RISI 3 (4) (2006)
Revista de Investigación de Sistemas e Informática
Facultad de Ingeniería de Sistemas e Informática
Universidad Nacional Mayor de San Marcos
ISSN: 1815-0268 (versión impresa) / ISSN: 1816-3823 (versión electrónica)

Criterios de Evaluación en los sistemas de Recuperación de Información

Nora La Serna Palomino [1], Cynthia Bautista Almeza [2]

Resumen El trabajo que se presenta en este artículo se desarrolla en el área de los Sistemas de Recuperación de Información (SRI). Fundamentalmente, se han realizado las siguientes actividades: 1) Presentación de la Arquitectura del Sistema de Recuperación de la Información que utilizará información de la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática; 2) Revisión de las investigaciones realizadas acerca de los Criterios de Evaluación en los SRIs; 3) Pruebas aplicando indicadores de evaluación al sistema Karpanta [6], y el análisis de los resultados. El trabajo se desarrolla en el marco del proyecto de investigación "Implementación de un Sistema de recuperación de Información para la Biblioteca de la FISI", cuyo objetivo es diseñar un SRI para la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática.

Palabras clave: sistemas de recuperación de información, modelo del espacio vectorial, criterios de evaluación, precisión y exhaustividad, karpanta.

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) Presentation of the architecture of the Digital Library of the Systems and Informatics Engineering Faculty, 2) The study of the main evaluation methodology and criteria for Information Retrieval System, 3) Tests applying evaluation criteria to the Karpanta System [6]. 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 The Universidad Nacional Mayor de San Marcos.

Key words: Information Retrieval Systems, Vector Model, evaluation criteria, precision and recall, Karpanta.

1 El presente trabajo se desarrolla en el marco del proyecto de investigación "Implementación de un Sistema de Recuperación de Información para la Biblioteca de la FISI" RR. N. ° 01766-R-06 del 10-04-2006.

2 Alumna del curso Seminario de Computación del semestre 2006-1 de la FISI, trabajo desarrollado como parte de su Proyecto de Tesis.

1. Introducción

El presente artículo forma parte del trabajo de investigación que se desarrolla y que tiene como objetivo principal implementar un sistema de almacenamiento y recuperación de información para un sistema de bibliotecas. El sistema que será implementado utilizará información de la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática, para realizar búsquedas de información como libros, tesis de pregrado, revistas, etc.

Acerca del diseño del Sistema de Recuperación de Información para la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática (FISI) de la UNMSM ha sido ampliamente descrito en publicaciones anteriores [9,10], sin embargo es importante destacar que se utilizan las técnicas definidas en el modelo de espacio vectorial [17], el que es uno de los modelos más utilizados en estos Sistemas.

Con el avance de la tecnología, computadores más potentes y software más eficientes, el almacenamiento de grandes volúmenes de información se esta dando en todas las disciplinas del quehacer humano. Internet, la red de redes, también alberga en sus computadores servidoras millones de documentos. Por lo tanto, cómo recuperar en forma eficiente documentos almacenados en formato 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, gobierno y público en general que necesita buscar información. Múltiples aplicaciones prácticas se están dando, algunos de los más conocidos son los buscadores Web y las bibliotecas digitales.

Varias medidas han sido propuestas para evaluar a los SRIs, sin embargo dos de esas medidas son ampliamente aplicadas: la exhaustividad y la precisión. En ambos casos, la medida se basa en la relevancia de los documentos recuperados; es decir, que tanto se

Ha satisfecho la necesidad de información de los usuarios y quiénes hacen la consulta. Otros criterios de evaluación que se consideran son aquellos relacionados con la estructura de datos y algoritmos de recuperación, son: La eficacia en la ejecución y la eficiencia del almacenamiento [1,2].

Específicamente, en los Sistemas de Recuperación de información en la Web, se plantean criterios bastante relacionados a los mencionados anteriormente, pero además se tienen en cuenta otros criterios básicamente caracterizados por el enorme tamaño de la web, por su estructura hipertexto y por su arquitectura distribuida. Además de los criterios precisión y exhaustividad, se incluyen: 1) Tiempo de respuesta, 2) Capacidad de búsqueda, y 3) Documentación e interfase.

Muchos estudios se han realizado acerca de las medidas para evaluar los SRIs, uno de los trabajos más destacados es la Tesis Doctoral realizado por Don Francisco Javier Martínez Méndez en el año 2002, Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en Internet [11]. Quien recoge de manera muy ordenada la mayor parte de los criterios de evaluación presentados por varios autores, y propone un planteamiento y desarrollo del procedimiento de evaluación en los Sistemas de recuperación de información basados en la Web.

En el presente artículo se recogen las dos medidas de evaluación más utilizadas: Precisión y exhaustividad, y con ellas se lleva a cabo la evaluación de Karpanta [6], el que es un motor de búsqueda y recuperación de información que adopta el modelo de espacio vectorial [17]; además es diseñado para entornos documentales y para su aplicación en la docencia y la investigación.

La estructura del presente artículo es la siguiente: En la sección 2 se presenta un bosquejo de la Arquitectura del SRI para la biblioteca de la FISI, en 3 se describen los criterios de evaluación que más destacan en el área, en 4 se explica de forma muy sucinta el

Sistema Karpanta que servirá para realizar las pruebas de evaluación. La sección 5 ilustra las pruebas realizadas, en 6 se realiza un análisis de los resultados de las pruebas y de los criterios de evaluación empleados, y finalmente en la sección 7 se presentan las conclusiones y cómo influirá en el sistema propuesto.

2. Arquitectura de un Sistema de Recuperación de la Información

Para el desarrollo del Sistema propuesto que utilizará información de la Biblioteca de la FISI, se ha seguido el modelo de espacio vectorial [17]. Según el modelo, cada documento es representado mediante un vector de n términos, en donde un término es la unidad mínima de información, por ejemplo una palabra o la raíz sintáctica de una palabra. Dado de que cada término puede ser más o menos significativo en un documento o en toda la colección de documentos, a cada término se le asigna un peso, de esta manera se mide la importancia de un término para distinguir un

documento de la colección. En la siguiente sección se describe el cálculo de la matriz de pesos para el sistema propuesto.

Siguiendo al modelo del espacio vectorial, las consultas de los usuarios también son representadas mediante un vector de términos, los términos deben ser los mismos que el de la colección de documentos. Y a la vez, se determina el peso de los términos de la consulta. De esta manera, la consulta del usuario y cada uno de los documentos están expresados en vector y matriz de pesos de términos respectivamente, lo cual mediante un cálculo de similaridad permite hallar aquellos documentos que se aproximan más a la pregunta del usuario.

En la Figura 1, se presenta la arquitectura del sistema propuesto. Si bien los documentos que se manejan en la biblioteca de la Facultad, principalmente son libros, revistas especializadas, tesis de pregrado, etc., los documentos en forma digital disponible en CD´s, son las tesinas de pregrado.

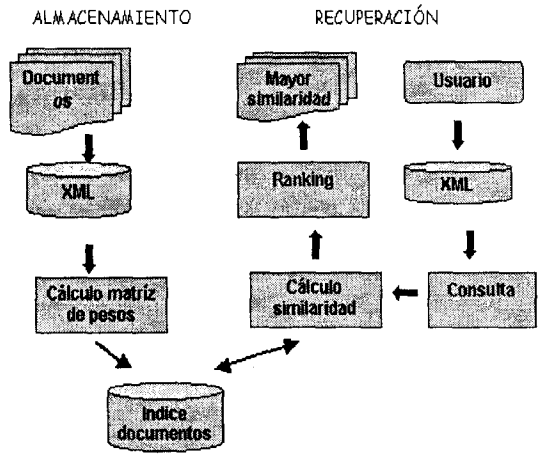

Figura 1. Arquitectura del Sistema propuesto

Hay alrededor de docientos tesis enformato digital, los que serán utilizados como los documentos iniciales para el sistema. Éstos documentos serán convertidos al formato XML (Extensible Markup Languaje).

Los documentos codificados en XML permiten etiquetar e identificar el contenido de los documentos, es decir, define los elementos de cada documento, y cómo tienen que estar organizados dentro del documento. De esta manera el sistema de búsqueda de información será más rápido y eficiente, además que facilitará el intercambio de información y la cooperación con otros sistemas de la Facultad.

Adicionalmente, en el desarrollo del sistema se plantea utilizar como herramientas de desarrollo software libre y tecnologías de información emergente como aplicaciones Web basadas en el Modelo Vista Controlador (MVC). Se codificará en lenguaje Java, y Sistemas de gestión de bases de datos MySql.

3. Criterios de Evaluación en los SRIs

Varias medidas han sido propuestas para evaluar a los SRI, sin embargo, dos de esas medidas son ampliamente conocidas: la exhaustividad y la precisión. En ambos casos, la medida se basa en la relevancia de los documentos recuperados, es decir, qué tanto se ha satisfecho la necesidad de información de los usuarios quienes 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 [17,18].

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 una 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. Una medida de evaluación combinada de exhaustividad y precisión es la desarrollada por [18], 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, 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.

Otros criterios de evaluación que se consideran, aquellos relacionados con la estructura de datos y algoritmos de recuperación, 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 bites que se precisan para almacenar los datos. El espacio general, una medida común de 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.

Específicamente, en los Sistemas de Recuperación de información en la Web, se plantean criterios bastante relacionados a los mencionados anteriormente, pero además se tienen en cuenta otros criterios básicamente caracterizados por el enorme tamaño de la Web, por su estructura hipertexto y por su arquitectura distribuida. En resumen, Lancaster; Chu y Rosental 4 proponen los siguientes criterios para la evaluación de esos motores de búsqueda:

Tiempo de respuesta, es el tiempo total que utilizan los sistemas para dar respuesta a una solicitud del usuario.

Precisión, como se mencionó anteriormente, se tienen en cuenta el número de documentos relevantes recuperados, entre el número total de documentos recuperados.

Capacidad de búsqueda, son las diferentes formas en que los motores presentan a los usuarios para que ellos realicen sus búsquedas. Esto incluyen: operadores booleanos, expresiones literales, y posibilidades para acotar la búsqueda en un determinado campo.

Documentación e interfase, esto es fundamental porque cuanto menor esfuerzo realice el usuario para una búsqueda, y mayor confianza en el sistema, los usuarios podrán tener mayor preferencia para usar la herramienta.

Adicionalmente, los autores en [8] 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.

4. Karpanta un SRI didáctico y experimental

Karpanta [6], es un motor de búsqueda y recuperación de información que adopta el modelo de espacio vectorial [17]; es diseñado para entornos documentales y para su aplicación en docencia e investigación.

Presenta dos Módulos básicos: 1) **Indización**, y 2) **Consulta**. En el proceso de Indización, construye una matriz de vectores de los documentos; en el proceso de **Consulta**, se calcula la similaridad entre la pregunta que se genera a partir de una consulta dada, y los documentos que se encuentran almacenados, para finalmente presentar los resultados de acuerdo a los valores obtenidos de la similaridad.

El proceso de **indización** a su vez consta de etapas: 1) Procesado de los documentos, y 2) Cálculo de los indicadores. En el primer caso, se realizan las siguientes actividades: la obtención de palabras de cada documento, filtrado y eliminación de palabras vacías, normalización de caracteres (mayúsculas, minúsculas, acentos), le matización (se aplica un s-stemmer), y se almacenan en una tabla cada término resultante, junto con la referencia o clave de los documentos en que aparece.

Durante la etapa del Cálculo de indicadores, se llevan a cabo las siguientes operaciones: hallar las frecuencias de cada término en la colección de documentos, determinar el IDF (Inverse Document frecuency), hallar la frecuencia en cada documento y calcular el peso de términos en cada documento.

El resultado del proceso de indización, es un archivo invertido en donde los documentos se almacenan en una tabla con los siguientes campos: término (palabra que aparece en cada documento), frecuencia (número de veces en que aparece un término en un documento), peso (de términos en cada documento), y clave (identificación del documento). Ver Figura 2.

doc_term : Tabla

termino	frecuencia	peso	clave
ACTIVIDADES	2	8,919612	1
AFINES	1	6,68443	1
AUTONOMA	3	17,30442	1
BIBLIOGRAFIA	1	8,070724	1
BIBLIOTECA	1	2,98932	1
BREVE	1	4,89267	1
CADA	1	3,95985	1
CIENCIAS	3	14,43788	1
CIENTIFICA	4	18,95408	1
CITAN	1	6,972112	1
CLASIFICACIO	1	5,126285	1
COLABORA	1	6,972112	1
COLABORACIC	1	4,851848	1
CONOCER	1	4,93523	1
CONSULTADA	1	8,070724	1
CREACION	1	4,08174	1
DA	3	15,54106	1
DEPARTAMEN	3	18,37444	1
DETALLAN	1	6,972112	1
DOCUMENTAC	6	18,40067	1
DOCUMENTAR	1	7,377577	1
EMPLEADOS	1	6,278965	1

Registro: 1 de 36925

Figura 2. Tabla de términos de los documentos de la colección.

En el **proceso de consulta**, se requieren de las siguientes operaciones: 1. Procesado del texto de la consulta (obtención de palabras, eliminación de palabras vacías, normalización de caracteres, lematización), 2. Determinar los pesos de los términos de la consulta, utilizando los datos del IDF obtenidos en la operación de indización, 3. Cálculo de similaridad entre la consulta y cada uno de los documentos, y 4. Presentación de los resultados. Ver Figura 3.

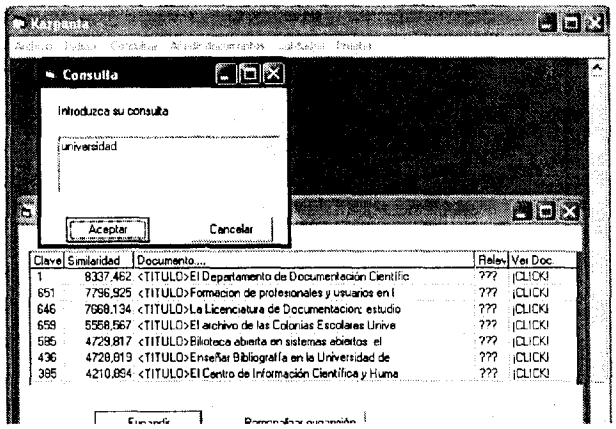

Figura 3. Proceso de consulta.

En el desarrollo del prototipo se han utilizado bases de datos relacionales, y el lenguaje SQL ( ) para el acceso a los datos durante el proceso de cálculo de los indicadores y durante el proceso de la consulta. Los documentos registrados corresponden a referencias bibliográficas y hay almacenados 1177 documentos.

5. Pruebas de Evaluación utilizando Karpanta

5.1 Preparación de los datos y selección de los indicadores

Para efectos de la simulación se seleccionaron 10 frases para realizar búsquedas con esas frases en el sistema karpanta. En los casos en que se consideró necesario, los términos fueron reemplazados, para efectos de la búsqueda, por un término natural más próximo al que utilizaría un usuario generalmente. En la tabla 1, se pueden observar en la columna Búsqueda, los términos empleados para las pruebas.

- Las variables analizadas fueron las siguientes:
  - Precisión, definida como el número de documentos relevantes recuperados sobre el total de documentos recuperados, se expresa así:

$$\text{Precisión} = \frac{\text{Documentos relevantes recuperados}}{\text{Documentos recuperados}}$$

  - Exhaustividad, es la proporción de material relevante recuperado del total de los documentos que son relevantes en la base de datos, independientemente de que éstos, se recuperen o no.

$$\text{Exhaustividad} = \frac{\text{Documentos relevantes recuperados}}{\text{Documentos relevantes}}$$

5.2 Experimentos realizados

Para evaluar el criterio de precisión se tiene en cuenta el número de documentos relevantes recuperados, entre el número total de documentos recuperados. En la tabla 1, para cada una de las frases de consulta al sistema, se presenta esta información, que además permitirá calcular el porcentaje de documentos relevantes en cada consulta.

N°	Búsqueda	Documentos Recuperados	Documentos Relevantes Recuperados	Documentos No Relevantes Recuperados	Documentos Relevantes No Recuperados
1	Lenguajes de Programación	15	9	6	6
2	Motores de búsqueda	15	10	5	1
3	Telefonía Celular	3	2	1	0
4	Literatura Infantil	14	12	2	2
5	Tecnologia Digital	15	10	5	3
6	Base de Datos	15	11	4	1
7	Recuperación de Información	15	10	5	2
8	Sistema Operativo	15	8	7	6
9	Cuento para niños	15	7	8	1
10	Biblioteca Universitaria	15	10	5	7

Tabla 1. Resultados por tema de búsqueda

Precisión

- El promedio de la tasa de precisión para el grupo de búsquedas fue de 65.24%, es decir, aproximadamente el 65% de todos los documentos recuperados fueron juzgados como relevantes. Esta cifra se basó en 10 búsquedas realizadas, en donde se encontró una búsqueda de

documentos recuperados mayores a 0. Ver Figura 4.

- La búsqueda identificado como 9 (cuento para niños) registró el menor grado de precisión obteniendo un 46.66%, esto es, 7 de 15 documentos recuperados.

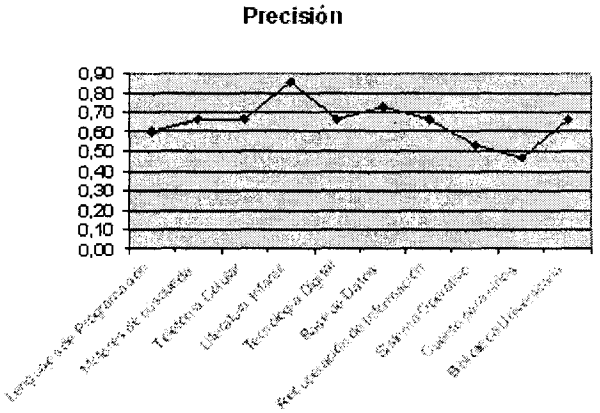

Figura 4. Precisión a partir de los datos de prueba

Exhaustividad

- El promedio de la tasa de exhaustividad para el grupo de búsquedas fue de 79.20%, es decir, aproximadamente el 79% de todos los documentos recuperados relevantes fueron recuperados. Esta cifra se basó en 10 búsquedas obtenidas, donde se encontró una búsqueda documentos recuperados mayores a 0. Ver Figura 5.
- La búsqueda 8 (sistema operativo) registró el menor grado de precisión obteniendo un 57.14%, esto es, 8 documentos de los 14 documentos relevantes fueron recuperados.

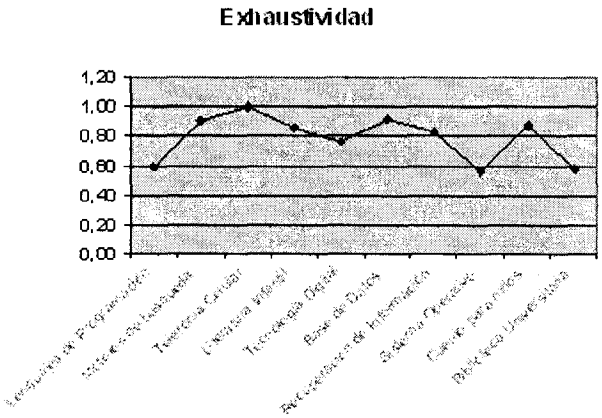

Figura 5. Exhaustividad a partir de las pruebas

6. Análisis de los resultados de las pruebas de evaluación

La base de datos contiene 1177 documentos, de los cuales se han recuperado en total 137 documentos en las 10 consultas realizadas al sistema. Después de las pruebas realizadas utilizando los criterios de evaluación: precisión y exhaustividad aplicados al motor de búsqueda documental Karpanta, se puede observar que el sistema posee una tasa de exhaustividad promedio de 0.65 y una tasa de precisión de 0.79. Lo cual es un resultado bastante positivo en éstos sistemas, lo que permite afirmar que Karpanta es un Sistema muy eficiente y óptimo.
Tomando como base la información teórica para el desarrollo de este estudio se afirma que la tasa exahustividad del sistema se encuentra entre el rango óptimo [0.6, 0.8]. Además la tasa de precisión se encuentra entre el rango óptimo [0.2, 0.8].

7. Conclusiones

- Las evaluaciones de los SRI se encuentran estrechamente vinculadas con la investigación y el desarrollo de la recuperación de información es la propia naturaleza de estos sistemas la que propicia esta necesidad crítica de evaluación, justo como ocurre en cualquier otro campo de trabajo que aspire a ser calificado y certificado como científico.
- Muchos trabajos se han centrado en analizar la efectividad del acceso físico a los datos contenidos en un SRI, cuando lo que verdaderamente cobra importancia hoy en día es analizar el comportamiento del acceso lógico a los datos, esto es, el contenido informativo de la recuperación.
- Aunque se ha desarrollado una respetable cantidad de medidas destinadas a llevar a cabo estas evaluaciones, la mayor parte de los estudios continúan empleando los juicios basados en la relevancia de los documentos para valorar esta efectividad. En la práctica, resulta más adecuado hablar de pertinencia de los documentos que de relevancia, en tanto que el primero de estos términos incorpora un sentido de utilidad para el usuario final. No obstante, este tipo de juicios está afectado de ciertas dosis de subjetividad a la hora de su realización.

Referencias Bibliográficas

[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 páginas.

[3] Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. Computer Netsorks and ISDN Systems, 30, 1998. p. 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] Lancaster, F. W. & Warner, A.J. Information retrieval today. Arlington, VA: Information Resources. 1973.

[9] La Serna P. N. y grupo, Estudio y evaluación de los Sistemas de Recuperación de Información. Vol 1 No 1 2004.

[10] La Serna P. N. y grupo, Diseño del Sistema de Recuperación de Información para la biblioteca FISI. Vol 2 No 2 2005.

[11] Martinez M.F. y Rodriguez 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

[12] Martínez Méndez Francisco Javier. Sistemas de Almacenamiento y Recuperación de Información, http://www.um.es/gtiweb/fjmm/sari2000.htm 2000.

[13] Martínez Méndez Francisco Javier. Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en Internet. Tesis Doctoral realizado en el año 2002.

[14] Notess, G.R. Search engine statistics. Bozeman, MT: Notess.com. http://www.searchengineshowdown.com/stats/2002

[15] Prieto-Diaz, R. and ARANGO, G. Domain Analysys: Acquisition of Reusable Information for Software Construction. New York: IEEE Press, 1991.

[16] Salton G. The SMART system. Encyclopedia of Library and Information Science 1980.

[17] Salton G. Y McGill M. Introduction to Modern Information Retrieval. Mc. Graw-Hill. 1983.

[18] Van Rijsbergen, C.J. Information Retrieval. London: Butterworths, 1979.

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

[20] SearchEngineWatch.com The major search engines Jupitermedia Corporation. http://www.searchenginewatch.com/links/major.html. 2002