RISI **2**(2), 26-31 (2005)
Rev. investig. sist. inform.
Facultad de Ingeniería de Sistemas e Informática
Universidad Nacional Mayor de San Marcos
ISSN: 1815-0268 (versión impresa)/ ISSN: 1816-3823 (versión electrónica)

DISEÑO DEL SISTEMA DE RECUPERACIÓN DE INFORMACIÓN PARA LA BIBLIOTECA FISI[1]

DESIGN OF THE INFORMATION RECOVERY SYSTEM FOR THE FISI LIBRARY

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

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) El diseño de la Arquitectura del Sistema de Recuperación de la Información para la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática; 2) La descripción de cada uno de los módulos del sistema; 3) El planteamiento de las actividades y tareas para la implementación del Sistema. El trabajo se desarrolla en el marco del proyecto de investigación «Construcción de un motor de recuperación de información para un Sistema de Bibliotecas», 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 del Espacio Vectorial, XML - Extensible Markup Languaje, Sistemas de Bibliotecas.

ABSTRACT The work that appears in this articulate is developed in the area of the recovery systems of information (RSI) Fundamentally the following activities have been made: 1) The design of the architecture system of recovery of the information for the library of the faculty of ingineering of systems and informatic. 2) The description of each one of the modulos system. 3)The position of the activities and tasks for the implementation system.

The work is within the framework developed of the investigation project «construction of a motor of information retrieval for a system of libraries» whose objective is to design a (RSI) recovery system of information for the library the ability engineering of system and computer later on for the central library of Greater San Marcos National University.

Key words: Information Retrieval Systems, Vector Model, XML - Extensible Markup Languaje, Library Systems.

1. INTRODUCCIÓN

El trabajo de investigación que se desarrolla tiene como objetivo principal diseñar e implementar un sistema de almacenamiento y recuperación de información para un sistema de bibliotecas. Para ello se utilizarán las técnicas definidas en el modelo de espacio vectorial [14] el que es uno de los modelos más utilizados en estos sistemas. Al mismo tiempo, con la finalidad de obtener eficiencia en el prototipo que se construye, se incorporarán metodologías y estándares actuales como XML (Extensible Markup Languaje), tecnologías web y el protocolo OAI (Open Archives Initiative).

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

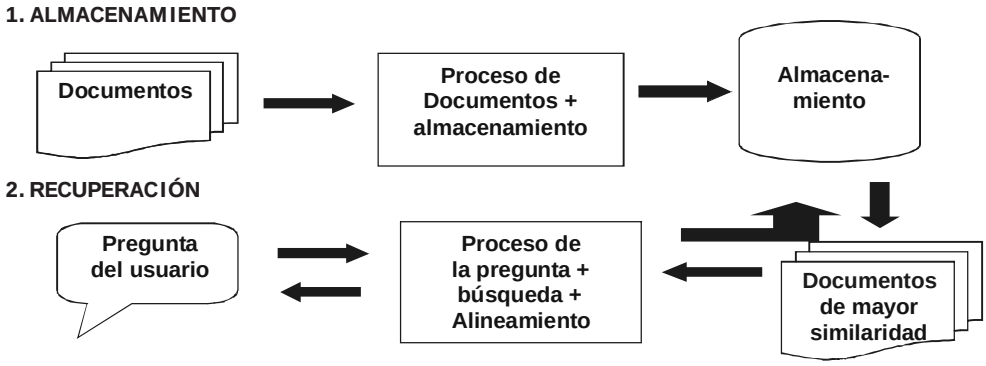

Figura 1. Modelo de Espacio Vectorial en los SRI.

El motor de búsqueda desarrollado será implementado en la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática, para realizar búsquedas de información como libros, tesis de pregrado, revistas, etc., y posteriormente incorporarlo a la Biblioteca Central de la Universidad Nacional Mayor de San Marcos. En un sistema de bibliotecas tradicional, tanto la pregunta del usuario como la búsqueda de información del sistema es sobre ciertos campos de la información, mientras que en los SRI la pregunta del usuario se realiza en lenguaje natural, y para la búsqueda de información se tiene en cuenta toda la información que contienen los documentos; esto permite que los resultados sean mejores en los SRI que utilizando los métodos tradicionales.

Según el modelo de espacio vectorial cada documento se registra en un vector de términos, y una colección de documentos forma una matriz 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; finalmente, se ordenan los documentos que tienen mayor valor de similaridad, y se muestran los resultados al usuario.

En la Figura 1 se muestra una vista funcional del modelo, en donde se observan las tareas que se realizan.

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.

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 computadores servidores millones de documentos. Por lo tanto, cómo recuperar, en forma eficiente, 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, gobierno y público en general que necesitan buscar información. Se están dando múltiples aplicaciones prácticas, algunas de las más conocidas son los buscadores web y las bibliotecas digitales. Dos de los autores más citados por los especialistas en la materia son Gerard Salton [14], y Ricardo Baeza-Yates [1].

La estructura del presente artículo es la siguiente: En la sección 2 se presenta la arquitectura del SRI para la biblioteca de la FISI, en las secciones 3 y 4 se describen cada uno de los módulos del sistema siguiendo el modelo del Espacio Vectorial. La sección 5 corresponde a la evaluación del trabajo que se desarrolla, y, finalmente, en la sección 6 se bosquejan las conclusiones y trabajos futuros en el sistema.

2. ARQUITECTURA DEL SRI PARA LA BIBLIOTECA DE LA FISI

Para el desarrollo del sistema de recuperación de información para la biblioteca de la FISI, que se plantea, seguimos el modelo de espacio vectorial [14]. 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 que cada término puede ser más o menos significativo en un documento o en toda la colección de

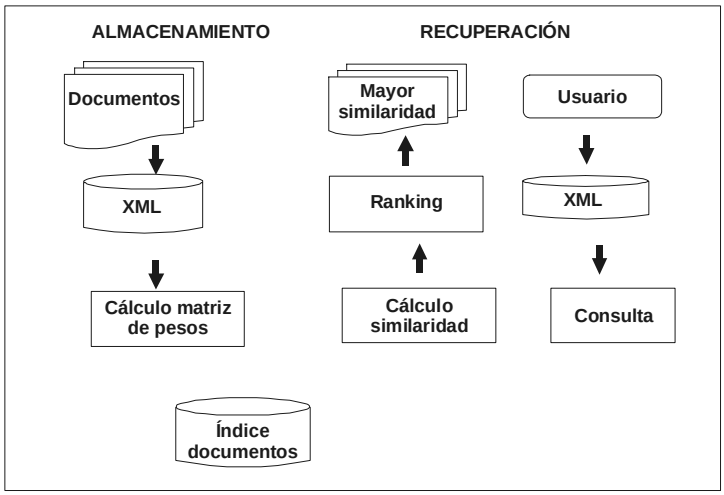

Figura 2. Arquitectura del sistema propuesto.

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 el 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 sección 4 se describe el cálculo de la similaridad entre la consulta del usuario y cada uno de los documentos en el sistema propuesto.

En la figura 2, 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 ROM, son las tesis de pregrado. Hay alrededor de cien tesis en formato digital, las que serán utilizadas como los documentos iniciales para el sistema. Estos documentos serán convertidos al formato XML (Extensible Markup Languaje).

XML permite 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 de que facilitará el intercambio de información y la cooperación con otros sistemas de la Facultad. Adicionalmente, XML es una versión abreviada de SGML (Standard Generalized Markup Language, ISO 8879), este último es el estándar internacional para la definición de la estructura y el contenido de diferentes tipos de documentos electrónicos. El XML es más simple y optimizado que el SGML para su utilización en ambientes de internet.

3. MÓDULO DE ALMACENAMIENTO DE LA COLECCIÓN DE DOCUMENTOS

3.1 Selección de términos

La selección de términos que mejor representen a los documentos, y que mejor discriminen unos documentos respecto de otros, conlleva al preprocesado de texto de la colección de documentos, cuyos objetivos son: 1) aumentar las tasas de exhaustividad del SRI, 2) reducir el espacio de almacenamiento, y 3) aumentar la velocidad de proceso. El preprocesado del texto normalmente sigue al menos los siguientes pasos [1]:

1. Análisis léxico del texto, es decir, el análisis de cada elemento del lenguaje, con el objetivo de determinar el tratamiento que se realizará sobre números, signos de puntuación, tratamiento de mayúsculas, minúsculas, nombres propios, etc.
2. Eliminación de palabras vacías (stop words), con el objetivo de reducir aquellos términos que tienen poca capacidad semántica o por su alta frecuencia, son poco significativos en el proceso de recuperación de la información. Es una forma de delimitar el número de términos que servirán como

términos índice. Generalmente se almacena en un archivo de palabras vacías.

3. Aplicar extractores de raíces *(stemmers),* programas que reducen cada palabra a su raíz eliminando prefijos, sufijos, terminaciones verbales, y obtener términos lematizados. El proceso también es denominado *lematización*. El objetivo es reducir el número de términos índice semánticamente muy parecidos.
4. Selección de los términos, que serán los términos índice, normalmente se realiza sobre la naturaleza sintáctica del término, resultarán como consecuencia de aplicar los pasos anteriores.
5. Utilizar *tesauros* que agrupan los términos en un solo concepto por término. Por ejemplo un tesauro geográfico, diccionario de sinónimos y antónimos. Los tesauros permiten ampliar las consultas, además de buscar en los archivos de índices un término, se busca dicho término en los tesauros relacionados.

3.2 Cálculo de la Matriz de pesos de la colección

Una vez seleccionado el conjunto de términos de la colección de documentos, el siguiente paso es realizar el cálculo de la matriz de pesos de la colección. Dos cálculos son importantes en la determinación de la matriz de pesos:

1. Hallar la frecuencia de un término (*tf*, frecuencia del término), es decir, el número de veces que aparece el término en un documento. Así, si un término aparece muchas veces en un documento, se supone que es importante en ese documento.
2. Determinar la frecuencia inversa de un término en la colección de documentos (idf), este cálculo se debe a que si un término aparece en muchos documentos, entonces ese término no es útil para distinguir ningún documento de los otros de la colección. Es decir, que la capacidad de recuperación de un término es inversamente proporcional a su frecuencia en la colección de documentos.

Hay varias técnicas para asignar pesos a los términos; dos de ellas, las que presentamos a continuación son las más representativas; las expresamos mediante las ecuaciones 1 y 2. En la ecuación 1, una de las más simples, para calcular el peso de cada elemento del vector, se tiene en cuenta la frecuencia del término dentro de cada documento *tf*, combinándola con la frecuencia inversa del término en la colección *idf* [1].

$$W_{ij} = tf_i * idf_j \qquad \text{(ecuación 1)}$$

Donde, W_{ij} es el peso del término j en el documento i.

En la ecuación 2, por cierto una de las técnicas más utilizadas, el peso del término se calcula como el producto de la frecuencia del término j en el documento i, multiplicado por el logaritmo de N / df_j.

$$W_{ij} = tf_{ij} * \log N / df_j \qquad \text{(ecuación 2)}$$

Donde, N es el número de documentos de la colección, y df_j es el número de documentos en que aparece el término j.

Para hallar la matriz de pesos de los términos, básicamente se realizan los siguientes pasos:

1. Calcular las frecuencias de cada término en la colección.
2. Cálculo del IDF (Inverse Document Frecuency).
3. Cálculo de la frecuencia en cada documento.
4. Hallar el peso del término en cada documento.

A continuación se ilustra con un ejemplo una matriz de términos para una colección de tres documentos, y luego se presenta la matriz de pesos correspondiente.

Seleccionamos como términos cada una de las palabras en los siguientes documentos:

doc1 = «Sistemas de recuperación de la información».

doc2 = «Clasificación de los Sistemas de Recuperación de la Información».

doc3 = «Motores de búsqueda».

Los términos que aparecen en los documentos son:

(sistemas, de, recuperación, la, información, clasificación, los, 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 se presenta en la siguiente tabla 1:

Tabla 1

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

Tabla 2. La matriz de pesos, sin las palabras vacías, queda de la siguiente manera:

sistemas		recuperación	información	clasificación	motores	búsqueda
doc1	0.11	0.11	0.11	0	0	0
doc2	0.11	0.11	0.11	0.11	0	0
doc3	0	0	0	0	0.11	0.11

4. MÓDULO DE RECUPERACIÓN

La consulta o solicitud del usuario hecha en lenguaje natural, también se expresa mediante un vector de términos; los términos deben ser los mismos que el de la colección de documentos. El mecanismo de obtención de pesos también se aplica a la consulta del usuario. De esta manera, la consulta y cada uno de los documentos están expresados en vector y matriz de pesos de términos respectivamente, los cuales después, mediante un cálculo de similaridad permiten hallar aquellos documentos que se aproximan más a la pregunta del usuario.

La resolución de la consulta Baeza-Yates y Ribeiro-Neto, 1999 [1] consiste en un proceso de establecer el grado de semejanza entre el vector consulta y el vector de cada uno de los documentos; aquellos cuyo grado de similitud sea más elevado se ajustarán mejor a las necesidades expresadas en la consulta. Sin embargo, es el usuario el que debe decidir la relevancia de los documentos recuperados, siendo ésta una característica totalmente subjetiva del mismo.

Hay varios métodos para calcular la similitud que existe entre un documento y una consulta; una de las más utilizadas es aquella que calcula la distancia que existe entre los vectores que los representan, y realiza el producto escalar de esos vectores; dicho producto, a su vez, corresponde al coseno del ángulo entre los dos vectores (ecuación 3).

$$\text{Sim (vector di, vector dj)} = \text{vector di} * \text{vector dj} \quad \text{(ecuación 3)}$$

$$= \cos(\text{vector di, vector dj}) = \frac{\text{vector di} * \text{vector dj}}{|\text{vector di}| * |\text{vector dj}|}$$

La similaridad es un valor entre cero y uno. Dos vectores iguales tienen similaridad uno; dos vectores que no comparten ningún término tienen similaridad cero. Se seleccionan los documentos que tienen mayor similaridad con la consulta del usuario.

Para hallar los documentos que un usuario solicita, básicamente se realizan los siguientes pasos:

1. Selección de términos de la consulta del usuario.
2. Cálculo de la matriz de pesos de la consulta del usuario.
3. Cálculo de similaridad entre la consulta del usuario y cada documento de la colección.
4. Se ordenan los documentos que tienen mayor valor de similaridad, y se muestran los resultados al usuario.

En el ejemplo presentado en la sección 3.2, para la colección de documentos doc1, doc2, y doc3, si la consulta es «recuperación de información», la representación de la consulta quedaría expresada con el siguiente vector (Tabla 3).

A continuación se realiza el producto escalar de la matriz de pesos con el vector de la consulta, es decir, se calculan las distancias del vector de la consulta con el vector de cada documento, y se devuelven los documentos ordenados de mayor a menor similitud. Para el ejemplo que seguimos, los documentos con mayor similaridad se presentan en la siguiente tabla:

	Similaridad
doc1	0.55
doc2	0.55
doc3	0

5. EVALUACIÓN DEL DISEÑO

En el presente trabajo describimos el avance del diseño e implementación de un sistema de almacenamiento y recuperación de información para un sistema de bibliotecas. Inicialmente, será implementado en la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática, para realizar búsquedas de información como libros, tesis de pregrado, revistas, etc., con base en otros trabajos desarrollados por especialistas en el

Tabla 3

	sistemas	recuperación	información	clasificación	motores	búsqueda
consulta	0	0.5	0.5	0	0	0

tema, y sobre la base de nuestros requerimientos, hemos propuesto la arquitectura del SRI para la biblioteca de la FISI. La implementación del diseño se realizará en dos etapas: 1) El almacenamiento de los documentos, 2) El proceso de recuperación de la información.

En la primera etapa, el primer paso consiste en seleccionar los tipos de documentos que vamos a considerar; como mencionamos en la sección 2, los documentos que se van a procesar serán las tesis de pregrado, las que se encuentran en formato digital; son alrededor de cien. El siguiente paso, en esta primera etapa, consiste en pasar los documentos al formato XML, para que a partir de ellos se realice el proceso de almacenamiento, es decir, seleccionar los términos de los documentos, que serán considerados como índices, y los cálculos para determinar los pesos de los términos.

En la segunda etapa, la pregunta del usuario también será expresada en formato XML, para después hacer el proceso de consulta, cálculo de similaridad y, finalmente, presentar los documentos con mayor similaridad a la pregunta del usuario.

6. CONCLUSIONES

La investigación ha dado lugar a una propuesta de diseño e implementación de un Sistema de almacenamiento y recuperación de información, que inicialmente será implementado en la Biblioteca de la Facultad de Ingeniería de Sistemas e Informática, para realizar búsquedas de información como libros, tesis de pregrado, revistas, etc. Para ello se utilizarán las técnicas definidas en el modelo de espacio vectorial (Salton, 1983), incorporando otras metodologías y estándares actuales como XML (Extensible Markup Languaje), tecnologías web y el protocolo OAI (Open Archives Initiative). En la sección 2, la arquitectura propuesta contiene todos los módulos que se van a desarrollar; y en las secciones 3 y 4 se describen ampliamente cada uno de los pasos que se van a considerar en la implementación de estos módulos. Como se puede observar, se presentan todas las tareas por desarrollar en adelante.

REFERENCIAS BIBLIOGRÁFICAS

[1] Baeza-Yates, R. y 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. y 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». http://www.asis.org/annual-96/ElectronicProceedings/chu.html

[5] Delgado Domínguez. «Mecanismos de recuperación de Información en la www», Universidad de Islas Baleares, 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 Núm. 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] 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

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

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

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

[13] Salton G. «The SMART system». Encyclopedia of Library and Information Science, 1980.

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

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

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

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