SISTEMA DE RECUPERACIÓN DE INFORMACIÓN BASADO EN EL MODELO VECTORIAL

Luis Ignacio Lizcano B. Profesor Titular Dpto de Sistemas Universidad Francisco de Paula Santander
llizcano@yahoo.com

Diego Armando Pérez Hernández Estudiante de Ingeniería de Sistemas IX Semestre
Universidad Francisco de Paula Santander
diegoarmandoh@hotmail.com

RESUMEN

Este artículo contiene la descripción de un diseño de un Sistema de Recuperación de Información (SRI) con el modelo conceptual basado en el espacio vectorial. Contiene una descripción de los fundamentos básicos de los conceptos de la recuperación de información (RI) y especifica una arquitectura de un SRI con base a dos módulos de indexación y de consulta. La intención es diseñar un motor de indexación y búsqueda completamente operacional, que pueda ser utilizado en entornos documentales. El diseño puede darse sobre una base de datos relacional (Figuerola, Berrocal y Zazó 2000), que facilita la observación, manipulación de estructuras y resultados intermedios; la realización de las operaciones fundamentales a partir de sentencias SQL, permiten una fácil modificación de su funcionamiento interno.

INTRODUCCIÓN

La escritura es probablemente uno de los medios más antiguos de almacenar y transmitir el conocimiento, y a partir de un cierto volumen de texto escrito se hace imprescindible un sistema organizativo que posibilite la localización de la información que se precise en cualquier momento. Esta necesidad ha estado cubierta por técnicas que no han variado en 200 años básicamente hasta que la disponibilidad de ordenadores cada vez más potentes, dispositivos de almacenamiento más rápidos y de mayor capacidad y las redes de ancho de banda han producido una explosión de la información que no puede ser afrontada sin un amplio conjunto de nuevas técnicas de almacenamiento, acceso, interrogación y manipulación de esa información.

El desarrollo de los sistemas automatizados de recuperación de información se inicio con el objetivo de facilitar el manejo de la enorme cantidad de literatura científica surgida de los años 40 (Mañas 94). No ha quedado restringida a este campo sino que se ha extendido a otras áreas: cualquier disciplina que base su trabajo en la utilización de documentos puede beneficiarse de las técnicas de recuperación de información textual. En los últimos 30 años se han desarrollado estructuras de datos eficientes para el almacenamiento de índices, sofisticados algoritmos de interrogación, métodos de compresión (Frakes y Baeza-Yates 1992) e incluso hardware específico; más recientemente, se han aplicado técnicas de procesamiento del lenguaje natural en aspectos tales como la extracción de información, formulación de interrogaciones amigables y la generación de respuestas (Rijsbergen 1779). La búsqueda de cadenas tanto exacta como aproximada, los métodos de construcción y manipulación de diccionarios (Baeza-Yates y Ribeiro 1999).

Un SRI permite la recuperación de la información, previamente almacenada, por medio de consultas a los documentos contenidos en la base de datos. Esta serie de preguntas se conceptúan como sentencias formales de expresiones de necesidades de información, y suelen venir expresadas por medio de un lenguaje de interrogación. Un documento es un objeto de datos, de naturaleza textual generalmente, aunque la evolución tecnológica ha propiciado la profusión de documentos multimedia, incorporándose al texto fotografías, ilustraciones gráficas, vídeo, audio, etc.

Un SRI debe soportar una serie de operaciones básicas sobre los documentos almacenados en el mismo, como son: introducción de nuevos documentos, modificación de los documentos almacenados y eliminación de los mismos. Debemos contar algún método de localización de los documentos, para presentárselos posteriormente al usuario. Los SsRI implementan estas operaciones en formatos muy diversos lo que provoca una diversidad en lo relacionado con la naturaleza de los mismos.

La RI es el conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución de un problema planteado (Fernández 1998).

La diferencia entre recuperación de datos (RDs) e RI se puede caracterizar por:

Según la forma de responder a la pregunta: en RDs se utilizan preguntas altamente formalizadas, cuya respuesta es directamente la información deseada. En RI las preguntas resultan difíciles de trasladar a un lenguaje normalizado, y la respuesta es un conjunto de documentos que pueden contener, sólo probablemente lo deseado, con un evidente factor de indeterminación.

Según la relación entre el requerimiento al sistema y la satisfacción de usuario: en RDs la relación es determinística entre la pregunta y la satisfacción. En RI es probabilística, a causa del nivel de incertidumbre presente en la respuesta.

Según el criterio de éxito: en RDs el criterio a emplear es la corrección y la exactitud, mientras que en RI el único criterio de valor es la satisfacción del usuario, basada en un criterio personal de utilidad.

Según la rapidez de respuesta: en RDs depende del soporte físico y de la perfección del algoritmo de búsqueda y de los índices. En RI depende de las decisiones y acciones del usuario durante el proceso de interrogación.

Este documento en la sección 2 y 3, presenta las ideas básicas referentes a RI, el modelo espacio vectorial es descrito en la sección 4, en la sección 5 y 6 se especifican los módulos de indexación y consulta respectivamente. En la sección 7 se presentan unas ideas sobre realimentación de consultas y finalmente se determinan unas sugerencias sobre la proyección del documento.

ARQUITECTURA DE UN SRI

La RI, puede definirse como la representación, almacenamiento, organización y el acceso a elementos de información (Fernández 1998). El campo de RI envuelve un conjunto bastante grande de conceptos, estructuras y métodos. Para seguir un orden lógico se verán las fases en las que se ve involucrado el tratamiento de la información que se puede resumir en: el modelo conceptual, la indexación, la transformación de consultas, las operaciones sobre los términos y la gestión de documentos (Frakes y Baeza-Yates 1992, Fernández 1998, Baeza-Yates y Ribeiro 1999).

Figura 1. Arquitectura de un SRI.

Básicamente, los SsRI se apoyan en dos módulos: uno de indexación, que construye los vectores de los documentos, y otro de consulta, que calcula la similaridad con una consulta dada (Figuerola, Berrocal y Zazó 2000). Tanto los documentos como los vectores resultantes, así como productos intermedios y auxiliares, se almacenan en una base de datos relacional. A pesar de que los sistemas de gestión de bases de datos relacionales han padecido en el pasado un cierto descrédito en lo que se refiere a su utilización en entornos documentales, creemos que constituyen un medio idóneo para esta tarea. A las ventajas genéricas ya conocidas de este tipo de sistemas (prevención contra la redundancia y la inconsistencia, facilidad de modificación de estructuras, flexibilidad de manejo, estandarización, etc.), hay que añadir el hecho de que en la actualidad han superado algunos de los inconvenientes que tradicionalmente se les han achacado: posibilidad de campos de tamaño variable (los conocidos campos memo), posibilidad de almacenar datos binarios (imágenes, sonido, referencias a objetos externos), y, desde luego, campos repetibles (aunque siempre ha sido posible tener campos repetibles con una base de datos relacional; de hecho, ésta es una de sus razones de ser. Adicionalmente, sistemas de gestión de bases de datos relacionales (con sus lenguajes estándar de manipulación de datos e interrogación) corren hoy de forma ágil y sin problemas en ordenadores personales.

MODELO CONCEPTUAL DE LA RI

En principio, la recuperación de información engloba las acciones encaminadas a identificar, seleccionar y acceder a los recursos de información útiles al usuario, el objeto documental se ha organizado y representado, utilizando una serie de normas y convenciones, en un soporte informático, mediante el diseño, creación y mantenimiento de bases de datos.

Como los SsRI implementan una gama diversa de estructuras de datos, algoritmos y técnicas de recuperación de información, por ello precisamos de un modelo conceptual donde se determinan el tipo de ficheros, operaciones sobre los términos, modelos de búsqueda con base patrones exactos o los modelos inexactos los cuales contendrán las técnicas probabilísticas, los espacios vectoriales (Fernández 1998, Baeza-Yates y Ribeiro 1999). La mayoría de los sistemas de información son de dos tipo, booleanos y búsqueda de información por patrones de texto. Las interrogaciones a los sistemas de búsqueda de texto se llevan a cabo por medio de cadenas de caracteres o por expresiones regulares. Los sistemas de patrones de textos son más utilizados comúnmente en pequeñas colecciones de datos y cuando hay que gestionar grandes volúmenes de documentos se destacan mayoritariamente los sistemas booleanos. Dentro de un sistema booleano, los documentos se encuentran representados por conjuntos de palabras clave, generalmente almacenadas en un fichero inverso.

Se ha tratado de mejorar el rendimiento de los SsRI por medio de la distribución estadística de los términos, en tanto que la frecuencia de aparición de un término en un documento o conjunto de documentos podría considerarse relevante a la hora de establecer al similitud entre la consulta y el dato que identifica el documento. La distribución de frecuencia de un término se implementa dentro de algunos modelos estadísticos como es el caso del modelo espacio vectorial, o el modelo probabilístico (Mañas 1994, Fernández 1998).

Una decisión fundamental a tomar en el diseño de un SRI es que tipo de estructura de fichero se va a usar para la base de datos subyacente, podemos enunciar: ficheros planos, ficheros inversos, ficheros de patrones de bits, árboles PAT y grafos (Frakes y Baeza-Yates 1992, Fernández 1998, Baeza-Yates y Ribeiro 1999).

Las operaciones que se pueden llevar a cabo sobre los términos en un SRI se conforman en: extracción de raíces, truncamiento, ponderación de pesos, palabras vacías y tesauros (Frakes y Baeza-Yates 1992, Fernández 1998, Baeza-Yates y Ribeiro 1999).

Los documentos son los objetos primarios en un SRI, por ello se le debe asignar un identificador único, deben dividirse en sus campos constituyentes, y estos campos deben ser introducidos dentro de identificadores de campos y conjuntos de términos. Otra operación con los do-

cumentos es ordenar por un campo determinado mostrar incluye la opción de salida por impresora de los documentos como a su visualización por pantalla del ordenador. La distribución de frecuencias de los términos puede ser usada para agrupar documentos similares en un espacio documental, por medio de las técnicas de clustering.

MODELO DE ESPACIO VECTORIAL

Partiendo de que se puede representar los documentos como vectores de términos, entonces los documentos pueden situarse en un espacio vectorial de n dimensiones, es decir, con tantas dimensiones como elementos tenga el vector.

Matemáticamente hace tiempo que se trabaja con espacios de n dimensiones, dando a n un valor superior a tres. Situado en ese espacio vectorial, cada documento cae entonces en un lugar determinado por sus coordenadas, al igual que en un espacio de tres dimensiones cada objeto queda bien ubicado si se especifica sus tres coordenadas espaciales. Se crean así grupos de documentos que quedan próximos entre ellos a causa de las características de sus vectores. Estos grupos o clusters están formados, en teoría, por documentos similares, es decir, por grupos de documentos que son relevantes para la misma clase de problemas de información. Grupos de clusters pueden organizarse, a su vez, en torno a un centroide, que es un documento representativo de las propiedades medias de los documentos del cluster. En una base de datos documental organizada de esta manera, resulta muy rápido calcular qué centroide se parece más a una pregunta, y es muy rápida también la ordenación por relevancia, ya que, de forma natural, los documentos ya están agrupados por su grado de semejanza. En la fase de interrogación, cuando se formula una pregunta, también se la deja caer en este espacio vectorial, y así, aquellos documentos que queden más próximos a ella serán, en teoría, los más relevantes.

CARACTERIZACIÓN DEL MODELO VECTORIAL

Un documento d_i se modeliza como un vector $d_j = (w_{1,i}, ..., w_{t,i})$, donde $w_{i,j}$ es el peso del término tj en el documento d_i.

La similaridad entre un documento y la consulta es un valor entre cero y uno.

Una consulta se puede ver como un documento por lo tanto se puede ver como un vector.

La similaridad entre dos documentos se calcula mediante el valor del coseno entre los dos vectores.

Permite hacer «matches» parciales; ordena los resultados por grado de relevancia. No incorpora la noción de correlación entre términos (problema de todos los modelos clásicos).

MODULO DE INDEXACIÓN

En la indexación de los documentos, no todas las palabras o términos que los componen se incluyen en los índices. A los términos que se incluyen en el índice se les llama elementos de indexación. Además hay que considerar que dichos elementos pueden sufrir una serie de transformaciones antes de acabar en el índice.

PROCESADO DE DOCUMENTOS

Después de elegir el modelo conceptual de un SRI es necesario optar por un modelo de organizar la información de los documentos. Y esta decisión no es independiente del modelo conceptual considerado, ya que ciertos modelos conceptuales va a implicar el uso de una organización de la información determinada, y viceversa.

OBTENCIÓN DE PALABRAS DE CADA DOCUMENTO

Los documentos son los objetos primarios en un IRS y hay muchas operaciones para ellos. En algunos IRSs, a los documentos añadidos a una base de datos se les debe asignar un identificador único, deben dividirse (en partes gramaticales) en sus campos constituyentes, y estos campos deben ser introducidos dentro de identificadores de campos y conjuntos de términos. Una vez en la base de datos, uno a veces quiere desenmascarar ciertos campos para buscarlos y mostrarlos, por ejemplo, un investigador puede desear buscar sólo los campos de título y resumen de un documento para una búsqueda dada, o puede desear consultar sólo el título y el autor de los documentos recuperados. En algunos sistemas gestores de bases

de datos documentales, a este tipo de operación se le denomina búsqueda por referencia cualificada.

FILTRADO Y ELIMINACIÓN DE PALABRAS VACÍAS

La lista de palabras vacías es una relación de términos considerados como valores no indexables, usados para eli-minar potenciales términos de indexación. Los términos de una lista vacía están carentes de todo significado a la hora de recuperar información, como ejemplo se puede tomar el determinante «la», que no posee ninguna funcionalidad a la hora de recuperar documentos, ya que en todos los documentos de la base de datos aparecerá este término de forma casi segura y no resalta nada del contenido del documento almacenado. Así, cada término potencial de indexación es comprobado previamente, verificándose su presencia en la lista de palabras vacías y es descartado si se encuentra en ella.

NORMALIZACIÓN DE CARACTERES

Se refiere a una estandarización de los textos de entrada, de tal manera que los términos identificados servirán como entrada al algoritmo de indexación cuya finalidad es la obtención de una representación interna del documento que será la estructura de datos sobre la que se realizará realmente el proceso de búsqueda.

LEMATIZACIÓN

Los algoritmos de extracción de raíces de los términos, o de eliminación de sufijos, se encuentran orientados a obtener un único término a partir de diferentes palabras que constituyen esencialmente variaciones morfológicas con un mismo significado. El resultado del algoritmo debe ser una misma forma canónica para las diferentes variaciones morfológicas de una palabra, que no tiene por qué ser, necesariamente, la raíz lingüística.

ALMACENAMIENTO DE TÉRMINOS Y REFERENCIAS A DOCUMENTOS

En el caso de indexación de los documentos a tratar, no todas las palabras o términos que los componen se incluyen en los índices. A los términos que se incluyen en el índi-ce se les llaman elementos de indexación. Además hay que considerar que dichos elementos pueden sufrir una serie de transformaciones antes de acabar en el índice. Exis-ten los que consideran elementos de indexación a las palabras que aparecen en el texto, mientras que otros modelos toman como elementos de indexación aquellos correspondientes a entidades distintas de las palabras, como por ejemplo los sufijos del modelo PAT. Una vez que un elemento de indexación ha pasado el filtro de las palabras vacías, pueden ir directamente al índice, o ser indexado después de sufrir algún tipo de transformación dirigida para aumentar su representatividad y a homogeneizar la base de términos a incluir en las consultas.

CÁLCULO DE FRECUENCIAS DE TÉRMINOS, IDFS Y PESOS, MEDIANTE SENTENCIAS SQL, Y ALMACENAMIENTO DE RESULTADOS EN UNA TABLA

A los términos que se consideran representativos de los documentos se les puede asignar un valor numérico basado en su distribución estadística, o sea, la frecuencia con la que los términos aparecen en documentos, colecciones de documentos, o en subconjuntos de colecciones de documentos, tales como documentos considerados relevantes en una búsqueda.

Esta tarea consiste en dos etapas: la primera determina los términos que se van a considerar capaces de representar el contenido de un documento, y la segunda consistirá en asignar a cada término un peso o valor que refleje la importancia del término como representante del contenido del documento.

La primera de las etapas es más o menos inmediata, ya que se basa en la extracción de los términos que componen el texto de los documentos Lo que ya no resulta tan inmediato es la asignación de pesos a esos términos.

La mayoría de los intentos de indexación automática se basan en la observación de que la frecuencia de ocurrencia de un término en un documento tiene alguna relación con la importancia de ese término como representante del contenido del documento.

Se puede obtener el factor de relevancia de un término basándose en

las características de la frecuencia de las palabras de la colección de documentos. Con los siguientes criterios:

Dada una colección de n documentos, calcular para cada documento i la frecuencia de cada término k en ese documento. Determinar la frecuencia de cada término respecto a la colección completa, sumando las frecuencias de cada término en los n documentos. Ordenar las palabras en orden decreciente de frecuencia y eliminar todas aquellas que tienen un valor por encima de un umbral dado. Esto elimina las palabras muy frecuentes. Del mismo modo, eliminar las palabras poco frecuentes. Las palabras que quedan, con una frecuencia media, se utilizarán para caracterizar los documentos indexados.

La frecuencia de documento inversa (Inverse Document Frequency: IDF). Consiste en asumir que la importancia del término es proporcionar a la frecuencia de ocurrencia de cada término k en cada documento i, e inversamente proporcional al número de documentos a los que se asocia ese término, el valor de discriminación: Esta es una medida del grado en el que el uso de ese término va a ayudar a distinguir un documento de otro. Además provee de un método objetivo para determinar el umbral de frecuencia: los términos con alta frecuencia y un valor de discriminación negativo son pobres y no se deberían utilizar en la indexación; los términos con baja frecuencia con un valor de discriminación cero pueden o no ser utilizados y los términos que son buenos discriminantes, tienen un valor de discriminación positivo y deben ser considerados en la indexación, coincidiendo con aquellos de frecuencia intermedia.

A partir de la información procedente de la distribución de frecuencias de los términos, es posible asignar una probabilidad de relevancia a cada documento dentro de un conjunto recuperado, permitiendo que los documentos recuperados sean organizados en orden a esta probable relevancia. La información de la distribución de frecuencias de los términos puede ser usada para agrupar documentos similares en un espacio documental, por medio de las técnicas de clustering.

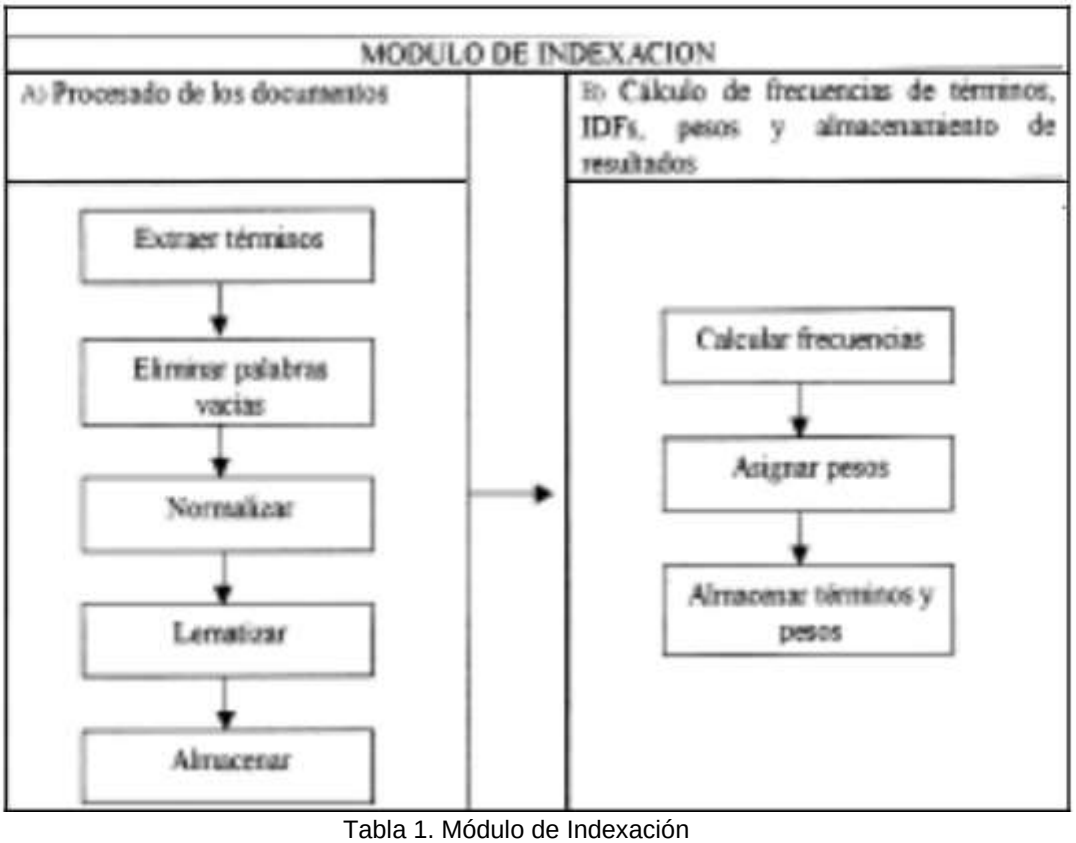

Tabla 1. Módulo de Indexación

MODULO DE CONSULTA

El módulo de consulta es aún más simple. Dado que una consulta en lenguaje natural ha de ser tratada como un documento cualquiera, requiere las mismas operaciones:

Obtención de palabras, eliminación de vacías, normalización de caracteres, lematización. Cálculo de pesos de los términos de la consulta, utilizando los datos de IDF almacenados en una tabla en la operación de indexado. Cálculo de similaridad entre consulta y cada uno de los documentos, mediante una simple sentencia SQL . Para realizar el cálculo de similaridad entre dos vectores existen diversas funciones, siendo las más conocidas la del producto escalar de dos vectores y los coeficientes del coseno.

Al hacer el cálculo del coeficiente de similaridad de los documentos y del vector de búsqueda, y someterlos a una comparación sistemática, se está en condiciones de establecer un orden descendente, colocando en primer término el documento cuyo valor es más cercano al del vector de búsqueda y así hasta concluir con todos los registros resultantes. Estos registros son los mismos que se obtienen al hacer un OR entre todos los términos que se utilizan en la interrogación.

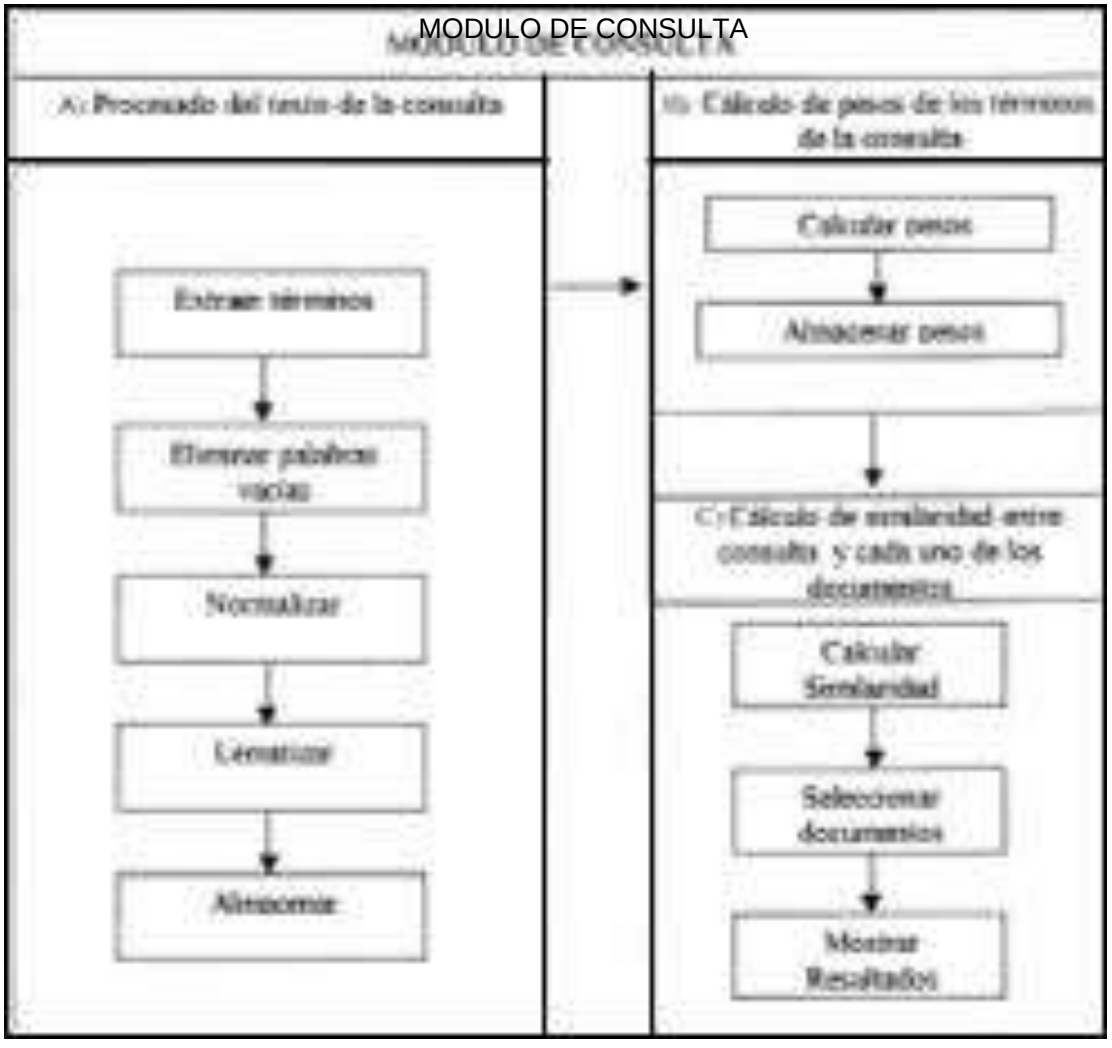

Tabla 2. Módulo de Consulta

REALIMENTACIÓN DE CONSULTAS

A partir de esta arquitectura, implementar realimentación de consultas es sencillo. Para una expansión simple, no hay más que añadir al texto original de la consulta el texto de los documentos con los que se desea realimentar, ya sea a través de selección por parte del usuario, ya sea tomando de forma automática los n primeros recuperados por la consulta original; tras esto, reejecutar la consulta con los añadidos hechos (Baeza-Yates y Ribeiro 1999).

Un enfoque algo más elaborado, puede requerir un re-ajuste de pesos de términos, en especial si se tienen en cuenta, entre los documentos recuperados por la consulta original, casos positivos y casos negativos; es decir, documentos recuperados que deben usarse para realimentar (ejemplos positivos), y documentos que, explícitamente, deben usarse como ejemplos negativos (es decir, que no se desean documentos como ésos). El recálculo posterior de pesos se efectúa mediante una simple sentencia SQL. Esto posibilita ajustar manualmente los coeficientes a aplicar (tanto negativos como positivos) sin necesidad de alterar el código del programa (Fernández 1998, Baeza-Yates y Ribeiro 1999).

CONCLUSIONES

Se ha mostrado la arquitectura y el diseño de un SRI con el modelo conceptual de espacio vectorial, lo suficientemente abierto y flexible para ser utilizado en labores docentes, así como de investigación. La sencillez de su arquitectura permitirá tan-to la fácil observación de resultados y estructuras intermedias como la modificación y añadido de nuevos módulos y, por consiguiente, la experimentación. De hecho, puede ser utilizado en la docencia de algunas materias relacionadas directamente con la recuperación automatizada de la información, y también en diversos trabajos de investigación, así, como para comprobar el efecto en la recuperación de nuevos sistemas de lematización para el castellano, teniendo un punto de comparación contra los que existen actualmente para el inglés.

BIBLIOGRAFIA

BAEZA-YATES R. y RIBEIRO N. Modern Information Re-trieval. Addison-Wesley. 1999.

FRAKES W. y BAEZA-YATES R. Information Retrieval: Data structures and algorithms. Englewood Cliffs: Prentice Hall. 1992.

FERNANDEZ LEAL J. Estudio Preliminar. 2: Modelos clásicos de recuperación de la información. Reporte técnico. Universidad de la Coruña. 1998

FIGUEROLA C; BERROCAL, José Luis A. y ZAZÓ Á. Diseño de un motor de recuperación de la información para uso experimental y educativo. Reporte técnico. Universidad de Salamanca. 2000

MAÑAS J- Búsqueda y Recuperación en la Internet. Reporte técnico. Universidad Politécnica de Madrid. 1994.

RIJSBERGEN C. J. Information Retrieval. London: Butterworths. 1979.