Metadatos en los documentos HTML: una ayuda para la recuperación de información

Angós Ullate, José María
Centro de Documentación Científica
Universidad de Zaragoza
E-mail: angos@posta.unizar.es
Salvador Oliván, José Antonio
Departamento de Ciencias de la Documentación e Historia de la Ciencia
Universidad de Zaragoza
E-mail: jaso@posta.unizar.es
Fernández Ruíz, María Jesús
Centro de Documentación
Ayuntamiento de Zaragoza
E-mail: mjferuiz@posta.unizar.es

0.1. Resumen

Se describe un sistema de indización de los recursos de la WWW basado en la información contenida en la cabecera de un documento escrito en HTML (*HyperText Markup Language*). La información de la cabecera contiene los datos primarios de cada recurso accesible en Internet. También se describe el sistema para hacer el registro y la base de datos distribuida que representa el catálogo común de los recursos de Internet. Esta base de datos se utilizará en un sistema de búsqueda que facilite la recuperación de información.

Palabras Clave: Indización. Búsqueda de información. WWW. Internet. Cabecera. Documento HTML. Metadatos.

0.2. Abstract

An indexing system for WWW resources based on the information contained in the head of a document written in HTML (HyperText Markup Language) is described. The document head contains information on the primary data of each accessible resource in Internet. The registration system and the distributed database that contains a union catalogue of the resources of Internet is analyzed. This database will be used in the search system that facilitates the retrieval of information.

Scire. 5 : 2 (jul.-dic. 1999) 47-61.

Keywords: Indexing. Information research. Web. WWW. Internet. Head. HTML document. Metatags

1. Introducción

El número de redes interconectadas en Internet continúa creciendo, y con el aumento de la potencia de servidores del tipo *workstation* conectados a estas redes, es posible soportar tanto la búsqueda de información local como la búsqueda remota. Estas redes permiten el intercambio de información, ya que utilizan un protocolo de intercambio común, el TCP/IP (*Transmission Control Protocol / Internet Protocol*). Muchas universidades, instituciones y empresas interconectan sus ordenadores utilizando redes, y de esta forma comparten los recursos existentes en cada una de ellas y, precisamente, una de las redes que se está utilizando de manera casi exclusiva es Internet.

En Internet se pueden utilizar fuentes de información públicas (generalmente gratuitas) o privadas (pagando por su utilización). Estas fuentes incluyen revistas electrónicas, periódicos, directorios, libros, archivos de sonido, imágenes, bases de datos científicas, de productos y servicios, textos, juegos, etc. Para acceder a todas esas fuentes, se necesita utilizar un sistema que permita un fácil acceso y búsqueda de los recursos disponibles en Internet, pero para ello hay que desarrollar y mejorar los existentes.

Los sistemas de información distribuidos, incluso aquellos cuyo control se realiza por un ordenador central (*mainframe*), tienen muchos problemas causados generalmente por la deficiente catalogación y por las diferencias existentes en la semántica y representación utilizadas. Estos problemas aumentan en los sistemas de información distribuidos que intentan integrar los recursos ofrecidos por distintos sistemas de información de Internet. Estos problemas podrían disminuir, o incluso evitarse, normalizando la estructura del índice y construyendo un sistema bibliográfico que use definiciones de control normalizadas.

En un sistema distribuido, como lo es Internet, la entrada de documentos debe ser distribuida y accesible a todos los proveedores de la información y a todos los usuarios de la red a través de una interfaz, un navegador como Netscape o Explorer. Los proveedores de información prepararán e introducirán la información bibliográfica en base a un sistema de indización normalizada, registrado en una base de datos distribuida en distintos nodos regionales o nacionales para aumentar la disponibilidad y respuesta. El sistema de búsqueda que los usuarios utilizarán tiene que ser sencillo y que ayude a localizar la información apropiada. Estos sistemas deben de incorporar los conocimientos de catalogadores, bibliotecarios y documentalistas expertos, para facilitar la recuperación y guiar al usuario en todos sus pasos.

La respuesta del sistema de búsqueda proporcionará el número de registros obtenidos en cada consulta y los datos (URL, *Uniform Resource Locator*) que permitan acceder a los documentos. La "navegación" por la base de datos, los recursos, protocolos y filtros usados serán seleccionados por el sistema de forma transparente para el usuario, facilitándole la tarea y proporcionándole acceso a todos los recursos, como si fuera hecho por un sistema de información centralizado.

2. Fuentes de Información

Las fuentes de información suelen ser el material original publicado: artículos, monografías, informes, tesis, programas, imágenes, películas, etc.; son las denominadas fuentes primarias. Las fuentes secundarias, llamadas a veces meta-información, se utilizan como índices para estas fuentes primarias de información y son creadas posteriormente, a los pocos meses o años. Una fuente terciaria de información es una combinación de información seleccionada y extraída de las fuentes primarias y secundarias.

El propósito de los índices y bibliografías (información secundaria) es la de inventariar la información primaria y permitir un acceso fácil a ella. La preparación de una bibliografía requiere encontrar la fuente primaria, identificar la materia, etc., describiéndola para que, posteriormente, los futuros usuarios la encuentren y la clasifiquen de acuerdo a normas aceptadas.

Ya que un índice va a ser usado por muchos usuarios, tiene que ser preciso, fácil de usar, debidamente clasificado (por autor, título, materia, etc.), actualizado y completo en su área de cobertura. Para que bibliográficamente sea útil, debe satisfacer una necesidad real. El éxito de Archie como sistema bibliográfico (para localizar ficheros disponibles en Internet via FTP) radica en que tiene una interfaz de usuario simple y sólo se necesita saber, para utilizarlo, el nombre de un programa, fichero o el tipo del fichero (la extensión), que probablemente está distribuido en uno o más lugares FTP anónimos. En el caso de la bibliografía en línea, los recursos de Internet como el Web, necesitan que el sistema esté actualizado en un breve período (minutos o como mucho horas), de manera que se conozca la existencia de un nuevo recurso rápidamente. Esta velocidad en la actualización es la característica que lo diferencia, fundamentalmente, del sistema bibliográfico de publicación impresa, que requiere semanas o meses en el caso de bases de datos en línea y más todavía para versiones en CD-ROM e incluso años para la versión impresa.

La manera en que se prepara y colecciona la bibliografía sobre una materia puede ser tradicional o semiautomática. El método tradicional consiste en que un grupo de especialistas evalúa las fuentes y realiza las entradas descriptivas para

cada registro. La exactitud de esta bibliografía es alta pero la cobertura suele ser limitada. Por otro lado, el método semiautomático consiste en que escanea los trabajos publicados por fuentes delimitadas (por dominio, idioma o región geográfica) y asigna a cada trabajo distintas submaterias. Generalmente se pueden proporcionar diferentes accesos al mismo documento con múltiples encabezamientos, siendo esto lo deseable ya que un registro puede tratar de más de un tema.

En general, en una búsqueda de información bibliográfica, un documento es más pertinente cuanto más significativo sea el título. La dependencia de los títulos como criterio de búsqueda, exige que deban ser indicativos del contenido del documento. Esto no siempre ocurre de esta manera, de ahí que alguien (el autor o el catalogador) tenga que añadir anotaciones, palabras o frases clave para indicar el contenido real. La exactitud o calidad de un documento puede indicarse incluyendo las opiniones de los evaluadores. Sin embargo, estas opiniones raramente son accesibles al catalogador. Otra característica importante para el usuario, es la existencia de un resumen preciso. Un resumen proporciona un sumario del material y por eso es más indicativo del contenido que el título o las palabras clave. Otras características, como la división de bibliografía por materias y submaterias, de interés en los sistemas manuales, no serán tan útiles en búsquedas en bases de datos; sin embargo, el acceso por estos criterios debe implementarse.

3. Sistema de Búsqueda y Catalogación

Los catálogos de las Bibliotecas están preparados por especialistas y, para cada entrada, se registra el autor, título, editor, lugar de publicación y otros detalles. El catálogo colectivo informa dónde está localizado cada registro.

Actualmente existe un gran número de documentos en Internet, además de los ficheros cuyos nombres pueden buscarse utilizando sistemas como Archie o Xarchie. La popularidad de la World Wide Web y de navegadores como Netscape o Explorer ha hecho que muchos investigadores publiquen sus trabajos en línea. Se han llevado a cabo diferentes intentos para proporcionar una búsqueda fácil de documentos relevantes, por ejemplo utilizando el WAIS, los Robots de Búsqueda —Yahoo (http://www.yahoo.com), Lycos (http://www.lycos.com/), Altavista (http://www.altavista.com), WebCrawler (http://webcrawler. com/), etc— o los Metabuscadores —Metacrawler (http://www.metacrawler.com/ index.html), Search.com (http://search.cnet.com/), Find-it (http://www.itools.com/find-it/find-it.html), etc.

Sin embargo, el problema que tienen muchos de estos índices es que la selección de los documentos que realizan es a menudo inexacta y demasiado abundante. Es muy difícil obtener los documentos correctos y muy fácil perder infor-

mación pertinente, ya que la indización de los documentos es deficiente. Además, se pide al usuario que acceda al recurso encontrado, basándose sólamente en información del título de la página encontrada y de la dirección URL (en algunos casos con unas breves líneas contenidas en la página localizada), y que, con tan escasa información, decida si el recurso satisface sus necesidades.

Estos problemas están tratados en el sistema propuesto usando una entrada de índice apropiada a la que llamaremos Información de Cabecera y proporcionando un mecanismo para registrar, controlar y buscar la bibliografía. El sistema es activo y necesita que el proveedor de la información registre el recurso produciendo una entrada en el índice para este recurso. Ya que el proveedor es responsable de la preparación de la entrada del índice, probablemente la información registrada sea buena.

El sistema total utiliza un tesauro para ayudar al usuario en el registro y en el proceso de búsqueda. La utilización del sistema evita el caos introducido por las diferentes percepciones de diferentes indizadores. Por lo tanto, tiene que estar presente (e imponerse) alguna forma de normalización de los términos usados, y nada mejor que utilizar o desarrollar un tesauro.

Para la generación del índice y el mantenimiento del tesauro se deben utilizar los conocimientos y experiencia de un catalogador o documentalista experto que ayude al proveedor del recurso a seleccionar términos correctos para los registros, como por ejemplo la materia, submateria y palabras clave. Igualmente, tiene que usarse un sistema experto en la búsqueda para ayudar al usuario en la localización de recursos de información apropiados. El último componente del sistema es la base de datos distribuida y duplicada de los recursos bibliográficos válidos en línea. La base de datos está en un segundo plano y los usuarios no se preocupan de su presencia y mucho menos de su naturaleza distribuida y duplicada. Estos componentes se describen posteriormente.

4. Información de la Cabecera del documento HTML

La parte más importante de cualquier sistema de indización es el registro que se hace para cada referencia que se está indizando. El sistema MARC, en el ámbito de las Bibliotecas, ha abordado el problema de la catalogación y en particular el de la catalogación en un formato electrónico o multimedia. Sin embargo, este sistema está alejado de la comprensión de muchos proveedores de información.

En este trabajo, proponemos una estructura de índice más simple, llamada Información de la Cabecera de los documentos HTML, para los recursos accesibles directamente en Internet. La estructura incluye información que se considera útil para sistemas en línea. La sintaxis es el lenguaje HTML (*HyperText Markup Language*) que está basado en el lenguaje SGML (*Standard Generalizad*

Markup Language). Sin embargo, el usuario, trabajando con el sistema de entrada al índice es guiado en el proceso por un sistema experto. Este sistema guía al usuario en la elección de téminos normalizados de un tesauro, por medio de una interfaz gráfica sencilla.

Esta información complementaria de la cabecera se recogerá en el elemento META de los documentos HTML. Como bien dice Ian Graham en *Introduction to HTML*, META es un elemento general para recoger la meta-información del documento, es decir, que sirve para recoger la información sobre el documento que no puede expresarse con otros elementos como el LINK, BASE o HEAD.

Este elemento del documento HTML, puede contener información META equivalente HTTP (*HTTP-Equivalent META Information*), o bien especificada arbitrariamente por el usuario (*Arbitrary User-Specified META Information*). El primer tipo de información, devuelta generalmente por el servidor como un campo de las cabeceras HTTP, no es utilizable para nuestro propósito. Sin embargo, el segundo tipo, la especificada por el usuario, se acopla perfectamente a nuestras necesidades de descripción bibliográfica. A continuación mostramos un ejemplo de algunos elementos META relacionados con esta publicación:

- <META NAME="author" CONTENT="Jose Maria Angós Ullate">
- <META NAME="keywords" CONTENT="html documentation web url">
- <META NAME="editor" CONTENT="Publicaciones Universidad Zaragoza">

El atributo NAME es utilizado para referirse a los nombres que hemos seleccionado para la Información de la Cabecera del documento. Desde luego, esta información será realmente útil en la recuperación de la información, si estos elementos son tenidos en cuenta por los robots que indizan los documentos HTML y analizan el contenido del elemento META. (Los robots de indización del Web son programas automáticos que encuentran páginas Web y crean los índices de búsqueda).

Con la Información de la Cabecera se pretende incluir aquellos campos que se utilizan más frecuentemente en la búsqueda de un recurso de información. Ya que la mayoría de las búsquedas comienzan con un título, nombre de un autor (70%), materia y submateria (50%), hemos hecho que la entrada de estos campos sean obligatorios en la información de la cabecera. También se incluye el resumen para decidir si el recurso es útil.

Los elementos que forman parte de la Información de la Cabecera, recogidos en el campo META, se escriben en inglés por coherencia con la especificación de

los documentos HTML, y los que nos parecen más adecuados (sin ser una lista cerrada) para la descripción y posterior recuperación son los siguientes:

1. *Título:* Obligatorio, es el título del recurso. Está marcado con *title*:

   <META NAME="title" CONTENT="Indización y Búsqueda en Web basadas en la información de la cabecera de los documentos HTML ">

2. *Temática:* Lista que incluye campos de materia y hasta dos niveles de sub-materia: es obligatoria una entrada por lo menos. Es un grupo repetitivo (un campo de múltiples partes con una o más ocurrencias de items en el grupo). Todos los recursos tienen, por lo menos, una ocurrencia para este campo. Podría estar marcado con *Subject* :

   <META NAME="Subject" CONTENT=" Almacenamiento de Información y Búsqueda ">

3. *Idioma:* Lenguaje en que está la información del recurso. Es opcional y su código puede ser *Language*:

   <META NAME="Language" CONTENT="Español Spanish">

4. *Juego de Caracteres:* Se recoge el conjunto (set) de caracteres utilizado. Es opcional y su marca puede ser *Character-Set* :

   <META NAME="Character-Set" CONTENT="ISO-8879">

5. *Autor:* Es el autor del recurso, es un grupo repetitivo. Los subcampos que contiene son: organización, dirección, números de teléfono y fax, y la dirección e-mail. Todos los subcampos, excepto el nombre que es obligatorio, son opcionales salvo cuando el autor es una organización que en ese caso se debe incluir la organización. El término autor se usa para dar la afiliación del programador, creador, artista, etc. Sus marcas serían, para cada autor, las siguientes: Author, Organization, Address, Phone, Fax, Email; por ejemplo:

   <META NAME="Author" CONTENT=" ANGÓS ULLATE, José María ">

   <META NAME="Organization" CONTENT=" Depto. Ciencias de la Documentación e Historia de la Ciencia, Universidad de Zaragoza">

   <META NAME="Address" CONTENT=" Pedro Cerbuna 12, Facultad de Filosofia y Letras, 50009 Zaragoza (España) ">

   <META NAME="Phone" CONTENT=" 976 761332 ">

   <META NAME="Fax" CONTENT=" 976 761088 ">

   <META NAME="Email" CONTENT=" angos@posta.unizar.es ">

6. *Descriptores:* Obligatorio, es la lista de palabras clave. Cada recurso debe tener al menos una palabra clave. Se puede usar el término *Keyword*:

   <META NAME="Keyword" CONTENT=" Indización Búsqueda de información Web Internet cabecera documento HTML ">

7. *Editor:* Opcional en el caso de una versión publicada. Se puede usar el término *Publisher*:

   <META NAME="Publisher" CONTENT=" Servicio de Publicaciones, Universidad de Zaragoza">

8. *Fecha de Creación:* Obligatorio, es la fecha en que se ha creado el documento. Podria usarse *Created-Date*:

   <META NAME="Created-Date" CONTENT=" 15-03-1999">

9. *Fecha de Caducidad:* Opcional, es la fecha en que el documento deja de ser válido. Podría usarse *Expiry-Date*:

   <META NAME="Expiry-Date" CONTENT=" ">

10. *Fecha de Actualización:* Generado por el sistema, es la fecha en que se actualiza el documento. Podría usarse *Updated-Date*:

    <META NAME="Updated-Date" CONTENT=" 18-03-1999">

11. *Versión:* Opcional, versión del recurso. Podría usarse *Version*:

    <META NAME="Version" CONTENT=" 1.00">

12. *Clasificación:* Opcional, nivel de seguridad del recurso. Podría usarse *Classification*:

    <META NAME=" Classification " CONTENT=" Público ">

13. *URL:* Lista de localizaciones, *Universal Resource Locator*. Podría incluir una lista de una o mas localizaciones donde el ítem es válido. Es obligatoria una dirección. Puede usarse *URL*:

    <META NAME=" URL " CONTENT=" http://www.unizar.es/JM/indexa.htm">

14. *URN:* El campo URN (*Unique Resource Name*) da el nombre único del recurso, si se tiene. Este nombre puede usarse en lugar de un URL si el documento es probable que se mueva o puede ser accesible desde múltiples localizaciones. Podría usarse *URN*:

    <META NAME=" URN " CONTENT=" No existe actualmente">

15. *UAS:* El campo UAS (*Universal Archive Site*) es utilizado para indicar la localización universal del archivo donde está el recurso. Se supone que

el recurso existirá más tiempo que la fecha de expiración del recurso. Podría usarse *UAS*:

<META NAME=" UAS " CONTENT=" No existe actualmente">

16. *Resumen:* Opcional pero recomendable. Podría usarse *Abstracts*:

    <META NAME=" Abstracts " CONTENT=" Este artículo describe un sistema de indización de los recursos de Internet, basado en la información contenida en la cabecera de un documento escrito en HTML (HyperText Markup Language). La información de la cabecera contiene los datos.....">

17. *Hardware:* Lista de Hardware requerido. Podría usarse *Hardware*:

    <META NAME=" Hardware " CONTENT=" Cualquier ordenador con acceso a Internet ">

18. *Software:* Lista de Software requerido. Podría usarse *Software*:

    <META NAME=" Software " CONTENT=" Navegador: Netscape, Explorer, Mosaic... ">

19. *Tamaño:* Tamaño del recurso en bytes. Podría usarse *Size*:

    <META NAME=" Size " CONTENT=" 65000">

20. *Coste:* Opcional, coste por acceder al recurso. Podría usarse *Cost*:

    <META NAME=" Cost " CONTENT=" Gratis">

21. *Password:* Obligatorio, password codificado o firma digital del proveedor del recurso introducido inicialmente y actualizable posteriormente. Cualquier cambio en la actualización de parte de la cabecera, necesita password o firma digital. Podría usarse *Password*:

    <META NAME=" Password " CONTENT=" deshollinador25">

22. *Firma:* Esta firma digital puede usarse para autentificar el recurso. Podría usarse *Signature*:

    <META NAME=" Signature " CONTENT=" h01110101 01101100101100111010001">

6. Registro del Índice

La entrada en el índice y el registro del documento se realizará por medio de una interfaz gráfica (Figura 1) que facilita al proveedor (autor/creador) del recurso registrar la información bibliográfica sobre el recurso. La interfaz permite al proveedor introducir la información y obtener ayuda por medio de una ventana de selección tipo *pop-up* y un mecanismo experto que le sugiere términos controlados extraídos de un tesauro.

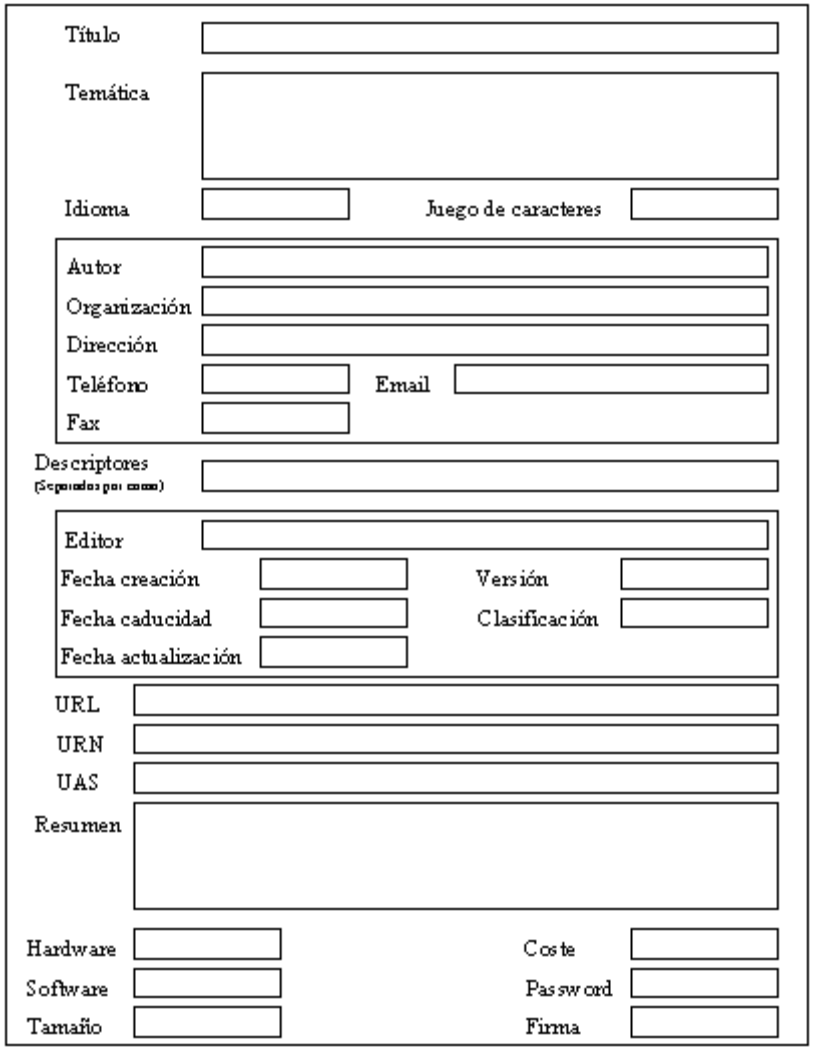

Fig. 1. Interfaz gráfica para la introducción de datos en la cabecera de documentos HTML

Una vez que la información se ha introducido correctamente, el autor puede

decidir registrar la entrada de la cabecera en la Base de datos que recoge dichas cabeceras. Cuando la información de la cabecera es aceptada por la base de datos, se notifica al autor/creador. Se proporciona una contraseña o firma digital la primera vez que se registra la cabecera y se utiliza para hacer cambios en ella. La entrada sólo puede actualizarse por personas autorizadas.

El sistema verifica que el recurso es accesible, comprobando que se haya añadido. La firma digital se añade también a la cabecera semántica. El propósito de esta información es establecer la veracidad del recurso cuando se recupera a través de la cabecera semántica. Si el recurso está corrupto, la validación de la veracidad fallará y se notificará al usuario

Cada recurso posee un nombre único (URN). La entrada al índice que es registrada se comunica a la base de datos descrita a continuación.

7. La cabecera del sistema de bases de datos distribuido

Las entradas del índice registradas por un proveedor de un recurso se almacenan en un sistema de base de datos distribuido (SBDD). Desde el punto de vista de los usuarios del sistema, la cabecera de la base de datos puede considerarse como un sistema monolítico. En realidad, estará distribuida y duplicada para permitir operaciones fiables y tolerantes a fallos. El interfaz oculta la naturaleza distribuida y duplicada de la base de datos.

Generalmente las bases de datos de diferentes materias serán mantenidas en diferentes nodos de Internet. Las localizaciones de estos nodos sólo necesitan ser conocidas por el correspondiente interfaz. Se deberá usar un catálogo para distribuir esta información; sin embargo, este catálogo podría distribuirse y duplicarse para sistemas de bases de datos distribuidas.

La información de la Cabecera introducida por el proveedor del recurso usando una interfaz gráfica es retransmitida de la estación de trabajo del usuario, por medio de un proceso cliente, al proceso servidor de la base de datos en uno de los nodos del SBDD. El nodo se elige en función de su proximidad a la estación de trabajo o de la materia del registro indizado. Una vez recibida la información, el servidor verifica la corrección y autenticidad de la información y si está todo en orden, envía la señal de correcto (ACK) al cliente.

El nodo servidor es el responsable de la localización de las particiones del SBDD donde deberá almacenarse la entrada y desde el que se enviará la información duplicada a los nodos apropiados. Por ejemplo, la entrada de la cabecera del ejemplo será parte del SBDD por materias de Almacenamiento de Información y Búsqueda.

Igualmente, el proceso servidor en la base de datos es responsable de proporcionar información al catálogo para el sistema de búsqueda. De esta manera,

los diferentes lugares de la base de datos trabajan de un modo cooperativo manteniendo la consistencia de la parte duplicada. La naturaleza duplicada de la base de datos también asegura la distribución de carga y el acceso continuo a la bibliografía cuando uno o más lugares de almacenamiento de la información no sean accesibles temporalmente.

8. El sistema de búsqueda

La principal guía para diseñar el sistema de búsqueda es la de utilizar la experiencia que tiene el documentalista. Se le pide que identifique las mejores fuentes de información de un determinado tema y que ayude en la selección de los materiales bibliográficos que se necesiten. El documentalista busca las repuestas a estas preguntas usando la información que obtiene de las búsquedas bibliográficas o de los documentos que le proporcionan bibliotecarios especializados y con conocimiento de las materias más relevantes.

En el componente de búsqueda del sistema propuesto, nosotros planteamos incorporar la experiencia del documentalista. De esta manera se guiará al usuario para que introduzca los distintos términos de búsqueda utilizando una interfaz gráfica similar a la usada por el sistema de entrada de índice. De manera análoga a la utilizada para crear el índice, se ayudará al usuario con un sistema experto que le permitirá elegir los términos más apropiados para la búsqueda, por medio de la consulta al tesauro o sugiriéndole términos del mismo. El sistema experto, basándose en estadísticas anteriores y en su "aprendizaje", podrá optimizar la búsqueda.

El sistema de la búsqueda también usa una interfaz gráfica y un proceso de tipo cliente. Una vez que el usuario ha pedido una búsqueda, el cliente comunica con el catálogo de SBDD más cercano para determinar el sitio apropiado de la base de datos SBDD. Seguidamente, el cliente comunica con esta base de datos y recupera los títulos más apropiados. Los resultados de la búsqueda se envían al ordenador del usuario y se muestran los contenidos de las cabeceras, y el usuario puede acceder a uno o a más de los documentos originales sin más que seleccionarlos en la pantalla. Si el documento seleccionado está almacenado en más de una dirección URL, el sistema decidirá de donde conseguirlo, basándose en la optimización de costes (en general, se elegirá el SBDD más cercano al cliente para proporcionarle la información pedida)

9. Conclusiones

Los sistemas actuales de indización de los documentos del Web se basan en el examen de las páginas de la red, por medio de robots de búsqueda, recogiendo los nuevos documentos encontrados e indizando sus contenidos. Los términos

extraídos de las páginas HTML, pasan a formar parte de sus índices. La desventaja más importante de este proceso es la poca fiabilidad de las entradas del índice producido y la falta de un resumen auténtico para el artículo. La poca fiabilidad es debida a la falta del control del vocabulario, puesto que se recoge cualquier término que aparece sin ningún tipo de comprobación.

Actualmente, estos planteamientos son relevantes para documentos texto del Web y no son aplicables a otros recursos. Cada vez más, los proveedores de información solicitarán el cobro de sus servicios: la creación de un índice se deberá pagar. Además, los usuarios, sin tener una idea adecuada de los contenidos, sólo pueden decidir a partir de sus títulos, no querrán recuperar (y por lo tanto pagar) aquellos que les parezcan poco relevantes.

En el sistema propuesto, el proveedor del recurso es el que prepara la información del índice. Por consiguiente, esta entrada del índice será más fiable que el realizado por otra institución o hecho por el simple escaneo del documento. La presencia de un resumen da al proveedor del recurso la posibilidad de proporcionar un buen resumen. Este sumario, que forma parte del índice, permite a los usuarios tomar decisiones mejor informadas sobre la pertinencia o no del documento localizado.

El sistema proporciona un sistema experto con una interfaz gráfica que ayuda al proveedor del recurso a producir la entrada del índice que forma parte de la base de datos. El sistema experto ayuda a seleccionar los términos apropiados del índice, como la materia, submateria, palabras clave, etc.

10. Bibliografía

Berners-Lee, T., Cailliau, R. WorldWideWeb : Proposal for a HyperText Project. URL= <http://info.cern.ch/hypertext/WWW/Proposal.html>.

Berners-Lee, T. URL and The Names and Addresses of WWW objects. URL= <http://info.cern.ch/hypertext/WWW/Addressing/Addressing.html>.

Berners-Lee, T. Hypertext Markup Language, Internet working draft. URL= <http://info.cern.ch/hypertext/WWW/MarkUp/HTML.html>.

Berners-Lee, T. Wide Web Initiative: The Project. URL= <http://info.cern.ch/hypertext/WWW/TheProject>.

Byrne, Deborah J. MARC manual: understanding and using MARC record. Englewood : Libraries Unlimited, 1991.

Crawford, Walt. MARC for Library Use: Understanding USMARC. Boston : G. K. Hall, 1989.

Cromwell, Willy (1994). The Core Record: A New Bibliographic Standard. // Library Resources and Technical Service. 38 : 4 (1994) 415-424.

Daviel, A. (1997). A Dictionary of HTML META Tags. URL= <http://vancouver-webpages.com/META/>.

De Bra, P., Houben, G-J. ; Kornatzky, Y. Search in the World-Wide Web. URL= <http://www.win.tue.nl/help/doc/demo.ps>

Desai, Bipin C. (1994). WebJournal : Visualization of Web. URL= <http://www.cs.concordia.ca/WebJournal.html>.

Desai, Bipin C., Shinghal, Rajjan (1994). A System for Seamless Search of Distributed Information Sources. URL= <http://www.cs.concordia.ca/w3-paper.html>.

Fletcher, J. (1993). Jumpstation. URL= <http://www.stir.ac.uk/jsbin/js>.

Gaynor, Edward (1994). Cataloging Electronic Texts: The University of Virginia Library, Experience. // Library Resources and Technical Services. 38 : 4 (1994) 403-413.

Giordano, Richard (1994). The Documentation of Electronic Texts Using Text Encoding Initiative Headers : An Introduction. // Library Resources and Technical Services. 38 : 4 (1994) 389-401.

Global Network Academy Meta-Library. URL= <http://uu-gna.mit.edu:8001/cgi-bin/meta>.

Horny, Karen L.(1986). Minimal-level cataloging: A look at the issues- A symposium. // Journal of Academic librarianship. 11 (1986) 332-334.

William A. Katz. Introduction to Reference Work. New York: McGraw-Hill, 1987.

Koster, M. ALIWEB(Archie Like Indexing the WEB). URL= <http://web.nexor.co.uk/aliweb /doc/aliweb.html>.

Koster, M. Simple Unified Search Interface (SUSI). URL= <http://web.nexor.co.uk/susi /susi.html>.

Koster, M. Configurable Unified Search Interface. URL= <http://web.nexor.co.uk/public/cusi /cusi.html>.

Library of Congress. MARC manuals used by the Library of Congress. Chicago: American Library Association, 1969.

McBryan, Oliver A. World Wide Web Worm. URL= <http://www.cs.colorado.edu/home/mcbryan/WWWW.html>.

Experimental Search Engine Meta-Index. URL= <http://www.ncsa.uiuc.edu/SDG/Software/ Mosaic/Demo/metaindex.html>.

NCSA Mosaic.URL= <http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSA MosaicHome.html>

Petersen, Toni ; Molholt, Pat (ed) (1990). Beyond the book: extending MARC for subject access. Boston: G.K. Hall, 1990.

Post, R. Lagoon : a WWW cache. URL= <http://www.win.tue.nl/lagoon>.

Ross, Rayburn M. ; West, Linda (1986). MLC: A contrary viewpoint // Journal of Academic librarianship, 11, 334-336.

Rhee, Sue (1986). Minimal-level cataloging : Is it the best local solution to a national problem?. // Journal of Academic librarianship. 11 (1986) 336-337.

Search WWW document full text. URL= <http://rbse.jsc.nasa.gov/eichmann/urlsearch.html>.

Taylor, Arlene G. (1994). The information universe: Will we have chaos or control?. // American Libraries. 25 : 7 (1994) 629-632.

Thau, R. SiteIndex Transducer. URL= <http://www.ai.mit.edu/tools/site-index.html>

WebCrawler. URL= <http://www.biotech.washington.edu/WebCrawler/WebQuery.html>.

World Wide Web Catalog. URL= <http://cui_www.unige.ch/cgi-bin/w3catalog>.