RISI **1**(1), 49-58 (2004)
Facultad de Ingeniería de Sistemas e Informática
Universidad Nacional Mayor de San Marcos
ISSN: 1815-0268 (impreso)

ESTUDIO Y EVALUACIÓN DE LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

Nora La Serna, Ulises Román, Norberto Osorio, Oscar Benito,
Jimy Espezúa, Hugo Vega*

RESUMEN

El trabajo que se presenta en este artículo se desarrolla en la línea de los Sistemas de Recuperación de Información (SRI). Básicamente se han realizado las siguientes actividades: 1) un estudio detallado de las principales técnicas, modelos y arquitecturas, así como de los criterios de evaluación de estos sistemas; 2) El estudio también ha llevado a un análisis de las técnicas de indexación, necesarios para el almacenamiento de los documentos; 3) Igualmente, el trabajo ha permitido la selección de cuatro aplicaciones de SRI para su análisis y evaluación: KARPANTA, SISA, DIALOG y SMART. El trabajo se desarrolla en el marco del proyecto de investigación «Sistema de Recuperación de Información», cuyo objetivo es diseñar un SRI para la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática y, posteriormente, para la Biblioteca Central de la Universidad Nacional Mayor de San Marcos.

Palabras clave: Sistemas de Recuperación de Información, Modelo Booleano, Modelo del Espacio Vectorial, buscadores WEB, evaluación de los SRI.

STUDY AND EVALUATION OF THE INFORMATION RETRIEVAL SYSTEMS

ABSTRACT

The research that is presented in this paper is developed in the Information Retrieval System (IRS) area. Basically, it has been done the following activities: 1) the study of the main techniques, models, and architecture, and the evaluation criterions of the systems 2) The indexing and searching for keeping and retrievaling documents, 3) Also, the work has let the study and evaluation of the following four Systems: KARPANTA, SISA, DIALOG and SMART. The research is developed inside of the «Information Retrieval System» project, whose main objective is to built a system for the Digital Library of the Systems and Informatics Engineering Faculty of greater National University of San Marcos.

Key words: Information Retrieval Systems, Boolean Model, Vector Model, Web search engines, Retrieval evaluation.

1. INTRODUCCIÓN

Los Sistemas de Recuperación de Información (SRI) permiten el almacenamiento óptimo de grandes volúmenes de información (principalmente documentos y últimamente también información multimedia) y la recuperación eficiente de la información ante las consultas de los usuarios. Este campo no es nuevo, pues, ha ido evolucionando desde la década de los años 50, cuando el objetivo era manejar información bibliográfica. Con el avance de la tecnología, computadores más potentes y software más eficientes, el almacenamiento de grandes volúmenes de información se está dando en todas las disciplinas del quehacer humano. Internet, la red de redes, también alberga en sus computadoras servidoras millones de documentos.

* Docentes de la Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima-Perú. E-mails: {nlasernap, uromanc, nosoriob, obenitop, jespezuac, hvegah}@unmsm.edu.pe

Por lo tanto, el cómo recuperar en forma eficiente los documentos almacenados en forma digital, que una persona necesita y solicita, es un tema no sólo de interés e importancia para la comunidad educativa (docentes, alumnos e investigadores), sino también para el sector empresarial el Gobierno y el público en general que necesita buscar información. Múltiples aplicaciones prácticas se están dando, algunas de las más conocidos son los buscadores web y bibliotecas digitales. El presente trabajo tiene dos objetivos principales: 1) Realizar un estudio detallado de las principales técnicas, modelos y arquitecturas, así como de los criterios de evaluación de los Sistemas de Recuperación de Información, y 2) Realizar un análisis y evaluación de cuatro aplicaciones de SRI que sobresalen en el medio. El resultado de los estudios que se realicen permitirá diseñar un Sistema de Recuperación de Información para la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática, y posteriormente para la Biblioteca Central de la Universidad Nacional Mayor de San Marcos.

La estructura del presente artículo es la siguiente: En la sección 2 se dan las definiciones más destacadas de los Sistemas de Recuperación de Información, en las secciones 3 y 4 se describen los dos modelos más utilizados en el diseño de los SRI: 1) El Modelo Booleano, y 2) El Modelo del Espacio Vectorial, en la sección 5 se presentan las principales técnicas de los SRI en la WEB, la sección 6 corresponde a la Evaluación de los SRI, en la 7 se hace un análisis de los resultados, y finalmente en la sección 8 se bosquejan las conclusiones y trabajos futuros.

II. DEFINICIONES DE SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

Desde el punto de vista de un Sistema de Información, podemos definir estos Sistemas como el conjunto de componentes (software, hardware, personas, procedimientos, datos, y comunicaciones) que interactúan, y cuyos objetivos son el almacenamiento óptimo de grandes volúmenes de información (principalmente documentos, últimamente también información multimedia) y la recuperación eficiente de información ante las consultas de los usuarios. Varias definiciones de los SRI se han dado desde su formalización en los años 1950, especialmente marcadas por el avance de la tecnología a través del tiempo, y también desde los puntos de vista de los autores, teniendo en cuenta que son multidisciplinarios, pues intervienen para su diseño generalmente la Bibliotecología, la Lingüística y las Ciencias de la Computación e Informática. Un resumen importante de las diferentes definiciones lo encontramos en Martínez M. Francisco (2000). Aquí presentamos algunas de las definiciones más destacadas.

Dos de los autores más citados por los especialistas en la materia son Gerard Salton y Ricardo Baeza-Yates. Este último autor –verdadera referencia en este campo donde ha venido preocupándose especialmente del tema de las estructuras de datos y de los métodos de acceso a los mismos– a la hora de definir la recuperación de información, en lugar de proponer una definición propia, hace uso de la elaborada por Salton: «La recuperación de la información tiene que ver con la representación, almacenamiento, organización y acceso a los item de información» [17].

Salton indica que, en principio, no deben existir limitaciones a la naturaleza del objeto informativo, y Baeza-Yates incorpora la reflexión siguiente: «La representación y organización debería proveer al usuario un fácil acceso a la información en la que se encuentre interesado. Desafortunadamente, la caracterización de la necesidad informativa de un usuario no es un problema sencillo de resolver» [1,2].

Algunos autores presentan la definición de Sistemas de Recuperación de Información como sinónimo de la Recuperación de Datos, influenciados por el punto de vista de las bases de datos; sin embargo, existen varias diferencias entre ambos términos. La tabla N.º 1 sintetiza las diferencias fundamentales entre ambos conceptos:

Tabla N.º 1. Diferencias entre recuperación de datos y recuperación de información.

	Recuperación de datos	Recuperación de información
Acierto (correspondencia)	Exacta	Parcial, la mejor
Inferencia	Algebraica	Inductiva
Modelo	Determinístico	Posibilístico
Lenguaje de consulta	Fuertemente estructurado	Estructurado o natural
Especificación de la consulta	Precisa	Imprecisa
Error en la respuesta	Sensible	Insensible

Los modelos de Sistemas de Recuperación de Información que se utilizan con mayor frecuencia en el diseño de los SRI son: 1) el Modelo Booleano, y 2) el Modelo del Espacio Vectorial. La descripción de ambos modelos son el tema de las siguientes dos secciones.

III. MODELO BOOLEANO

El Modelo Booleano es uno de los primeros modelos y el más utilizado de los SRI. En este modelo, un documento se encuentra representado por un conjunto de *palabras clave* (palabras con un valor semántico), las cuales pueden ser extraídas de un documento, de una parte de éste o de sus meta datos. Igualmente, la consulta es un grupo de palabras clave [17]. Generalmente se utilizan *archivos inversos* para almacenar las palabras clave.

Los archivos inversos contienen los siguientes campos: palabra clave o término índice (describe al documento), un identificador de documento (debe ser único para cada documento) y un identificador de campo (donde se encuentra la palabra clave) [12]. En un sistema booleano las consultas de los usuarios contienen operadores lógicos (Y, O, NO), y así un motor de búsqueda regresa aquellos documentos que cumplen con los aspectos lógicos de la consulta.

3.1. Arquitectura

En un SRI hay dos instancias: 1) el almacenamiento de los documentos, y 2) la recuperación de información desde la solicitud del usuario. En la figura N.º 1 se ilustran las dos instancias del proceso de almacenamiento y recuperación basado en el Modelo Booleano [11].

a) Desde el punto de vista del almacenamiento del documento en el SRI van a ocurrir los siguientes procesos:

1. A cada documento que entra se le asigna un Identificador.
2. Se identifican las palabras contenidas en el documento.
3. Se excluyen las palabras vacías.
4. Se «cortan» las palabras, es decir, se extraen las raíces de las palabras.
5. Se establece un peso de ponderación para cada raiz.
6. Finalmente las raíces debidamente ponderadas se introducen en la base de datos.

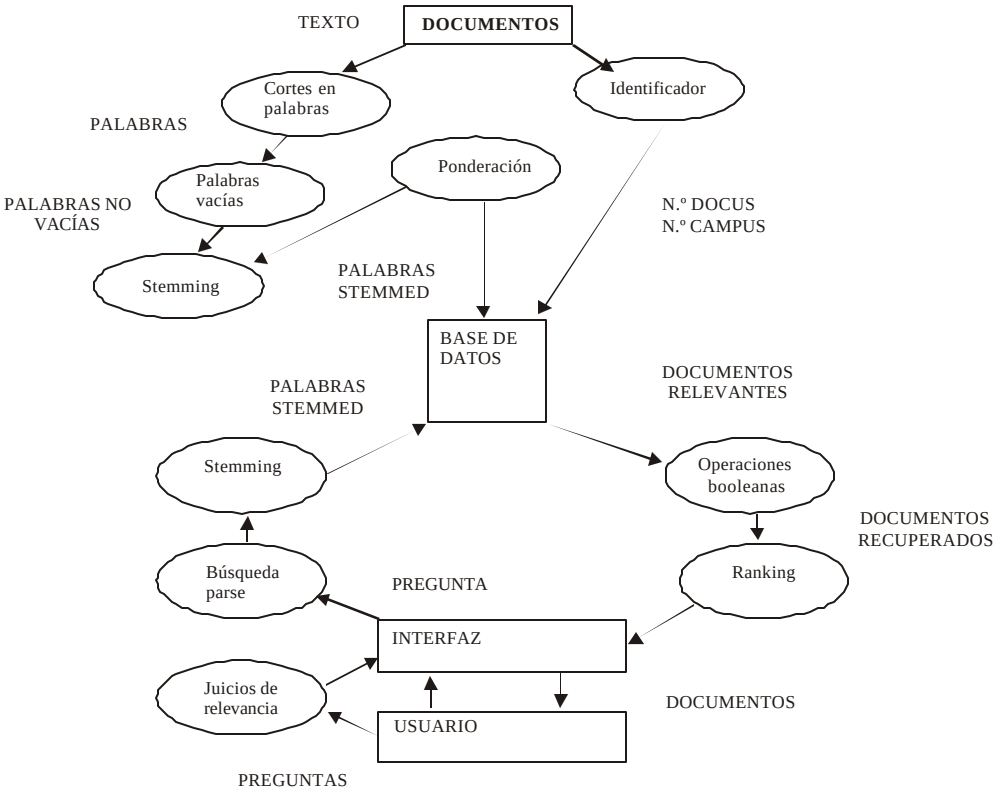

Figura N.º 1. Vista funcional del Modelo Booleano.

b) Cuando el usuario lleva a cabo una operación de recuperación de información, se realizarán los siguientes procesos:

1. El usuario, en función de sus necesidades y conveniencias, lleva a cabo una serie de juicios de relevancia para confeccionar su ecuación de búsqueda, ayudándose de las prestaciones que le proporciona el interfaz de búsqueda.
2. La ecuación de búsqueda, una vez introducida, se descompone en sus partes fundamentales.
3. Los términos clave empleados en la ecuación de búsqueda son «cortados» para extraer de ellos sus raíces y de esta forma proceder a su localización en la base de datos.
4. Una vez localizados los distintos subconjuntos de documentos asociados a los términos clave, se llevan a cabo las operaciones booleanas pertinentes, que han sido introducidas por el usuario en la ecuación de búsqueda.
5. Posteriormente, los documentos pueden alinearse para su presentación según un ranking determinado.

El Modelo Booleano da como resultado los documentos que parecen relevantes ante la consulta de un usuario, sin embargo, no dice qué tan relevante es un documento y así se asume que éstos tienen el mismo grado de importancia; es ahí donde radica una de las principales desventajas del modelo. Otra de las dificultades que presenta el Modelo Booleano tradicional es la dificultad en la elaboración de consultas, debido a que éstas están basadas en operadores booleanos y no todos los usuarios están familiarizados con esta forma de consulta. Otro problema radica en el poco control que hay sobre el tamaño de la salida producida por una consulta; esto ocasiona que se tenga una cantidad muy pobre de ellos. Adicionalmente, en el modelo booleano no hay provisiones para lograr una asignación de pesos a los términos, esto quiere decir que todos los términos son considerados siempre como de igual importancia.

Algunas de las desventajas descritas fueron eliminadas en el modelo vectorial, aunque éste no presenta la capacidad de formular consultas utilizando los diferentes operadores booleanos. Es a partir de esta falta que nace la idea de extender el modelo para obtener el Modelo Booleano Extendido.

IV. MODELO DEL ESPACIO VECTORIAL

Según este modelo, cada expresión del lenguaje natural puede representarse como un vector de pesos de términos, en donde un término es la unidad mínima de información, por ejemplo una palabra o la raíz sintáctica de una palabra. La asignación de pesos a los términos indica su presencia o importancia en el documento o en la colección de documentos. Habiendo varias técnicas para asignar pesos, una de ellas es la frecuencia del término, es decir, el número de veces que aparece el término en un documento. En el siguiente ejemplo se muestra la representación de un documento y una consulta mediante vectores de pesos:

Documento = (peso_de_término_l, peso_de_término_2, ..., peso_de_término_n)

Consulta = (peso_de_término_l, peso_de_término_2, ..., peso_de_término_n)

Para determinar la similitud que existe entre un documento y una consulta se calcula la distancia que existe entre los vectores que los representan; a menor distancia, mayor similitud. Para calcular esa distancia se aplica el Teorema del Coseno:

$$\text{COS}(\text{Vector X} * \text{Vector Y}) = \frac{\text{VECTOR X} * \text{VECTOR Y}}{|\text{VECTOR X}| * |\text{VECTOR Y}|}$$

Cuando el resultado de la aplicación de la fórmula anterior se aproxima a la unidad, quiere decir que los vectores son muy similares. Como acabamos de ver, calcular la similitud entre un documento y una consulta es tan fácil como calcular la distancia entre dos vectores. Sin embargo, esos vectores deben representar lo mejor posible tanto a los documentos como a la consulta [7].

4.1. Arquitectura

En la figura N.º 2 se presenta la gráfica de la vista funcional del modelo, en donde se realizan las siguientes tareas:

1. Se analizan los documentos y se transforman a una representación interna de cada uno.
2. Se analiza la consulta y se transforma a una representación interna.
3. A partir de las representaciones obtenidas en los pasos anteriores se calcula el grado de similitud entre cada documento y la consulta.
4. Se recuperan los documentos que guardan mayor similitud con la consulta del usuario.

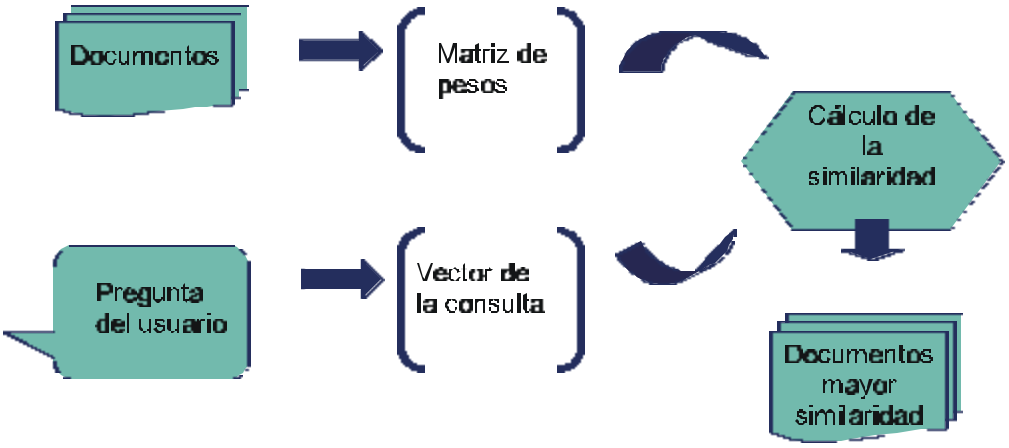

Figura N.º 2. Vista funcional del Modelo del Espacio Vectorial.

Los vectores están formados por «pesos de términos». El primer paso es escoger los términos. Por ejemplo, seleccionamos como términos cada una de las palabras en los siguientes documentos:

doc1 = «Sistemas de Recuperación de la Información»

doc2 = «Clasificación de los SRI»

doc3 = «Motores de búsqueda»

Los términos que aparecen en los documentos son: sistemas, de, recuperación, la, información, clasificación, los, SRI, motores, búsqueda.

El siguiente paso es asignar un 1 si el término aparece en el documento y un 0 si no aparece. La matriz de términos queda de la siguiente manera:

Luego, a cada término de cada uno de los documentos y al vector de la pregunta se le asignaría un peso. A continuación se calcularían las distancias del vector de la consulta con el vector de cada documento y se devolverían los documentos ordenados de mayor a menor similitud.

Procesos más detallados de una vista funcional del modelo seguirían al menos los siguientes pasos:

1. Eliminar signos de puntuación, etiquetas HTML, etc., dejando solamente las palabras de cada documento.
2. Aplicar listas de parada (listas con las palabras de uso más frecuente del idioma del texto, como artículos, preposiciones, etc.) para eliminar las palabras más habituales que aportan menos representatividad al documento.

	sistemas	de	recuperación	la	información	clasificación	los	SRI	motores	búsqueda
doc1	1	1	1	1	1	0	0	0	0	0
doc2	0	1	0	0	0	1	1	1	0	0
doc3	0	1	0	0	0	0	0	0	1	1

Si la consulta es «recuperación de información».

La representación de la consulta quedaría expresada como el siguiente vector:

	sistemas	de	recuperación	la	información	clasificación	los	SRI	motores	búsqueda
consulta	0	1	1	0	1	0	0	0	0	0

3. Aplicar extractores de raíces *(stemmers),* es decir, programas que reducen cada palabra a su raíz eliminando prefijos, sufijos, y terminaciones verbales.
4. Calcular el poder de discriminación de cada término, es decir, la capacidad de separar documentos consultando si tiene o no cada término.
5. Utilizar *thesauri* que agrupan los términos en un solo concepto por término; de esta manera todos los términos sinónimos se sustituyen por uno solo.
6. Para calcular el peso de cada término suelen realizarse cálculos basados en la frecuencia con que aparece cada término, tanto en un documento como en toda la colección.
7. Asignar a cada documento los pesos correspondientes a cada término.
8. Representar la consulta y calcular la similitud.
9. Ordenar y mostrar resultados.
10. Aplicar realimentación por relevancia, es decir, recoger información del usuario acerca de los resultados para que el sistema la aplique en sus cálculos.

V. SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN EN LA WEB

Internet, la red de redes, alberga en sus computadoras servidoras millones de documentos de información. Varias de las técnicas de almacenamiento y recuperación de información que se utilizan en los SRI tradicionales se han heredado de Internet. La mayoría de los sistemas de búsqueda en Internet utilizan el Modelo de Espacio Vectorial para el almacenamiento de los documentos; mientras que hay dos formas básicas de buscar información en la web, 1) los **motores de búsqueda**, y 2) los **directorios** [1].

Los **motores de búsqueda** (*search engine*) son sofisticadas aplicaciones que manejan grandes bases de datos de referencias a páginas web, recopiladas por medio de un proceso automático, es decir sin intervención humana. Uno o varios agentes de búsqueda (robots o crawlers) recorren la web. A partir de una dirección inicial de un documento extraen las direcciones de todos los documentos que estan referenciados por enlaces. De esta manera, los robots recopilan direcciones y generan etiquetas que permiten su indexación y almacenamiento en la base de datos. Avanzados algoritmos de búsqueda analizan las páginas que tienen en sus bases de datos y proporcionan el resultado más apropiado a una búsqueda. Los motores más populares son: Google, Altavista, Lycos, etc.

Los **directorios** son aplicaciones controladas por humanos que manejan grandes bases de datos que contienen direcciones de páginas, títulos, descripciones, etc. Las direcciones son clasificadas en subdirectorios de categorías temáticas. Las categorías presentan un listado de enlaces a las páginas referenciadas en el buscador. El directorio más grande y famoso es Yahoo.

5.1. El motor de búsqueda GOOGLE

GOOGLE es uno de los sistemas de recuperación en la web más utilizados, no sólo por la eficiencia en la búsqueda de información de los usuarios, sino tambien por el diseño de su arquitectura, el que es concebido para realizar un uso eficiente del espacio de almacenamiento y para proteger a los índices, de que se conviertan en un elemento lento y operativo. Este motor de búsqueda que fue desarrollado en la Universidad de Stanford en California, utiliza el Modelo del Espacio Vectorial para el proceso de almacenamiento y recuperación de la información [3].

El objetivo primordial del diseño de Google no es otro que mejorar estos índices de precisión en la recuperación de la información y, además, mejorar la presentación de los documentos recuperados de manera que, los primeros sean los más directamente relacionados con las necesidades de información planteadas por los usuarios.

Destacan dos grandes características en Google:

- En primer lugar, Google hace uso de la conectividad de la Web para calcular el grado de calidad de cada página. Esta graduación se denomina «**PageRank**» (coincide con el nombre del algoritmo de ranking empleado por este motor de búsqueda).
- En segundo lugar, Google utiliza esta propia capacidad de conexión de los documentos webs para mejorar los resultados de búsqueda.

El algoritmo PageRank (PR) asume que el número de enlaces que una página proporciona tiene mucho que ver con la calidad de la misma. PageRank puede ser pensado como un modelo del comportamiento del usuario. Otra justificación intuitiva de PageRank es que una página puede

tener un alto coeficiente de PageRank si existen muchas páginas que apuntan a ella, o si hay un número algo menor de páginas que apuntan a ella pero que posean, a su vez, un alto nivel de PageRank. De forma intuitiva, aquellas páginas muy citadas son páginas que vale la pena consultar y, en cambio, aquellas que sólo posean un enlace son páginas de poco interés para su consulta.

Se debe de recordar que el objetivo de la búsqueda no es otro que proporcionar una alta efectividad, y que lo primero que percibe el usuario es la precisión de los resultados de la búsqueda. El proceso de evaluación de la pregunta que lleva a cabo Google es el siguiente:

1. Descomposición *(parsing)* de la pregunta.
2. Conversión de las palabras a *wordIDS* (identificadores de palabras).
3. Localización de la posición de cada palabra en un *«barril de almacenamiento».*
4. Exploración de las listas de documentos hasta localizar un documento que contenga todos los términos de búsqueda.
5. Cálculo del rango de este documento para esta pregunta.
6. Una vez llegados al final del barril de almacenamiento, se vuelve al inicio repitiendo los pasos 4 y 5 para cada palabra de la ecuación de búsqueda.
7. Una vez calculados todos los rangos, se procede a ordenarlos de mayor a menor y presentarlos al usuario.

VI. EVALUACIÓN DE LOS SRI

Varias medidas han sido propuestas para evaluar a los SRI, sin embargo, dos de esas medidas son ampliamente utilizadas: la exhaustividad y la precisión. En ambos casos, la medida se basa en la relevancia de los documentos recuperados, es decir, en qué tanto se ha satisfecho la necesidad de información de los usuarios que hacen la consulta. Y aunque siempre se dice que la relevancia es un criterio subjetivo debido a que diferentes personas asignarían diferentes valores de relevancia a un documento, siempre se toma en cuenta en cualquier método de evaluación de los SRI.

La exhaustividad o *recall*, cuyo valor asociado se obtiene de dividir el número de documentos relevantes que satisfacen una consulta entre el total de documentos relevantes contenidos en la base de datos. Por ejemplo, suponiendo que en la base de datos existen 40 documentos relevantes para la consulta de un usuario y que el sistema de recuperación obtiene 20 documentos relevantes, por lo tanto la exhaustividad es de 20/40, es decir 50%.

La precisión se obtiene de dividir el número de documentos relevantes recuperados entre el número total de documentos recuperados. Por ejemplo, suponiendo que un SRI contiene 40 documentos relevantes que satisfacen una consulta dada, y el sistema de recuperación solamente obtiene 30 documentos, de los cuales sólo 20 son relevantes; entonces la presición del sistema es de 20/30, es decir 67%.

Los SRI tienden a maximizar la exhaustividad y la precisión de forma simultánea, para ello se han presentado diferentes métodos que han ayudado a que los sistemas actuales puedan atender las solicitudes de los usuarios cada vez en menos tiempo. Un método comprende el uso de grafos de exhaustividad-precisión, donde un eje es para la exhaustividad y otro para la precisión. Una medida de evaluación combinada de exhaustividad y precisión es la desarrollada por Van Rijsbergen (1979), que se define de la siguiente manera:

$$E = 1 - [(1 + b2)\ P\ R\ /\ (b2\ P + R)]$$

Donde P = precisión, R = exhaustividad o rellamada, y b es una medida de la importancia relativa para un usuario, de exhaustividad y precisión. Los investigadores eligen valores de E que ellos esperan que reflejarán la rellamada y precisión que interese al usuario típico. Por ejemplo, si los valores de b se encuentran en niveles de 0.50, esto nos indica que un usuario estuvo dos veces tan interesado en la precisión como en la rellamada, y si el valor de b fuera 2, nos indica que un usuario estuvo tan interesado en la rellamada como en la precisión [18].

Otros criterios de evaluación que se consideran son aquellos relacionados con la estructura de datos y algortimos de recuperación; éstos son: la eficacia en la ejecución y la eficiencia del almacenamiento. La eficacia en la ejecución es medida por el tiempo que toma un SRI para realizar una operación. Este parámetro es importante en un SRI, debido a que un largo tiempo de recuperación, interfiere con la utilidad del sistema, llegando a alejar a los usuarios del mismo si es lento.

La eficiencia del almacenamiento es medida por el número de bytes que se precisan para almacenar los datos. El espacio general, una medi-

da común para medir la eficacia del almacenamiento, es la razón del tamaño del índice de los archivos más el tamaño de los archivos del documento sobre el tamaño de los archivos del documento. Los valores del espacio general que oscilan entre los valores 1,5 y 3 son típicos de los SRI basados en los archivos inversos.

Adicionalmente, Lancaster (1973) propuso que los criterios para la evaluación de los SRI deberían estar basados en los siguientes factores: 1) cobertura o alcance, 2) exhaustividad, 3) precisión, 4) tiempo de respuesta, 5) esfuerzo del usuario, y 6) formato de presentación [10].

VII. RESULTADOS

El presente trabajo nos ha permitido hacer una exhaustiva revisión de los SRI. Las funciones más importantes en estos sistemas son:

a) El almacenamiento óptimo de grandes volúmenes de información (principalmente documentos, últimamente también información multimedia).

b) La recuperación eficiente de información ante las consultas de los usuarios.

Dos son los modelos más utilizados en su diseño: a) el Modelo Booleano, y b) el Modelo de Espacio Vectorial.

- El Modelo Booleano está caracterizado por la utilización de palabras clave y tablas de índices para el almacenamiento y recuperación de la información; así como también por el uso de operadores lógicos para las consultas de los usuarios. En el proceso de recuperación de un documento, el criterio de *relevancia* prima para la selección de un documento. Para ello, varias técnicas estadísticas han sido implementadas para determinar la relevancia de un documento.
- Según el Modelo de Espacio Vectorial cada documento se registra en un vector de términos, y una colección de documentos forma una matríz de términos en donde un término es la unidad mínima de información, por ejemplo una palabra. Para medir la importancia de un término en un documento se asignan pesos a cada uno de los términos. El modelo establece ciertos criterios de similitud para comparar qué tan parecidos son dos términos o dos documentos. Un criterio para determinar la similitud que existe entre un documento y una consulta es calcular la distancia que existen entre los vectores que los representan.

Los sistemas de recuperación en la web utilizan generalmente el Modelo de Espacio vectorial para el almacenamiento de los documentos. Dos formas básicas de buscar información en la web son los motores de búsqueda y los directorios. Ambas formas manejan grandes bases de datos que contienen principalmente direcciones e información de páginas.

- Los motores de búsquedas son sofistificados programas que realizan la búsqueda de información en la web de forma automática, mediante los robots de búsqueda.
- Los directorios son aplicaciones controladas por humanos que manejan subdirectorios de categorías temáticas con enlaces a páginas referenciadas.

El estudio de los SRI también nos permite plantear el análisis y evaluación de cuatro sistemas de recuperación que sobresalen: KARPANTA, SISA, DIALOG y SMART.

- KARPANTA es un SRI basado en el Modelo de Espacio Vectorial desarrollado en la Universidad de Salamanca, España [6]. En la figura N.º 3 se observa el proceso de indización en el sistema KARPANTA.
- SISA, (Sistema para la indización Semiautomática) es un sistema de indexación desarrollado en la Universidad Politécnica de Valencia, España [8].
- DIALOG, es un SRI web comercial para distribución de información electrónica.
- SMART es un sistema de análisis automático y de recuperación de textos, uno de los sistemas pioneros de los SRI [16].

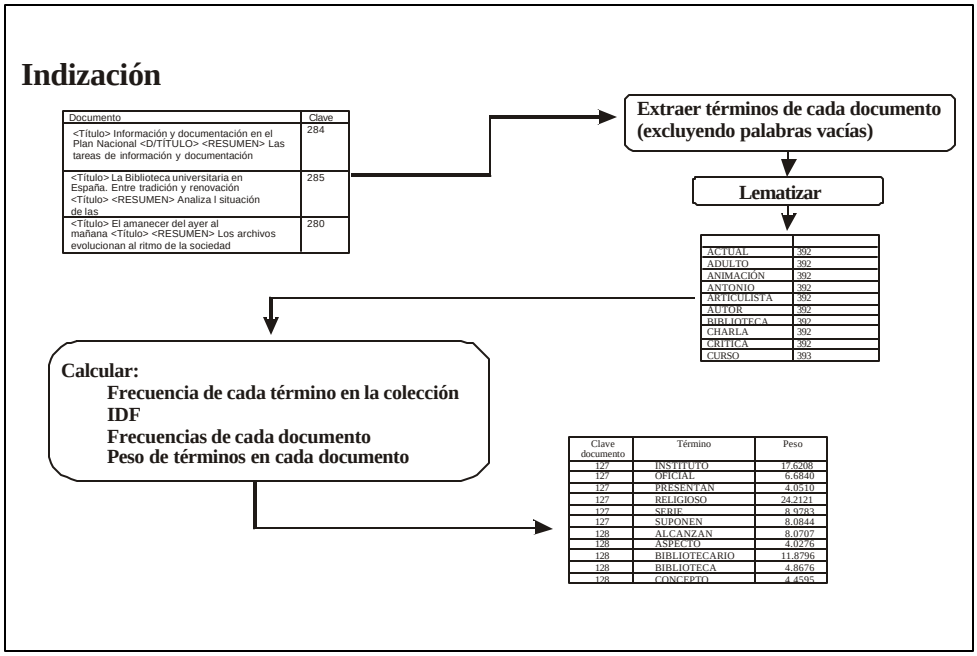

Figura N.º 3. Proceso de Indización en el sistema KARPANTA.

VIII. CONCLUSIONES Y TRABAJOS FUTUROS

La investigación ha dado lugar al estudio detallado de la evolución, técnicas de almacenamiento y recuperación, así como de los criterios de evaluación de los Sistemas de Recuperación de Información. El estudio también ha llevado a un análisis de las técnicas de indexación necesarios para el almacenamiento de los documentos.

El trabajo ha permitido la selección de cuatro sistemas de recuperación desarrollados para su estudio y evaluación: KARPANTA, SISA, DOMAIN y SMART. El resultado de este estudio permitirá presentar las bondades y limitaciones de cada uno de ellos, y seleccionar el diseño más adecuado, de acuerdo a nuestras necesidades, de un Sistema de Recuperación de Información para la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática, y posteriormente para la Biblioteca Central de la Universidad Nacional Mayor de San Marcos.

IX. BIBLIOGRAFÍA

1. Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval***.** Maryland: Addison-Wesley-Longman Publishing co, 1999.
2. Baeza-Yates R.y Davis Emilio. *Ranking Global de Páginas Web Basado en Atributos de los Enlaces*; CLEI 2004, 8 pp.
3. Brin, S. and Page, L. *The anatomy of a large-scale hypertextual Web search engine*. Computer Netsorks and ISDN Systems, 30, 1998. pp. 107-117.
4. Chu, H. and Rosenthal, M. *Search engines for the WWW: A comparative study and evaluation methodology* En http://www.asis.org/annual-96/ElectronicProceedings/chu.html
5. Delgado Domínguez *Mecanismos de recuperación de Información en la www*», Universidad de Islas Baliares, España. 1998. http://dmi.uib.es/people/adelaida/tice/modul6/memfin.pdf
6. Figuerola C., Alonso J., y Zazo A. *Diseño de un motor de recuperación de la información para uso experimental y educativo*. BID Num. 4 junio 2000.
7. Frakes W.B. y Baeza Yates R. *Information Retrieval: data structures and algorithms*. Prentice Hall, 1998.
8. Gil-Leiva I. *Sistema para la Indización Semiautomática de Artículos de Revista sobre Biblioteconomía y Documentación (SISA*), II Jornadas sobre Tratamiento y Recuperación de Información, Madrid (Leganés), Septiembre 2003.
9. Lancaster, F. W. & Warner, A.J. *Information retrieval today*. Arlington, VA: Information Resources. 1973.

10. Martínez M.F. y Rodríguez M. J. *Síntesis y crítica de las evaluaciones de la efectividad de los motores de búsqueda en la Web*. 2003. http://InformationR.net/ir/8-2/paper148.html

11. Martínez Méndez Francisco Javier. *Sistemas de Almacenamiento y Recuperación de Información*. http://www.um.es/gtiweb/fjmm/sari2000.htm 2000.

12. Medina Nieto, María Auxilio. Tesis de Maestría, *EGRAI: Espacio Grupal con Referencistas y Agentes como apoyo a la Investigación*, http://info.pue.udlap.mx/~tesis/msp/

13. Notess, G.R. *Search engine statistics*. Bozeman, MT: Notess.com. http://www.searchengineshowdown.com/stats/ 2002

14. Prieto-Díaz, R. and ARANGO, G. *Domain Analysys: Acquisition of Reusable Information for Software Construction*. New York: IEEE Press, 1991.

15. Salton G. *The SMART system*. Encyclopedy of Library and Information Science 1980.

16. Salton G. Y McGill M. *Introduction to Modern Information Retrieval*. Mc. Graw-Hill. 1983.

17. Van Rijsbergen, C.J. *Information Retrieval*. London: Butterworths, 1979.

18. Zhang, D. and Dong, Y. *An efficient algorithm to rank web resources.* En http://www9.org/w9cdrom/251/251.html

19. DIALOG. www.dialog.com

20. SearchEngineWatch.com *The major search engines Jupitermedia Corporation*»

. http://www.searchenginewatch.com/links/major.html. 2002