Limitaciones del *WWW* en el ámbito de la información documental

Por **Juan Antonio Pastor Sánchez**

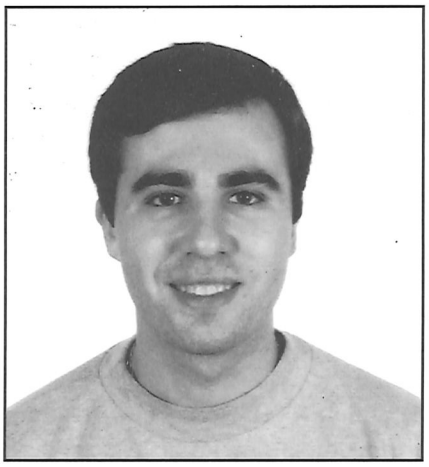
HOY DÍA podemos afirmar que el *World Wide Web* se ha convertido en la herramienta más difundida para el *almacenamiento, recuperación y difusión de la información (ardi)* en Internet. El éxito del *WWW* se debe principalmente a la naturalidad con la que se accede a la información: de forma no secuencial. De este modo el web se comporta como un sistema de hipertexto.

Sin embargo el *WWW* se concibió en un principio como un medio para el intercambio de información, para la difusión de resultados científicos y la comunicación entre investigadores y no para el *ardi*. Es por este motivo por lo que el web tiene algunas disfuncionalidades cuando se aplica con esta finalidad. Aún así hay que tener en cuenta que se trata de una herramienta en continua evolución, abierta y fácilmente ampliable.

Por otro lado su arquitectura cliente/servidor nos ofrece, en definitiva, un sistema flexible y adaptable a las nuevas exigencias que aparezcan con el paso del tiempo. La problemática del *WWW* se basa en que este sistema es una implementación muy reducida del concepto de hipertexto.

Limitaciones principales: planteamiento de posibles soluciones

En primer lugar existe una escasez en la tipología de relaciones del *WWW*. Básicamente sólo existe un tipo de enlace informativo: la *referencia*. Con esta relación nos movemos entre documentos web de un modo rápido e intuitivo; pero el núcleo del acceso descentralizado y no secuencial a la información exige nuevas relaciones.



Juan Antonio Pastor Sánchez

«Sería necesaria la coordinación entre distintas organizaciones para la creación y gestión de un tesauro que pudiera ser utilizado en el ámbito de la información existente en Internet»

Una relación *nota* nos permitiría disponer de una ventana virtual para aclaraciones; la relación *expansión/contracción* permitiría ampliar información desplazando hacia abajo el texto existente a continuación del enlace. De este modo se podría bascular entre dos estados: estado expandido y estado normal (contraído). También debería desarrollarse la relación de *incluye/excluye* con un efecto visual similar al anterior, con la diferencia de que la información utilizada estaría contenida en otro documento; de este modo se produce un efecto de reutilización de la información. Sería recomendable dotar a los distintos documentos de algún tipo de estructuración jerárquica entre ellos. Esto se conseguiría a partir de una relación *todo/parte* que indicaría cuál es el documento (o documentos) «padre» del que estemos consultando y cuáles son sus respectivos documentos «hijo».

En el fondo el *WWW* es muy limitado en cuanto a la presentación y consulta de la información. Además carece de herramientas conceptuales propias de búsqueda de documentos. Hay que incidir en que este sistema se creó para comunicar e intercambiar información y no para almacenarla. En el momento que se utiliza para el *ardi* se aplican herramientas de gestión textual, extrayendo palabras para construir índices. Estos índices se utilizan a través de un lenguaje de búsqueda. En el momento que se habilitan este tipo de utilidades para buscar páginas web se crea un curioso híbrido entre sistema de hipertexto y gestor documental. Sin embargo esta unión que podría parecer tan fecunda se vuelve ineficaz debido a que ambos conceptos se aplican de un modo muy limitado.

La estructuración de la información a partir de enlaces de hipertexto entre documentos multimedia requiere métodos de búsqueda más adecuados, partiendo para ello del propio concepto de enlace, lo cual nos conduce al concepto de *red semántica*. Una *red semántica* para buscar información en el web se traduciría en una serie de términos enlazados entre sí por relaciones de distinto tipo. En el campo de la Documentación este tipo de herramienta ya existe y es muy utilizada: el tesauro.

Aunque la noción de tesauro tendría que ser adaptada para su uso en Internet ofrecería un modo más natural de acceso a la información, sin necesidad de recurrir necesariamente a herramientas de búsqueda textual, que siempre podrían ser utilizadas de forma alternativa. En realidad existen «servidores de información» que complementan el uso de un lenguaje de búsqueda con clasificaciones más o menos elaboradas. El inconveniente es que cada servicio ha desarrollado una clasificación propia y totalmente distinta a las demás. Además estas clasificaciones son muy genéricas e inservibles en áreas del conocimiento especializadas.

Cont. en la pág. 12

Cambios generales

Pese a estar hablando de un sistema basado en la arquitectura cliente/servidor hay que tener en cuenta que entre ambos elementos se encuentra un tercero: la información en forma de documentos web.

Sin embargo, un análisis descendente de los documentos web nos permitirá observar que la estructura de cada uno de ellos se basa en un formato cuyo aspecto visual viene dado por las especificaciones *html*. El *html* es el auténtico núcleo del *WWW*. Su evolución junto con la de los clientes permitirá una mayor flexibilidad en la creación de documentos web, y por tanto el abanico informativo es susceptible de ser ampliado. Esto es debido a que en muchas ocasiones el propio contenido informativo se ve limitado por el medio a través del cual se comunica, y el web no es una excepción. El servidor también se ve afectado por estos cambios, ya que muchas de las tareas que anteriormente realizaba ahora las lleva a cabo el cliente. Esto permite la descarga de tareas con la consiguiente reducción de los tiempos de transmisión (ver figura 1).

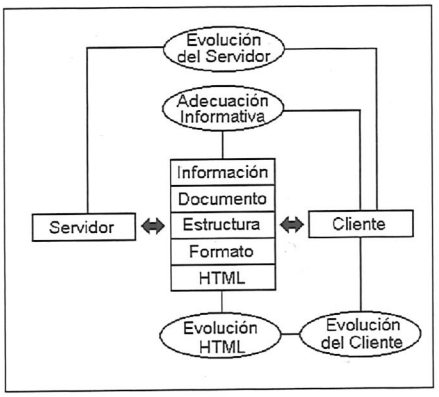


Figura 1

Por todo ello las soluciones se podrían plantear a partir de la ampliación del *html* (siempre y cuando el programa cliente asuma los cambios). Cabría la posibilidad de implementar nuevas funciones —como las que a continuación se exponen— en un ámbito reducido con un lenguaje de propósito general como *Java*, pero lo que realmente se necesita es un cambio en la organización de la información en el web a nivel general. Sin embargo se podría realizar una serie de experiencias iniciales utilizando *Java* u otro tipo de lenguaje de programación e incluso con aplicaciones que permitan interactuar con los clientes web.

Nuevas relaciones

En el caso de la implementación de nuevas relaciones podrían crearse fácilmente nuevas marcas que sólo afectarían al documento del propio usuario y a partes concretas del mismo documento (ejemplo de las relaciones *relnota*, *relexpansión* y *relincluye* del cuadro 1 para las relaciones de *nota*, *expansión/contracción* e *incluye*. La puesta en práctica de aquellas relaciones que afectaran al documento ajeno al usuario sería más complicada, ya que la propia arquitectura del *WWW* impide modificar los documentos remotos si no se poseen los correspondientes permisos. Sin embargo existen ciertas marcas *html* muy interesantes si fueran aprovechadas al máximo por los distintos clientes: *link* y *meta*.

La marca *link* podría ser utilizada para indicar relaciones partitivas o jerárquicas entre documentos (relaciones *parte/todo* o *incluido*); por su parte *meta* sería de utilidad como medio descriptivo de las páginas web. Sus atributos *Http-equiv* o *Name* podrían contener un resumen o algún tipo de indización. Por su parte los atributos *content* o *URL* podrían incluir dicha información dentro de la propia página o en otra externa a ella (ver cuadro 1).

La unión de dos redes

La aplicación del concepto de *red semántica*, y más concretamente de un tesauro, puede abordarse de varias formas. En primer lugar sería necesaria la coordinación entre distintas organizaciones para la creación y gestión de un tesauro que pudiera ser utilizado en el ámbito de la información existente en Internet. Esta tarea se puede dividir en áreas temáticas asignadas a entidades especializadas. La consulta del tesauro y de la indización se realizaría en forma de documentos web. El proceso de indización podría ser realizado por los administradores y creadores de los documentos indizados, que ya no tendrían que dar de alta su páginas en cada servidor/buscador de información sino que sólo debería recurrir a una única clasificación, lo cual permitiría el intercambio entre servidores de información con una estructura y codificación común. Los motores de búsqueda serían uniformes, permitiendo gran flexibilidad en la creación de interfaces adaptados a las necesidades de distintos tipos de usuarios.

```
CREACIÓN DE NUEVAS MARCAS

<RELNOTA> Texto de la nota <ANOTACION> Texto a mostrar en la ventana virtual </ANOTACION> </RELNOTA>
      <RELEXPANSION> Texto de la expansión <EXPANSION> Texto a mostrar a continuación del texto de la expansión </EXPANSION> </RELEXPANSION>
      <RELINCLUYE URL=url del documento incluido> Texto que indica el documento incluido </RELINCLUYE>

ADAPTACIÓN DE MARCAS EXISTENTES

      <LINK HREF=URL del documento con el que se relaciona TITLE=texto que describe la relación>
      <META NAME=resumen CONTENT=Texto del resumen URL=Documento que contiene el resumen>
      <META NAME=índices CONTENT=Descriptores, términos o palabras clave de la indización URL=Documento que contiene la indización>
```

Cuadro 1. Nuevas relaciones

La otra posibilidad es quizás la más interesante y potente: incluir la consulta del tesauro, y opcionalmente la indización, en el programa cliente. Para ello habría que crear una serie de extensiones que configuren la estructura de un tesauro y las asignaciones realizadas entre descriptores y páginas web. De este modo sería posible que un conjunto de documentos utilicen un *microtesauro* e incluso un pequeño corpus documental totalmente indizado. Sería necesaria la ampliación del *html* o la creación de un nuevo tipo de lenguaje de marcas.

En el cuadro 2 podemos observar un ejemplo de este tipo de lenguaje, que es capaz de recoger información sobre el nombre del tesauro, los términos que contiene y las relaciones entre ellos. Asimismo contiene información sobre la indización o asignación de descriptores a determinadas direcciones *URL* en donde se encuentran los documentos web. Estas especificaciones a las que podría-

```
- Marcas Meta adaptadas para la aplicación de tesauros:

<META NAME=Tesauro URL=Documento con las especificaciones del Tesauro>
<META NAME=Indización URL=dirección del motor de interrogación del Tesauro>

- Ejemplo de un fichero html:

<TESAURO>
      <TITULO>Tesauro de Información Universitaria</TITULO>
      <TERMINOS>
            <TERM TIPO=Descriptor COD=1 TEXTO=Proceso de Matriculación>
            <TERM TIPO=Descriptor COD=2 TEXTO=Información al Estudiante>
      </TERMINOS>
      <RELACIONES>
            <REL TIPO=NT RELATOR=2 RELACIONADO=1>
            <REL TIPO=BT RELATOR=1 RELACIONADO=2>
      </RELACIONES>
</TESAURO>
<INDIZACION>
      <IND COD=2 URL=http://www.um.es/~um-siu/infoestu/infoestu.htm>
</INDIZACION>
```

Cuadro 2. Thesaurus Markup Language

mos llamar *TML (thesaurus markup language)* se componen de dos marcas principales: *<tesauro>* que estructura la información relativa al tesauro, e *<indización>* con los datos de asignaciones.

Se puede considerar un tesauro como un conjunto de términos relacionados entre sí; para ello existen dos secciones *<términos>* y *<relaciones>*. Cada término viene especificado por una marca *<term>* en donde se indica el tipo de término (descriptores, no-descriptores o campos semánticos), a continuación un código único e individual para cada elemento y por último su descripción textual. Las relaciones se incluyen dentro de la marca *<rel>*, que permite especificar el tipo de relación y los dos términos que se enlazan entre sí.

Las dos primeras líneas del cuadro 2 son un ejemplo de cómo se puede hacer referencia desde un documento al fichero que contenga las especificaciones del tesauro. En ocasiones es preferible utilizar los datos de indización contenidos en una base de datos que se puede interrogar a partir de un programa *CGI* o similar. Para esto también se puede utilizar una marca *<meta>* que indique dónde se encuentra el motor de interrogación.

Aunque a primera vista los ficheros *TML* puedan parecer engorrosos de gestionar no hay que olvidar que existen programas gestores de tesauros que pueden ser adaptados para generar listados en este tipo de formato.

¿Hacia dónde vamos?

Los profesionales de la documentación se han enfrentado a la saturación de información. Hoy día esta saturación está alcanzando a Internet. En ambos casos la aplicación de *Tecnologías de la Información* clásicas, basadas en bases de datos relacionales no ha tenido todo el éxito esperado debido a problemas de coordinación entre organismos y la velocidad de crecimiento del volumen de información. Las nuevas tecnologías de la información deben abandonar el concepto de «tabla» y explotar el de «red» y debe dejar de tratar la «interrogación» para centrarse en la «navegación».

En el caso de Internet la adopción de estructuras de almacenamiento en donde existan dos redes —la de documentos y la de conceptos— puede ayudar a controlar el problema. Estas redes deben tener una riqueza semántica de la que carece actualmente el web. La ampliación de la tipología de relaciones del *WWW* y la aplicación del tesauro a la red de documentos, junto con una indización de calidad, normalizada, homogénea y fácilmente accesible nos hacen concebir esperanzas en el caótico futuro que se puede vislumbrar.

Bibliografía sobre el tema

Ardö, A.; **Falcoz, F.**; **Morten, N.** y **Sandfaer, M.** 1995.«Improving resource discovery and retrieval on the Internet: The Nordic Wais/World Wide Web Project - Summary Report», 1995. En: *Documento Internet. http://www.ub2.lu.se/W4/summary. html*

Canals Cabiró, I. 1990. «Introducción al hipertexto como herramienta general de información. Concepto, sistemas y problemática». En: *Revista Española de Documentación Científica,* 1990. pp.685-709.

Carr, L. A. 1994. «Structure and hypertext», 1994. En: *Documento Internet. http://journals.ecs.soton.ac.uk/lacethesis.html*

Chen, H. y **Schatz, B. R.** 1995. «Semantic retrieval for the Ncsa Mosaic». En: *Documento Internet. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/chen/chenschatz.html,* 1995.

De Bra, P. M. E. y **Post, R. D. J.** 1995. «Information retrieval in the world-wide-web: making client-based searching feasible», 1995. En: *Documento Internet. http://www.win.tue.nl/win/cs/is/reinpost/www94/www94. html*

De las Heras, A. R. 1991. *Navegar por la información.* Madrid: Anonymous, 1991.

Eklund, J. 1995. «Cognitive models for structuring hypermedia and implications for learning from the world-wide-web», 1995. En: *Documento Internet. http://www.scu.edu.au/ausweb95/papers/hypertext/eklund/index.html*

García Marco, F. J. 1994. «Hypertexto y lenguajes documentales: retos y sinergias». En: *Documat' 94,* 1994. pp. 417-426.

Jones, S. 1993. «A thesaurus data model for an intelligent retrieval system». En: *Journal of Information Science,* 1993. pp. 167-178.

Mathe, N. and **Chen, J.** 1995. «Adaptive dynamic hypertext based on paths of traversal», 1995. En: *Documento Internet. http://www.cd.bgsu.edu/hypertext/adaptive/Mathe.html*

Mayfield, J. 1995. «Two-level hypertext models as an underpinning for AHSs», 1995. En: *Documento Internet. http://www.cs.bgsu.edu/hypertext/adaptive/Mayfield.html*

Nelson, T. H. 1988. «Managing immense storage: project Xanadu provides a model for the posible future of mass storage». En: *Byte,* 1988. pp. 225-238.

Pastor Sánchez, J. A. y **Saorín Pérez, T.** 1995. «El hipertexto documental como solución a la crisis conceptual del hipertexto. El reto de los documentos cooperativos en redes». En: *Cuadernos de Documentación Multimedia,* 1995. pp. 41-56.

Rada, R.; **Akmal, Z.**; **Geeng-Neng, Y.**; **Michailidis, A.** y **Mhashi, M.** 1991. «Collaborative hypertext and the MUCH system». En: *Journal of Information Science,* 1991. pp. 191-196.

Juan Antonio Pastor Sánchez. *Servicio de información universitario. Unidad de información y documentación. Universidad de Murcia.*
Campus de Espinardo. 30071 Murcia.
Tel.: +34-68-30 71 00, ext. 2612; fax: ext. 2613
pastor@siu.um.es