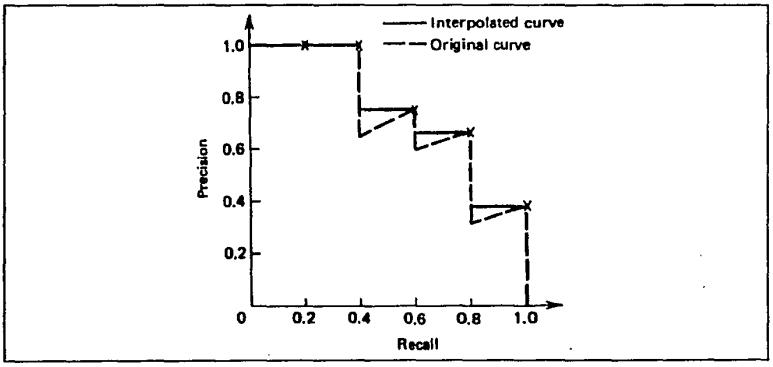


Fig. 6 Ajuste de la gráfica anterior a una gráfica tipo escalonada. Figura extraída de la referencia [2].

Para obtener un mejor criterio de evaluación es conveniente tomar en cuenta no sólo una, sino varias peticiones de búsqueda. Así, dado un conjunto de diferentes curvas de **recall** y **precisión** similares a la de la figura 6, obtenidas a partir de cada petición, es posible obtener una gráfica que muestre el comportamiento promedio de **recall** y **precisión** para todas las peticiones.

Un promedio que refleje el funcionamiento medio que un usuario puede esperar del sistema, puede obtenerse mediante el cálculo de las medias aritméticas de **recall** y **precisión** sobre un total de **NUM** peticiones:

$$recall_P = \frac{1}{NUM} \sum_{i=1}^{NUM} \frac{NDRR_i}{NDRel_i}$$

$$precisión_P = \frac{1}{NUM} \sum_{i=1}^{NUM} \frac{NDRR_i}{NDRec_i}$$

Donde $\mathbf{NDRR_i}$, $\mathbf{NDRel_i}$ y $\mathbf{NDRec_i}$, son los valores correspondientes a **NDRR**, **NDRel** y **NDRec** para la petición **i**. Una gráfica que refleje los valores de $\mathbf{recall_P}$ y $\mathbf{precisión_P}$ puede así obtenerse y sería semejante a la que se muestra en la fig. 5. En donde la parte superior izquierda de la gráfica corresponde a las formulaciones de peticiones de búsqueda muy específicas con las cuales son recuperados sólo algunos cuantos documentos, y de ahí que se espere que la **precisión** tome un valor alto, mientras que el valor de **recall** sea bajo. La parte inferior derecha de la gráfica representa las formulaciones de peticiones de búsqueda que son muy generales, de aquí que sean recuperados un gran número de documentos y que esto a su vez repercuta en una disminución en el valor de **precisión**, mientras que el valor de **recall** es alto.

Las curvas de **recall** y **precisión**, semejantes a la fig. 5, pueden ser usadas para evaluar el funcionamiento de los Sistemas de Recuperación de Información mediante el cálculo de los valores de **recall** y **precisión** para dos o más sistemas, o para el mismo sistema operando bajo diferentes condiciones. En estas circunstancias, las curvas producidas por los sistemas **A** y **B** pueden ser sobrepuestas en la misma gráfica para determinar cuál sistema es superior y qué tanto lo es. En general, la curva que más se acerque a la esquina superior derecha de la gráfica, donde los valores de **recall** y **precisión** son máximos, indica el mejor funcionamiento. Un típico ejemplo es mostrado en la fig. 7, donde se muestra el funcionamiento de un sistema bajo dos métodos diferentes de indexación, para una colección de documentos en una librería científica, el cual es evaluado sobre 35 peticiones de búsqueda.

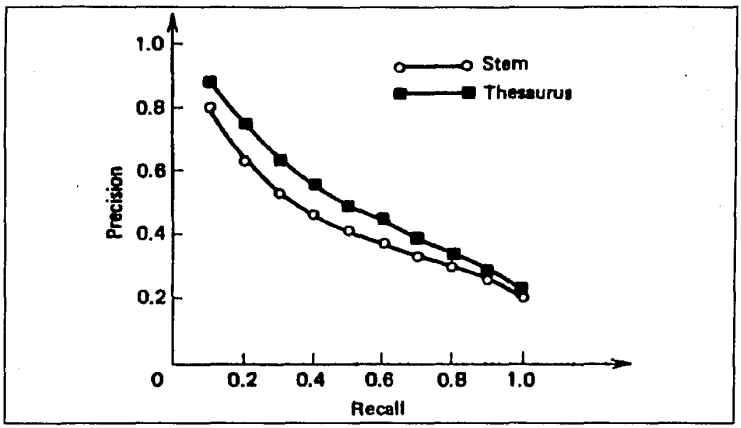


Fig. 7 Ejemplo que muestra el funcionamiento de un SRI bajo dos métodosdiferentes de indexación. Figura extraída de la referencia [2].

CAPITULO II Un SRI basado en triadas

Toda persona ha tenido la experiencia de encontrarse sumergido dentro de un conglomerado de grupos de personas que conversan simultáneamente entre sí. En tal situación y a pesar del aparente desorden y diversidad de los temas de conversación, uno es capaz de identificar casi inmediatamente a los grupos que conversan sobre un tópico en particular. La razón por la cual se es capaz de identificar el tema de conversación con tan sólo escuchar unas cuantas palabras es debido a que uno no necesita escuchar exactamente una determinada palabra para estar seguro del tema de conversación, en lugar de ello, el oído humano está a la caza de patrones de voz que sean semejantes a un determinado patrón que identifica al tema.

A pesar de la naturalidad del razonamiento anterior, la mayoría de los actuales SRI utilizan las palabras de los documentos para identificar su contenido. Esto trae como consecuencia la necesidad de idear complejos algoritmos que sean capaces de "comprender" la morfología del idioma en que esté escrita la colección de documentos. El presente trabajo de investigación tiene como objetivo el mostrar que es posible generar por cada documento un cierto patrón semejante a un patrón de voz y utilizar este patrón para la identificación del documento dentro del esquema de un SRI.

2.1.Indexación por triadas.

La indexación por triadas es una novedosa técnica de indexación automática, la cual consiste en descomponer cada una de las palabras de un documento en triadas de letras y utilizar a la frecuencia de éstas para construir un patrón de triadas (histograma) que conformará la indexación del documento. Para entender con mayor precisión lo anterior se convendrá que una triada válida para una palabra dada es cualquier fragmento continuo de tres letras que esté contenido en esta palabra. Por ejemplo el conjunto de triadas que puede ser construido en base a la palabra **sofia** (sin acento) es:

sof
ofi
fia

Esta nueva técnica cuenta con dos características que la hacen muy atractiva. La primera característica es la facilidad de su implementación y la segunda y más importante es que el proceso de indexación deja de depender casi en su totalidad de la lengua alfabética en que estén escritos los documentos. La razón por la que el proceso de indexación por triadas sigue dependiendo del idioma es debido a que no todas las posibles combinaciones de tres letras (triadas) son igualmente válidas en todas las lenguas alfabéticas. Naturalmente se podría salvar esta dependencia si se permitiera dar validez a cualquier combinación de tres letras en cualquier idioma, sin embargo el número de triadas válidas sería demasiado grande y debido a que las frecuencias de las triadas son las que se emplean para la indexación de los documentos, esto implicaría que los archivos que se generaran como resultado del proceso de indexación serían también grandes. Otro inconveniente quizá más grave que el anterior es que al considerarse un número mayor de triadas los tiempos de recuperación de un SRI basado en triadas aumentan. Por tal razón es recomendable realizar un análisis estadístico de las triadas válidas del idioma alfabético en que está escrita la colección de documentos. Este análisis no sólo permitirá determinar las triadas válidas sino también permitirá conocer con qué frecuencia es usada una triada en el idioma. Esto último es de gran utilidad si se extrapola la idea de **Luhn** acerca del poder de resolución de las palabras (capítulo anterior) para el caso de las triadas. De esta manera, al eliminar aquellas triadas poco frecuentes así como las muy frecuentes, el conjunto de triadas válidas (filtradas), que conformarán el lenguaje de indexación, puede ser reducido considerablemente sin mermar de manera significativa el rendimiento de un SRI basado en triadas.

El análisis estadístico de las triadas en el caso del idioma español, fue realizado con ayuda de un conjunto de artículos de periódicos cuyo volumen fue de aproximadamente 8 MBytes. Este archivo muestra fue preprocesado antes de su análisis de tal manera que fue eliminado todo acento y sustituida toda ñ por n, pues se considera que los acentos así como las ñ's dentro de los documentos escritos contribuyen muy poco a la interpretación del mismo. Como resultado del análisis estadístico se encontró que empleando este archivo muestra existían 6262 triadas válidas en este español sin acentos cuyas frecuencias de aparición oscilaban de 1 hasta 41226. Para reducir el conjunto de triadas válidas se decidió introducir un filtro para las 6262 triadas de tal manera que fueron consideras como válidas sólo 3636 triadas cuyas frecuencias se encontraran dentro de un determinado intervalo, el cual fue establecido de manera un tanto arbitraria pues fue en base a prueba y error (ver la sección de evaluación de este capítulo). De esta manera, se decidió que como un primer intento, se emplearan estas 3636 triadas para conformar el lenguaje de indexación, el cual como se verá a continuación es la base para la construcción de las representaciones de los documentos.

2.2. Estructuras de información usando triadas.

Una vez que ha sido establecido el conjunto de triadas que conformaran el lenguaje de indexación es posible construir representaciones o indexaciones de los documentos basados en este lenguaje de indexación. Para cada documento tal representación estará dada como un vector de dimensión **m** (cardinalidad del lenguaje de indexación) en el cual se asumirá una correspondencia biunívoca entre los elementos del lenguaje de indexación y las entradas de este vector en las cuales se habrá de almacenar la frecuencia de aparición de cada triada dentro del documento. Al unir las representaciones de cada documento en la colección se obtiene una estructura matricial la cual es el resultado final de la indexación de toda la colección de documentos, tal estructura será denominada como matriz de indexación.

	$Triada_1$	$Triada_2$	$Triada_3$		$Triada_m$
$Documento_1$	f_{11}	f_{12}	f_{13}	. . .	f_{1m}
$Documento_2$	f_{21}	f_{22}	f_{23}	. . .	f_{2m}
$Documento_3$	f_{31}	f_{32}	f_{33}	. . .	f_{3m}
	.	.	.	. . .	.
	.	.	.	. . .	.
	.	.	.	. . .	.
$Documento_n$	f_{n1}	f_{n2}	f_{n3}	. . .	f_{nm}

Fig. 8 Estructura matricial empleada para albergar la indexación de toda la colección de documentos en donde cada renglón contiene la indexación de sólo un documento. Con f_{ij} se denota la frecuencia de ocurrencia de la triada j en el i ésimo documento de la colección.

Debe observarse que la anterior estructura es semejante a la estructura empleada en el caso de la indexación por palabra o término (capítulo anterior).

De igual forma en que son indexados los documentos, será también indexada la petición de búsqueda, de tal manera que tanto los documentos como la petición serán mapeados a un espacio vectorial de dimensión finita **m**. ¿De qué manera se decidirá cuándo un determinado documento en la colección es relevante a una petición de búsqueda dada?. Tal pregunta quedará resuelta al definir una medida de similitud entre elementos de este espacio vectorial. En el capítulo anterior se mencionó a la medida *coseno* como una posible medida de similitud. Sin embargo, para el caso de la indexación por triadas se han detectado algunos inconvenientes al usar esta

medida de similitud. El problema fundamental que está detrás de estos inconvenientes es que la medida *coseno* de una petición **P** y un documento **D** depende de la norma del vector **D**, es decir la medida de similitud varía aún cuando se varían exclusivamente los valores de las frecuencias de las triadas del vector **D** que tienen un valor de cero en el vector de petición **P**. Para ejemplificar lo anterior supóngase la siguiente situación:

P = (1,1,0)
D1 = (1,1,v) donde **v** es cualquier natural.
D2 = (1,0,0)

Para tal situación se tienen los siguientes valores de similitud S.

$$S(P,D1) = ||P||Cos(A1) = P \cdot D1/||D1||$$
$$S(P,D2) = ||P||Cos(A2) = P \cdot D2/||D2|| = 1$$

Dado que el documento **D1** contiene todas las triadas de la petición **P**, debería considerarse a este documento como más próximo a la petición **P**, sin embargo, se observa que cuando $v > 1$ entonces $S(P,D1) < S(P,D2)$. Para resolver este problema podría pensarse en utilizar como medida de similitud al producto punto y emplear ahora vectores binarios para la representación de los documentos en donde deje de importar la frecuencia de las triadas y sólo sea importante si están o no están (1 ó 0) presentes en el documento que se indexa. Sin embargo, en este trabajo se ha supuesto que la idea de **H. P. Luhn** acerca del esquema de ponderación de las palabras dentro de un documento, es también válido cuando es aplicado a triadas en lugar de palabras.

Así pues, sería deseable contar con una medida de similitud que sin dejar de considerar a la frecuencia de las triadas le diera más importancia a los documentos que contengan el mayor número de triadas diferentes que estén presentes en la petición. Es decir, lo que se busca es una medida de similitud **S** tal que dados los documentos **D1** y **D2** y una petición **P** cumpla con los siguientes lineamientos:

1.- Si $P^* \cdot D1^* > P^* \cdot D2^*$ entonces $S(P,D1) > S(P,D2)$
2.- Si $P^* \cdot D1^* = P^* \cdot D2^*$ entonces
Si $P \cdot D1 > P \cdot D2$ entonces $S(P,D1) > S(P,D2)$
de lo contrario $S(P,D1) <= S(P,D2)$

$$\text{Donde } \mathbf{V}^{\cdot}=(v_1^{\cdot},v_2^{\cdot},\ldots,v_n^{\cdot}) \quad \text{con} \quad v_k^{\cdot} = \begin{cases} 0 & \text{si la k ésima entrada del vector } \mathbf{V} \text{ es cero} \\ 1 & \text{de lo contrario} \end{cases}$$

Una expresión algebráica para tal medida de similitud es como sigue:

$$\mathbf{S(P,D) = P^{\cdot} \cdot D + ma^2 P^{\cdot} \cdot D^{\cdot}} \qquad (3)$$

En donde **m** es igual a la cardinalidad del lenguaje de indexación y **a** es una constante cuyo valor deberá ser mayor al valor máximo permitido para las frecuencias de las triadas, es decir esta constante **a** habrá de ser una cota superior para el valor de las frecuencias de las triadas. La demostración de que la expresión 3 cumple con las condiciones de la medida de similitud buscada se encuentra en el apéndice de este trabajo.

Lo más recomendable sería pues, emplear la expresión algebráica anterior como medida de similitud para un SRI basado en triadas. Sin embargo, para lograr una implementación eficiente se vió la conveniencia de imponer a la medida de similitud **S** la restricción de que la cardinalidad de su contradominio sea igual a 256, es decir, lo que se desea es que un byte sea suficiente para albergar el resultado de cualquier medida de similitud entre dos vectores. Esto último entra en conflicto con la medida de similitud propuesta en 3, pues debido a que el valor del factor $\mathbf{f = ma^2}$ es grande el rango de valores será también grande para la medida **S**. Debido a lo anterior se decidió por un lado restringir a la constante **a** a un valor pequeño (por ejemplo 4) y por otro lado experimentar con medidas de similitud que aunque semejantes a la expresión 3 tuvieran como valor del factor **f** un valor alrededor de la constante **a**. Esto último es un tanto arbitrario pues esta familia de medidas de similitud no tienen que cumplir los lineamientos que satisface la expresión 3, sin embargo y dependiendo de la colección de documentos, se estima que es posible encontrar dentro de esta familia de medidas de similitud una cuyo comportamiento sea lo más aproximado al que se obtendría con la expresión 3. Una metodología para determinar valores apropiados para **a** y **f**, puede ser a través de las pruebas de evaluación, y en ese sentido se ha observado que para el caso particular de una colección de más de medio millón de documentos los cuales tienen una extensión promedio de 235 bytes, se obtienen buenos resultados si es fijado para **a** y **f** un valor de 4.

Ahora bien, debido a que en general la probabilidad de no presencia o escasa presencia de una determinada triada en un determinado documento es mayor que la probabilidad de una alta presencia, es recomendable, para evitar que la matriz de

indexación sea muy grande, utilizar códigos de compresión tales como los códigos de **Hoffman** [3] para la codificación de esta matriz de indexación. Es claro que al comprimir la matriz de indexación se afectarán negativamente los tiempos de respuesta de las búsquedas, pues de esta manera será necesario descomprimir la matriz de indexación antes de identificar a los documentos cuyas representaciones muestran los valores más grandes de similitud con respecto a la representación de la búsqueda. Una manera de lograr que los tiempos de búsqueda no se vean tan afectados con esta medida es no comprimir a la matriz de indexación como un todo y en su lugar realizar la compresión de la matriz por columnas independientes. De esta forma, cada vez que sea iniciada una búsqueda será necesario sólo descomprimir de la matriz de indexación aquellas columnas cuyas triadas correspondientes se encuentran presentes en la petición de búsqueda.

A continuación se da una descripición más detallada del algoritmo empleado para la indexación de los documentos así como también el algoritmo para la búsqueda de documentos sobre la colección ya indexada.

2.3 Algoritmo de indexación.

El proceso para generar la matriz de indexación es sumamente complejo y su complejidad radica principalmente en la necesidad, como ya se hizo mención, de comprimir por separado las columnas de la matriz. Para llevar un control preciso del proceso de indexación se empleó una estructura de datos cuya función es semejante a la que realiza la *Process Status Word* (**PSW**) en el caso del funcionamiento global, a nivel de hardware, de una computadora digital. Esta estructura de datos, la cual fue denominada con el nombre de **estado**, tiene concretamente como función la de registrar el estado final de la matriz de indexación después de la última indexación. Así pues la información que esta estructura **estado** maneja es la siguiente:

- Un contador global, con el cual se lleva cuenta del número total de documentos que han sido indexados hasta el momento.

- Un apuntador por cada columna de la matriz de indexación, a través de estos apuntadores se especifica el inicio de cada columna dentro de la matriz de indexación. Se debe recordar que las columnas de la matriz de indexación se hayan comprimidas y por tal razón, no existe una regularidad en cuanto al tamaño y ubicación de cada columna dentro de esta matriz.

- Un contador por cada columna de la matriz de indexación, este contador registra el tamaño en bytes de la columna comprimida que le corresponda. La función que estos contadores realizan es muy importante pues a través de éstos puede saberse con precisión el número de bytes que deben ser leídos de la matriz de indexación cuando se desee manipular una de sus columnas.

El proceso de indexación de la colección de documentos es en esencia un proceso secuencial en el sentido de que la indexación de la colección es realizada al indexar uno por uno todos los documentos de la colección. De esta manera la indexación de la colección es el resultado de una sucesión de indexaciones. En cada una de estas indexaciones se indexa uno y sólo uno de los documentos de la colección y el resultado de esta indexación es agregado como un nuevo renglón al final de la matriz de indexación. Los pasos que se realizan para indexar y agregar un nuevo documento a la colección de documentos ya indexados, son los siguientes:

- Se extraen del texto del documento a indexar todas aquellas palabras nulas (ver indexación automática en el capítulo anterior).

- A partir del texto resultante del paso anterior, se construye un vector de frecuencias de triadas, el cual es equivalente a cualquier renglón de la matriz de indexación de la figura 8.

- Con la ayuda de la estructura de **estado** se realizan para cada columna de la matriz de indexación o entrada del vector de frecuencias los siguientes 4 pasos:

    - Se lee la **i** ésima columna de la matriz de indexación la cual es descomprimida a continuación.

    - A la columna resultante del paso anterior se le agrega al final la **i** ésima entrada del vector de frecuencias.

    - Esta columna agrandada es comprimida a continuación para obtener de esta forma una nueva columna comprimida que incluya ya la **i** ésima entrada del vector de frecuencias.

    - La nueva columna comprimida es ahora escrita como parte de una nueva matriz de indexación.

- Una vez que han sido leídas todas las columnas de la matriz de indexación y por tal razón también hayan sido escritas todas las columnas de la nueva matriz de indexación, es borrada la matriz de indexación anterior y declarada a la nueva matriz como la actual.

- La estructura **estado** es modificada de tal manera que refleje las condiciones de la nueva matriz de indexación.

2.4. Algoritmo de búsqueda.

Una vez que ha sido indexada toda la colección de documentos, es posible la búsqueda de documentos relevantes a una petición dada. Esta búsqueda será realizada sobre la matriz de indexación y bajo el siguiente algoritmo:

- Se extraen del texto de la petición todas las palabras nulas y con el texto restante es construído el vector de frecuencias de triadas de la petición el cual se denotará por $\mathbf{P} = (p_1, p_2, ..., p_m)$ (donde **m** es igual a la cardinalidad del lenguaje de indexación).

- Es inicializado con ceros un vector el cual deberá tener tantas entradas como documentos en la colección ya indexada. En las entradas de este vector habrán de depositarse los resultados de las medidas de similitud de todos los documentos de la colección con respecto a la petición. Por esta razón, este vector será denominado **vector de similitudes** y será además denotado como $\mathbf{S} = (s_1, s_2, ..., s_n)$ (donde **n** es igual al número de documentos en la colección).

- Con la ayuda de la estructura **estado** se realizan para cada **j** ésima entrada del vector de frecuencias que sea diferente de cero los siguientes 2 pasos:

    - Se lee la **j** ésima columna de la matriz de indexación la cual es descomprimida inmediatamente después. En su momento esta columna descomprimida será denotada por $\mathbf{C_j} = (c_{1j}, c_{2j}, ..., c_{nj})$.

    - El **vector de similitudes S** será actualizado como sigue:

$$S_i = S_i + P_j C_{ij} + (f=4)\ P_j^* C_{ij}^* \qquad i=1,2,...,n$$

- Se ordenan de mayor a menor los valores del **vector de similitudes** y se imprimen finalmente los nombres de los primeros **N** documentos que correspondan a las primeras **N** entradas del vector reordenado **S**.

2.5. Evaluación de un SRI basado en triadas.

Para la evaluación de un SRI basado en triadas fué empleado un banco de información de más de medio millón de fichas bibliográficas del cátalogo de la biblioteca de la DGB de la UNAM, las cuales tienen una extensión promedio de 235 bytes. Este banco de información se indexó utilizando la técnica de indexación por triadas descrita anteriormente y como resultado se obtuvo la matriz de indexación cuyo tamaño es de aproximadamente 49 Mbytes. Dado que el tamaño del banco de información es de aproximadamente 122 Mbytes esto significa que la matriz de indexación ocupa aproximadamente el equivalente al 40% de este espacio. La indexación de este banco de información fue realizada por una supercomputadora **Cray Y-MP/432** y el tiempo de CPU que ésta requirió para indexar todo el banco fue de aproximadamente 7 horas. Podría parecer que es demasiado el tiempo de indexación y sobre todo para un equipo de cómputo como lo es una **Cray**, sin embargo, a este respecto se deben mencionar tres puntos. Primero, la indexación del banco de información se realiza sólo una vez. Segundo, la razón por la que se empleó la **Cray** obedece más a una necesidad de contar con una buena cantidad de disco y no a una necesidad de velocidad de proceso, en este sentido se comenta que debido a la naturaleza del algoritmo de indexación se estima que el tiempo de indexación en otros equipos de cómputo no serán extremadamente diferentes a los reportados por la **Cray**. Tercero, por la manera en que se ha hecho la implementación del algoritmo de indexación los resultados que con este algoritmo se obtienen son independientes de la plataforma de cómputo que se empleó para generarlos y por esta razón, estos resultados pueden, sin nigún problema, ser usados para la búsqueda de documentos bajo otra plataforma de cómputo incluyendo computadoras personales.

Por otro lado, es importante mencionar que dentro de este banco de información existen fichas bibliográficas escritas en otros idiomas diferentes al español como son el inglés, francés e italiano. A pesar de tener conocimiento de lo anterior, se empleó para la indexación del banco de información el lenguaje de indexación construido especialmente para el idioma español. Dos son las razones por las que se decidió indexar al banco de información con este lenguaje de indexación, la primera razón es simplemente por que la gran mayoría de las fichas bibliográficas están escritas en español, la segunda razón es por que se pensó que sería interesante evaluar el comportamiento del SRI cuando algunos de los documentos de la colección no estuvieran escritos en español.

Como se recordará del capítulo anterior, la manera en que se suele evaluar un SRI es a través de las medidas de **recall** y **precisión**. Sin embargo, para llevar a cabo esta evaluación es necesario primero conformar un padrón heterogeneo de peticiones de búsqueda y emplear este padrón en las pruebas de evaluación. Debido a la manera en que se define la medida **recall** es necesario que antes de iniciar las pruebas de evaluación sea calculado o en el mejor de los casos estimado mediante técnicas de muestreo el número total de documentos dentro de la colección que son relevantes a cada petición dentro del padrón (cálculo de **NDRel** para cada petición en el padrón). Para conformar un buen padrón heterogeneo de peticiones así como para estimar el factor **NDRel** para cada petición en el padrón, se requiere de un grupo de personas que efectuen un trabajo de análisis de la colección de documentos. Por esta razón y debido a la imposibilidad de contar por el momento de un grupo de trabajo, se decidió en primera instancia omitir el cálculo de **recall** y sólo realizar el cálculo de **precisión** en las pruebas de evaluación, en segunda instancia se decidió que el padrón de peticiones de búsqueda fuera conformado con ayuda de un programa de computadora que "analizara" a la colección de documentos y en base a este análisis conformara el padrón de peticiones. La manera en que operó este programa sobre la colección de documentos para construir este padrón fue como sigue:

1.- Se eligieron aleatoriamente 1245 documentos de la colección.

2.- Para cada uno de los 1245 documentos seleccionados en el paso anterior, se construyó una petición de búsqueda la cual fue construida como una cadena de entre una y cuatro palabras que fueron extraídas al azar de alguno de los 1245 documentos.

A continuación se muestra un fragmento del padrón de peticiones que fue construído con ayuda de este programa de computadora (el número al principio de cada petición indica cuantas palabras conforman a la petición).

```
3 mexico ignacio nacional
3 investimenti giuffre della
4 latin gramatica venezia giovanni
4 metcalfe mechanic and york
4 press new interviewes review
1 manual
3 cristianismo nikander modern
2 shivers ingenieria
4 york philip notes chinese
3 and hoffman john
3 mexico exteriores relaciones
2 gardening macmillan
3 manuales documentacion mariano
2 guillermo mesa
3 deliberazione delle leyes
2 riccardo impresa
3 geral aires losada
4 obras tomas polemicas doctrinales
2 directa barcelona
3 ineficacia graficas garcia
```

Debido a la imposibilidad de realizar manualmente 1245 peticiones de búsqueda y evaluar los resultados, se desarrolló otro programa de cómputo que realizara este trabajo de evaluación. Este nuevo programa analiza, por cada petición de búsqueda en el padrón, los documentos que se encuentran en la lista recuperada y entregada por el SRI. Este análisis consistió en determinar el número que ocupaba aquel primer documento, dentro de la lista entregada por el SRI, que contuviera todas las palabras incluídas en la petición de búsqueda. Se decidió por simplicidad que el programa sólo analizara los diez primeros documentos de la lista ordenada, de tal manera que de no encontrarse un documento que contuviera todas las palabras de la petición se reportaría como una búsqueda fallida.

A continuación se muestra un fragmento del reporte generado por este programa evaluador.

```
ref = 2  3  mexico  ignacio  nacional
ref = 1  3  investimenti  giuffre  della
ref = 1  4  latin  gramatica  venezia  giovanni
ref = 3  4  metcalfe  mechanic  and  york
ref = 1  4  press  new  interviewes  review
ref = 1  1  manual
ref = 1  3  cristianismo  nikander  modern
ref = 1  2  shivers  ingenieria
ref = 1  4  york  philip  notes  chinese
ref = 9  3  and  hoffman  john
ref = 1  3  mexico  exteriores  relacines
ref = 1  2  gardening  macmillan
ref = 7  3  manuales  documentacion  mariano
ref = -  2  guillermo  mesa
ref = 2  3  deliberazione  delle  leyes
ref = 4  2  riccardo  impresa
ref = -  3  geral  aires  losada
ref = 1  4  obras  tomas  polemicas  doctrinales
ref = 1  2  directa  barcelona
ref = 1  3  ineficacia  graficas  garcia
```

Como puede observarse este fragmento es casi igual al fragmento del padrón de peticiones salvo por la primera columna numérica con la cual se ha indicado como resultado para cada petición de búsqueda, el número que ocupa el primer documento dentro de la lista ordenada entregada por el SRI que contiene todas las palabras de la petición. Cuando un guión aparece en lugar del número, esto significa que no se encontró tal documento dentro de la lista de los diez primeros documentos.

Una manera de resumir los resultados contenidos en el reporte de evaluación es a través de la siguiente tabla:

```
total de pruebas de 1 palabras =  304
  1   1   176   57.89 %
  1   2    34   11.18 %
  1   3    15    4.93 %
  1   4     6    1.97 %
  1   5     4    1.32 %
  1   6     2    0.66 %
  1   7     4    1.32 %
  1   8     4    1.32 %
  1   9     1    0.33 %
  1  10     4    1.32 %
  1   -    54   17.76 %

total de pruebas de 2 palabras =  304
  2   1   166   54.61 %
  2   2    38   12.50 %
  2   3    16    5.26 %
  2   4     9    2.96 %
  2   5     4    1.32 %
  2   6     7    2.30 %
  2   7     8    2.63 %
  2   8     6    1.97 %
  2   9     5    1.64 %
  2  10     2    0.66 %
  2   -    43   14.14 %

total de pruebas de 3 palabras =  325
  3   1   239   73.54 %
  3   2    29    9.00 %
  3   3     7    2.15 %
  3   4     7    2.15 %
  3   5     3    0.92 %
  3   6     7    2.15 %
  3   7     4    1.23 %
  3   8     2    0.62 %
  3   9     3    0.92 %
  3  10     3    0.92 %
  3   -    24    7.38 %

total de pruebas de 4 palabras =  312
  4   1   250   80.13 %
  4   2    28    8.97 %
  4   3     5    1.60 %
  4   4     7    2.24 %
  4   5     3    0.96 %
  4   6     1    0.32 %
  4   7     0    0.00 %
  4   8     3    0.96 %
  4   9     4    1.28 %
  4  10     0    0.00 %
  4   -    11    3.53 %
```

En esta tabla como puede observarse se han separado los resultados en cuatro grupos, el primer grupo sumariza los resultados de todas aquellas peticiones de búsqueda que están formadas por una sola palabra, el segundo grupo las peticiones con dos palabras, el tercero con tres palabras y el cuarto con cuatro palabras. Cada uno de estos cuatro grupos de resultados esta conformado a su vez por once renglones. En el primer renglón se muestra el número de documentos, que conteniendo todas las palabras (1,2,3 y 4 palabras, dependiendo el grupo) de la petición, fueron encontrados en el primer lugar de la lista ordenada entregada por cada petición por el SRI. En el segundo renglón se muestra el número de

documentos que fueron encontrados en la segunda posición dentro de la lista. El n-ésimo renglón (1 <= n <= 10) muestra el número de documentos que fueron encontrados en la n-ésima posición dentro de la lista ordenada. El último renglón muestra el número de peticiones de búsqueda fallidas.

Debe observarse que los resultados presentados en la tabla anterior no representan una evaluación de la precisión del SRI tal como se sugiere en el capítulo anterior. Sin embargo, se considera que esta tabla refleja de manera muy aproximada la precisión real del SRI basado en triadas.

Estas pruebas de evaluación fueron realizadas en un equipo **SUN 630/MP** con el cual se encontró que el tiempo de respuesta promedio de una búsqueda sobre toda la colección (más de medio millón de fichas) es de alrededor de tres segundos.

Los resultados de precisión y tiempos de respuesta expuestos anteriormente se obtuvieron utilizando un lenguaje de indexación de 3636 elementos. Se decidió ampliar el filtro que determina el lenguaje de indexación (leer al inicio de este capítulo) con lo cual se obtuvo un nuevo lenguaje de indexación con 3797 elementos. Utilizando este nuevo lenguaje de indexación se reindexó la colección de documentos y hasta este punto se observó que la matriz de indexación se incrementó con respecto a la anterior en 4.5 Mbytes. A continuación se efectuaron, para este nuevo lenguaje de indexación, las pruebas de evaluación descritas anteriormente con las cuales se obtuvieron los siguientes resultados:

```
total de pruebas de 1 palabras =  304
 1  1  185  60.86 %
 1  2   34  11.18 %
 1  3   11   3.62 %
 1  4    7   2.30 %
 1  5    4   1.32 %
 1  6    2   0.66 %
 1  7    3   0.99 %
 1  8    6   1.97 %
 1  9    1   0.33 %
 1 10    5   1.64 %
 1  -   46  15.13 %

total de pruebas de 2 palabras =  304
 2  1  176  57.89 %
 2  2   41  13.49 %
 2  3   16   5.26 %
 2  4    9   2.96 %
 2  5    5   1.64 %
 2  6    4   1.32 %
 2  7    8   2.63 %
 2  8    4   1.32 %
 2  9    4   1.32 %
 2 10    2   0.66 %
 2  -   35  11.51 %
```

```
total de pruebas de 3 palabras =  325
 3   1   249   76.62 %
 3   2    27    8.31 %
 3   3    11    3.38 %
 3   4     3    0.92 %
 3   5     7    2.15 %
 3   6     3    0.92 %
 3   7     4    1.23 %
 3   8     3    0.92 %
 3   9     6    1.85 %
 3  10     3    0.92 %
 3   -     9    2.77 %

total de pruebas de 4 palabras =  312
 4   1   261   83.65 %
 4   2    21    6.73 %
 4   3     9    2.88 %
 4   4     5    1.60 %
 4   5     4    1.28 %
 4   6     1    0.32 %
 4   7     0    0.00 %
 4   8     1    0.32 %
 4   9     0    0.00 %
 4  10     1    0.32 %
 4   -     9    2.88 %
```

Se desprende de estos resultados que la precisión mejoró en aproximadamente 3 puntos porcentuales. En cuanto al tiempo promedio de respuesta se observó que fue ligeramente mayor al obtenido con el anterior lenguaje de indexación.

Conclusiones y trabajos pendientes

Podría parecer para algunas personas que los resultados de precisión presentados en la sección anterior no son tan buenos, sin embargo, debe recordarse que las pruebas de evaluación de la precisión no fueron hechas de la manera más correcta por falta de recursos y en su lugar fueron empleados programas de cómputo que realizaron el trabajo de evaluación. Por esta razón, se tiene la certeza de que si un grupo de personas construye un padrón de peticiones más cercano a la realidad y si además se emplean personas que analicen los documentos recuperados en cada búsqueda del padrón, los resultados de precisión serán mejores.

En cuanto a la evaluación del recall del SRI basado en triadas, se comenta que aunque no ha sido efectuada una prueba para tal medida, se estima de las observaciones del comportamiento del SRI que el valor de esta medida sería bastante aceptable.

Es obvio que además de quedar pendiente una evaluación rigurosa del comportamiento de un SRI basado en triadas, es recomendable también utilizar este tipo de pruebas para refinar el criterio de selección del lenguaje de indexación así como también para la selección de una buena medida de similitud.

Para concluir se debe mencionar que un SRI basado en triadas para la consulta de fichas bibliográficas de la **UNAM** se encuentra actualmente dando servicio a todo público que tenga acceso a la red universitaria. Este servicio se encuentra en una máquina **SUN 630/PM** cuya dirección es **condor.dgsca.unam.mx** (dirección IP **132.248.10.3**). Para accesar a este servicio es necesario conectarse a este nodo vía *telnet* e ingresar **info** como *login*, inmediatamente después de lo cual serán presentados una serie de menúes. Para consultar finalmente las fichas bibliográficas deberá seleccionarse:

- **Consulta de catálogos en línea de biblioteca.**

en el primer menú,

- **Bibliotecas de la UNAM.**

en el segundo menú y en el tercer menú:

- **Acervo de la Biblioteca Central (DGB).**

Apéndice

Sean **P**, **D**, **D1** y **D2** elementos de un espacio vectorial de dimensión **m** sobre el campo de los enteros, en donde se pide además que las entradas de estos vectores sean mayores o iguales a cero y menores a una costante **a** > 0, entonces si todo lo anterior se cumple, la función de similitud **S** dada por:

$$S(P,D) = P^* \cdot D + ma^2 P^* \cdot D^*$$

(donde P^* y D^* son los vectores caracteristicos de **P** y **D** respectivamente)

cumple con lo siguiente:

1.- Si $P^* \cdot D1^* > P^* \cdot D2^*$ entonces $S(P,D1) > S(P,D2)$

2.- Si $P^* \cdot D1^* = P^* \cdot D2^*$ entonces

Si $P \cdot D1 > P \cdot D2$ entonces $S(P,D1) > S(P,D2)$

de lo contrario $S(P,D1) <= S(P,D2)$

Demostración:

Es claro que si $P^* \cdot D1^* = P^* \cdot D2^*$ entonces

$S(P,D1) > S(P,D2)$ si $P \cdot D1 > P \cdot D2$ y

$S(P,D1) <= S(P,D2)$ si $P \cdot D1 <= P \cdot D2$

Por lo tanto sólo es necesario demostrar que:

si $P^* \cdot D1^* > P^* \cdot D2^*$ entonces $S(P,D1) > S(P,D2)$

lo cual por otro lado siempre se cumple mientras se tenga la certeza de que:

$P \cdot D1 >= P \cdot D2$.

Asi pués sólo resta demostrar que:

si $P^* \cdot D1^* > P^* \cdot D2^*$ entonces $S(P,D1) > S(P,D2)$

aún cuando $P \cdot D1 < P \cdot D2$. En tales circuntancias es claro que:

$$\sum_{i=1}^{M} P_i > 0 \qquad y \qquad P^* \circ D1^* \geq P^* \circ D2^* + 1$$

y por tanto:

$$a^2mP^{\bullet}\circ D1^{\bullet} \geq a^2mP^{\bullet}\circ D2^{\bullet} + a^2m\ .$$

Ahora bien debido a que:

$$P\circ D1 \geq \sum_{i=1}^{m} P_i \quad , \quad P\circ D2 \leq a\sum_{i=1}^{m} P_i \leq a^2m \quad y \quad \sum_{i=1}^{m} P_i > 0$$

Se tiene que:

$$a^2mP^{\bullet}\circ D1^{\bullet} + P\circ D1 \geq a^2mP^{\bullet}\circ D1^{\bullet} + \sum_{i=1}^{m} P_i > a^2mP^{\bullet}\circ D2^{\bullet} + a^2m$$

$$\geq a^2mP^{\bullet}\circ D2^{\bullet} + a\sum_{i=1}^{m} P_i \geq a^2mP^{\bullet}\circ D2^{\bullet} + P\circ D2$$

lo cual quiere decir finalmente que **S(P,D1) > S(P,D2).**

Bibliografía

A continuación se presenta una lista de referencias bibliográficas de los documentos en que está basado parte del presente escrito. Sin embargo, como puede observarse la lista es muy reducida y de hecho estas referencias fueron de utilidad sólo para la realización del primer capítulo de este trabajo. La razón por la que no se emplearon documentos en los que se basara el trabajo del segundo capítulo es simplemente porque no existen (de hecho se intentó búscarlos) y por esto la investigación que aquí se presenta es totalmente original.

[1] **Information Retrieval.**
Second Edition.
C.J. van Rijsbergen.
University of Cambridge.
Butterworth & Co (Publishers) Ltd., 1979.
London.

[2] **Introduction to modern information retrieval.**
Gerald Salton.

[3] **Coding and information theory.**
R. W. Hamming & Englewood Cliffs, N. J.
Prentice - Hall, 1980.