Recuperación de información en el World Wide Web: planteamiento, herramientas y perspectivas

Jesús Tramullas Saz
Universidad de Zaragoza (Zaragoza)
webmaster@jabato.unizar.es

RESUM

S'aborda la problemàtica existent en els processos de recuperació d'informació en l'àmbit del World Wide Web. S'estableixen les bases que sustenten les pàgines Web, veritables documents hipermèdia virtuals. Són objecte de definició i anàlisi els diferents enfocs que l'usuari pot adoptar per a enfrontarse'n a la recuperació d'informació en WWW, i es proposen unes línies d'actuació per a instrumentar mecanismes, a diversos nivells, que millorin el procés d'identificació, accés i recuperació dels documents.

RESUMEN

Se aborda la problemática existente en los procesos de recuperación de información en el ámbito del World Wide Web. Se establecen las bases que sustentan a las páginas Web, verdaderos documentos hipermedia virtuales. Son objeto de definición y análisis los diferentes enfoques que el usuario puede adoptar para enfrentarse a la recuperación de información en WWW, y se proponen unas líneas de actuación para instrumentar mecanismos, en varios niveles, que mejoren el proceso de identificación, acceso y recuperación de los documentos.

1. La recuperación de información

El planteamiento de la recuperación de información (*Information Retrieval,* IR), en su moderno concepto y discusión, hay que buscarlo en la realización de los test de Cranfield, y en la bibliografía generada desde ese momento y referida a los mecanismos más adecuados para extraer, de un conjunto de documentos, aquellos que fuesen pertinentes a una necesidad informativa dada.[1] Las propias características de las entidades del mundo real, así como del tratamiento al que son sometidas, proveen a la representación de las mismas de un cierto nivel de indefinición. Es decir, que el proceso documental, por muy alto nivel de perfección que pueda alcanzar, siempre introduce un factor de distorsión en la representación del documento. Si se considera que el acceso al documento se realiza casi por completo utilizando esta representación como intermediario, puede deducirse que los mecanismos en los que se basa la IR no son perfectos, sino que se verán influenciados por ese factor, independientemente de su validez técnica.

Los problemas de la IR resaltan todavía más cuando el usuario se sitúa en un entorno informatizado (Rijsbergen, 1979). Los niveles de definición y de relación presentes en un sistema de bases de datos relacional, por ejemplo, son más precisos y rigurosos que los existentes en un sistema de bases de datos documentales, y este rigor y control permite la recuperación de los datos necesarios en un momento dado de una forma rápida y eficaz, utilizando criterios casi inequívocos. Por contra, la recuperación de información en un sistema documental, documático», si bien a nivel técnico es rigurosa, depende en gran manera de los criterios utilizados en la representación del contenido del documento, y en los criterios utilizados en la representación ideal del documento que se utiliza en la recuperación. Influyen entonces factores ajenos al propio sistema informático, más relacionados con el intermediario humano entre el documento original y el sistema informático.

Ha sido Blair (Blair, 1990) quien ha resumido las diferencias entre *data retrieval* (recuperación de datos, RD) e *information retrieval* (recuperación de información, RI), utilizando como criterios las siguientes cuestiones:

1. Resulta imposible abordar en este trabajo la rica investigación inter y transdisciplinar desarrollada en el campo de la Recuperación de Información, por lo que se realiza un breve resumen. Pueden encontrarse referencias bibliográficas básicas sobre RI en la bibliografía final. Véase García Marco, 1995.

a. Según la forma de responder a la pregunta: en RD se utilizan preguntas altamente formalizadas, cuya respuesta es directamente la información deseada. En RI las preguntas resultan difíciles de trasladar a un lenguaje normalizado, y la respuesta es un conjunto de documentos que pueden contener, sólo probablemente, lo deseado, con un evidente factor de indeterminación.

b. Según la relación entre el requerimiento al sistema y la satisfacción de usuario: en RD la relación es determinística entre la pregunta y la satisfacción. En RI es probabilística, a causa del nivel de incertidumbre presente en la respuesta.

c. Según el criterio de éxito: en RD el criterio a emplear es la corrección y la exactitud, mientras que en RI el único criterio de valor es la satisfacción del usuario, basada en un criterio personal de utilidad.

d. Según la rapidez de respuesta: en RD depende del soporte físico y de la perfección del algoritmo de búsqueda y de los índices. En RI depende de las decisiones y acciones del usuario durante el proceso de interrogación.

La comparación entre ambos tipos de recuperación resulta un punto clave para nuestro análisis, si se tiene en cuenta a qué criterios, principios y métodos responde la información dinámica en soporte electrónico, dispuesta a través de redes de telecomunicaciones. Es forzoso aceptar la diferencias entre las bases de datos distribuidas o federadas (Saltor, 1992), y las características que se encuentran en los repositorios de información (ya que consideramos que determinados elementos presentes en Internet deben ser considerados así, y no propiamente bases de datos federadas).

2. Internet: el World Wide Web como espacio de información electrónica

En Internet pueden encontrarse dos tipos principales de recursos: informativos y herramientas. Por recursos del tipo herramientas se entienden todos aquellos ficheros que ofrecen o contienen útiles capaces de solucionar problemas o necesidades del usuario, generalmente aplicaciones de *software*, bajo las denominaciones *freeware o shareware*. Estas se obtienen mediante transferencia de ficheros de un ordenador remoto, se instalan en el ordenador del usuario para resolver los problemas planteados, y no son objeto de interés en este momento.

Más complejos, por su variedad, son los recursos informativos. Hasta el momento, se han diferenciado atendiendo al método de acceso a la información que debía utilizarse para aprovechar ésta, lo que daba como resultado la división clásica en correo electrónico, terminal remoto, Gopher y World Wide Web (Ubieto, 1995). La propia evolución de Internet se ha encargado de limitar en su justa medida esta división. En el momento actual, un usuario puede acceder de forma centralizada y única a todos los tipos de recursos informativos existentes en Internet casi con una única herramienta (en este caso los clientes Web) (García y Tramullas, 1996).

2.1. Los repositorios de información

El concepto de repositorio de información es más reciente, complejo y potente que el tradicional concepto de bases de datos. Aunque se ha asociado tradicionalmente a los sistemas y herramientas *Computer Aided Software Engineering* (CASE), y ampliado hacia la idea de diccionario de recursos de información (Miguel y Piattini, 1995), los términos repositorio de información adquieren su pleno significado en un entorno en el que se desarrollan y explotan bases de datos multimedia (Chorafas, 1994). En este ámbito, un repositorio de información es un almacén en el que se encuentran situados diferentes ficheros que pueden contener información y datos en diferentes presentaciones. Estos datos e informaciones son combinados en un documento multimedia, que actúa como marco integrador de los mismos, de forma que el usuario pueda tener una visión integrada de la información, independientemente de la disposición y estructura física de la misma.

En realidad, un documento o página Web se construye siguiendo los principios explicados arriba. La página Web en sí es un marco, un contenedor, en el cual se coloca la información, independientemente del tipo que se trate, de acuerdo a los criterios del creador de la página. Las etiquetas HTML cumplen la función de indicar de que tipo de información se trata y cómo debe mostrarse al usuario, señalando además dónde se encuentra situada, ya que las imágenes, sonidos, etc. que se muestran en una página Web no forman parte de la misma físicamente: son ficheros independientes que pueden modificarse o cambiarse en un momento dado, alterando o no el contenido informativo de la página.

2.2. Los documentos multimedia e hipermedia

Los documentos multimedia se basan en la integración, en un mismo marco cognitivo, de informaciones de diferentes tipo, complementarias unas de otras, que ofrecen un contenido informativo completo. Este tipo de documentos pueden añadir elementos y utilidades hipertextuales, con lo que el usuario se sitúa entonces ante un documento hipermedia (Martin, 1990). En todos los documentos hipermedia se pueden diferenciar dos jerarquías, interesadas en la lógica del contenido del documento y en la física del documento formateado (Tramullas, 1996). Atendiendo a las mismas, una página Web, tal y como la percibe un usuario, corresponde a la jerarquía lógica del contenido del documento, y puede llegar a ser muy diferente de la jerarquía física del documento formateado, definida por las etiquetas HTML y por la localización y organización de los ficheros en el ordenador (u ordenadores).

Las páginas o documentos Web responden meridianamente al concepto y contenidos de lo que debe ser un documento hipermedia. Además, también ofrecen características clásicas de los sistemas de bases de datos distribuidas. Sin embargo, resulta curioso resaltar que en todo este análisis faltaría un elemento clásico en estos esquemas: la aplicación que gestiona el documento hipertextual, que gestiona la base de datos documental e hipermedia distribuida. Sólo son necesarios un lector o visualizador (el cliente Web), y un simple editor de textos (aunque haya herramientas especializadas como WebEdit). La aplicación que envía la información tampoco es un sistema hipermedia o de bases de datos; es un «demonio», sólo ocupado en enviar y recibir.

Todo esto nos da base para considerar Internet, desde el momento en el que apareció la herramienta Gopher, y por supuesto en la actualidad con World Wide Web, como un espacio de información electrónica inmenso, compuesto en su mayor parte por documentos electrónicos hipermedia, los cuales no limitan sus funciones a meramente informativas, sino que son capaces de interactuar con el usuario y con herramientas clásicas de gestión de bases de datos, sirviendo como intermediario. En alguna ocasión hemos definido Internet como una gigantesca base de datos documental distribuida (Tramullas, 1996a); sin embargo, este mismo concepto, que debe formularse de otra manera, falla en la suposición de que toda base de datos necesita un sistema de gestión de bases de datos para su desempeño, sistema que no aparece en Internet. Por ello, es preferible hablar de Internet como de un gigantesco sistema hipermedia distribuido, a nivel mundial, de publicación y acceso a la información, en el cual los límites entre documento, información, acceso a la información o soporte resultan todavía difusos.

3. Recursos para la recuperación de información en Internet

La carencia de un sistema hipermedia o de sistema de gestión de bases de datos, aunque es una ventaja notable, ya que permite publicar de manera libre y descentralizada, con una total libertad a nivel de presentación y de contenido, da como resultado la carencia de sistemas de recuperación de información propios, por lo que es necesario utilizar mecanismos independientes para realizar esta tarea. La gran cantidad de recursos existentes en Internet ha provocado la aparición y desarrollo de herramientas que buscan facilitar al usuario la localización y acceso a aquella información o ficheros que le sean necesarios. Estas herramientas se pueden encontrar para cada uno de los servicios y aplicaciones existentes en Internet.

Archie, WHOIS, Netfind, Veronica, WAIS... son los nombres que identifican a estas aplicaciones (Gilster, 1996), cuya organización y funcionamiento responde a la de un sistema de bases de datos que reúne, de forma automática (utilizando un programa agente[2]), en una base de datos el nombre, localización, y en ocasiones contenido, de los diferentes tipos de recursos. La bases de datos son actualizadas regularmente según criterios establecidos por sus administradores. Estas bases de datos son consultables por cualquier usuario utilizando un conjunto de expresiones regulares (en numerosas ocasiones derivadas de las *regexp* de UNIX), el cual obtiene como respuesta a su requerimiento un listado de los recursos que cumplen las condiciones y la localización de los mismos, utilizando una notación URL.

Los servicios señalados no ofrecen grandes problemas, ya que la recuperación utiliza términos y expresiones simples. Mención especial merece WAIS (Marchionini, Barlow y Hill, 1994), una herramienta magnífica para la gestión de todo tipo de información, especialmente textual, que no ha alcanzado el éxito que se le auguraba, en parte por el auge de los robots o engines de los que se tratará más adelante, en parte por el paso a una empresa privada de los nuevos desarrollos de WAIS.

---

2. Entendemos como *agente*, en el modelo cliente-servidor, la parte del sistema que realiza la preparación e intercambio de información por cuenta de una aplicación del cliente o del servidor. (tomado de RFC 1208, *A Glossary of Networking Terms*. 1991).

4. Recuperación de información en el ámbito del World Wide Web

4.1. Enfoques en la localización de información

Las herramientas de recuperación de información han experimentado un desarrollo muy importante con el auge del World Wide Web. La expansión de documentos hipermedia directamente accesibles demandaba la creación de sistemas que permitiesen recuperar los documentos Web que fuesen de interés para un usuario, de entre todos los existentes. Como respuesta, los usuarios de Web fueron desarrollando varios enfoques que diesen cumplida satisfacción a las necesidades de información.

En un primer momento se adoptó la solución de recopilar índices temáticos, que agrupaban los servidores Web según su contenido, de forma acorde a unas categorías dadas, que pretendían reflejar el contenido informativo de la infoestructura (University of Michigan School of Information, 1996). Sin embargo, este enfoque, del cual es conocido representante el servicio ofrecido por Yahoo!, ofrecía problemas en la adecuada adscripción temática de los servidores, por los diferentes conceptos utilizados por los usuarios, así como por la forzosa limitación a una descripción primaria del contenido. Estas limitaciones no deben ser óbice para reconocer el importante papel que cumplen en el complejo proceso de recuperar información en Internet.

Sin embargo, la utilización de directorios imposibilitaba el acceso a los recursos informativos a través del contenido de los documentos que ofrecían. Resultaba necesario encontrar un mecanismo que permitiese acceder directamente no ya a los servidores, sino a aquellos documentos Web que satisficiesen las necesidades informativas de los usuarios. Este es el origen de los robots, *spiders* o *wanderers* (Koster, 1996), como se les denomina, que rastrean automáticamente el ámbito Web para acceder a los servidores HTTP, recuperar los documentos Web que contienen, indizar su contenido textual, incorporarlo a una base de datos propia, y utilizar los punteros hipertextuales e hipermedia que incorporan para localizar nuevos servidores no incluidos en la base de datos general (Cheong, 1996). Los documentos obtenidos de esta forma se indizan y se tratan de manera similar a lo que haría un sistema de gestión de bases de datos documentales (Tramullas, en prensa), incorporándose a una base de datos. Por último, se provee una interfaz Web para que el usuario consulte los contenidos de la base de datos del motor de búsqueda (Tyner, 1996). La respuesta adopta la forma de un nuevo documento Web que el usuario puede utilizar como punto de partida para iniciar un nuevo proceso de exploración.

Una especialización de este sistema es aquél en el cual son los propios servidores HTTP los que indizan su contenido, para enviar posteriormente esta indización a una base de datos central, sobre la que los usuarios ejecutan sus consultas.

4.2. Problemas presentes en la recuperación de información en el ámbito Web

La recuperación de información en Internet ofrece similares problemas a los planteados en la RI clásica. En primer lugar, debe considerarse que los documentos Web ofrecen diferentes tipos y contenidos, no respondiendo todos ellos a lo que podría esperarse de un documento tradicional. Pueden encontrarse documentos con contenido informativo, documentos que contienen índices o directorios, documentos que sólo ofrecen enlaces o punteros hipertextuales, o combinaciones de todos ellos, en la más pura concepción hipertextual (Landow, 1995). En segundo lugar, los elementos de HTML (etiquetas) utilizados para crear documentos Web están pensados para la publicación, como derivados del SGML que son, más que para la estructura informativa de un documento. Además, gran parte de los documentos Web existentes han obviado la creación de etiquetas o campos específicos para la inclusión de palabras clave, descriptores o resúmenes, y sólo la lenta implantación de HTML 3.2 está solventando, en parte, esta limitación. Puede deducirse que la no consideración de los problemas presentes en la RI, durante el proceso de diseño y creación de los documentos Web, es la fuente de las limitaciones existentes en la búsqueda, desde cualquier perspectiva, de documentos que resuelvan las necesidades de los usuarios.

A esta limitación se unen factores psicológicos presentes en el usuario final. Latente ya de por sí en numerosos usuarios una ansiedad notable cuando se enfrentan a herramientas informáticas, por la teórica falta de control sobre el proceso informático o por la falta de conocimientos (Fariña y Arce, 1993), a esta ansiedad se une el problema que genera la avalancha de información, en dos facetas: la desorientación y la sobrecarga cognitiva (Díaz, Catenazzi y Aedo, 1996). La primera de ellas se relaciona con la posible falta de puntos de referencia en las estructuras de las páginas Web, mientras que la segunda se produce por la gran cantidad de información valiosa que el usuario puede recibir, pero que resulta inabarcable de una forma integrado, siendo necesario iniciar una navegación que puede derivar lejos del objetivo fijado inicialmente.

5. Herramientas de recuperación en el World Wide Web

Han sido dos los enfoques principales utilizados para la recuperación de información en el World Wide Web, como se ha señalado en un punto anterior (*v. supra*). Para el objeto de este trabajo van a resultar de interés aquellos en los cuales el usuario utiliza como elemento fundamental, en un primer momento, la composición y ejecución de ecuaciones de búsqueda.[3] Por esta razón no se va a abordar la problemática de los directorios, su contenido y posibles desarrollos. Además, debe considerarse que cada vez mayor cantidad de directorios están incluyendo motores de búsqueda, aunque sean internos.

Por lo tanto, van a ser objeto de análisis aquellas herramientas que ofrecen al usuario la posibilidad de formular, interactivamente, ecuaciones de búsqueda, cuya ejecución supone la recepción, como respuesta, de un listado de documentos o páginas Web que, teóricamente, responden a los criterios establecidos en la ecuación. Esta interacción puede tener lugar de cuatro formas diferentes:

1. **Directa.** El usuario, utilizando un cliente Web, se conecta al servidor Web que contiene el motor de búsqueda correspondiente a la base de datos que desea consultar. Éste le envía una página Web que actúa como interfaz de interrogación, y mediante la cual se formula y envía la consulta. El servidor la recibe, procesa, y envía a su vez como respuesta una nueva página Web en la que reúne las diez o veinte respuestas más acordes, según criterios internos, al requerimiento formulado por el usuario.

2. Por **intermediario.** El usuario, utilizando un cliente Web, formula la consulta utilizando una interfaz ofrecida por un servidor Web diferente del que ofrece la base de datos, es decir, un intermediario para el acceso. Éste adopta la forma de una página Web que contiene un mecanismo de interrogación similar al ofrecido por el robot original. De esta forma el motor de búsqueda se libera del trabajo de enviar la página que contiene la interfaz de interrogación, y sólo procesa la consulta y envía al usuario la respuesta. Pueden encontrarse incluso aquellos que la envían a varios motores de búsqueda al mismo tiempo, utilizando una única consulta.

3. Por **agente.** El usuario instala en su ordenador una aplicación que permite formular ecuaciones de búsqueda, y remitirlas directamente a los motores de búsqueda, sin necesidad de utilizar un cliente Web. La respuesta puede adoptar varias formas (listado general, página Web nueva...), y puede ser objeto de mecanismos de filtrado y de depuración para eliminar duplicados, etc. Estas aplicaciones ofrecen funcionalidades relacionadas con el almacenamiento, procesado y ordenación de las respuestas de forma combinada, utilizando los clientes Web para visualizar los documentos o páginas Web correspondientes a las respuestas.

4. Por **robot personal.** Se trata de aplicaciones que se instalan en el ordenador del usuario, y que son capaces de acceder a un servidor Web, construir un mapa e índices de sus contenidos, y utilizar los mismos para acceder a la información que sea interesante para el usuario. El mapa, índice o base de datos creada se mantiene en el ordenador del usuario, y puede (y debe) ser actualizada regularmente. Cuando el usuario requiere esa información, el robot lanza al cliente Web local en busca de la misma.

5.1. Directa

En esta aproximación el usuario utiliza directamente la interfaz ofrecida por el servidor Web que provee acceso a la base de datos (Codina, 1996). Cada motor de búsqueda ofrece una interfaz Web propia, con opciones de búsqueda simple y avanzada, y envía como respuesta una (o varias páginas Web) con punteros que enlazan con aquellas páginas contenidas en su base de datos que responden a los requerimientos del usuario (Zorn, 1996). Pueden encontrarse numerosos trabajos que versan sobre la cobertura de los diferentes motores de búsqueda, los periodos de actualización, los criterios de ponderación y refinamiento que emplean, la estructura de sus algoritmos, etc. (Vaquero y García, 1996). Algunos de estos servicios están ofreciendo ahora a sus usuarios la posibilidad de acceder a ellos directamente, utilizando complementos (o *add-ons*) para los clientes Web. De esta forma AltaVista[4] ofrece, a los miembros del AltaVista Visionary Club, la utilidad VistaPass (programada en JavaScript), Excite[5] ofrece Excite! Direct (que se añade a la barra de botones del cliente), y Lycos[6] ofrece Lycos Quick Search.

---

3. Véase la introducción de Bocher e Ihlenfeldt (1996), y la recopilación disponible en <http://www.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Searching_the_Web/>.

4. AltaVista <http://www.altavista.digital.com>.

5. Excite <http://www.excite.com>.

6. Lycos <http://www.lycos.com>.

5.2. Por intermediario

Evolución de la anterior es la disposición, en numerosos servidores Web, de formularios o interfaces de interrogación, de forma que el usuario no tenga que acudir al original cada vez que desea ejecutar una búsqueda, sino que puede utilizar el situado en un servidor más cercano, con lo que se descarga la red de tráfico, y se agiliza el trabajo de los motores de búsqueda, que se ven liberados del trabajo de enviar los interfaces de interrogación, ya que esta tarea la realizan terceros. Pueden encontrarse dos tipos principales:

1. Aquellos que ofrecen un formulario independiente para cada motor de búsqueda.
2. Aquellos que ofrecen un único formulario, que envía la consulta a varios motores de búsqueda de forma simultánea.

La respuesta es enviada directamente por el motor de búsqueda, aunque en ocasiones es tratada por el intermediario con la finalidad de, por ejemplo, evitar duplicados. Otra de las ventajas de estas interfaces es la posibilidad de traducirlas al idioma del país en el que se sitúan, o de incluir referencias y ayudas a la interrogación en el mismo, lo que alivia la situación de aquellos usuarios con dificultades en la comprensión del inglés escrito. En esta categoría de intermediarios deben englobarse servicios como los ofrecidos en All4One Search Machine, Savvy Search, Metacrawler o MetaSearch,[7] y en nuestro ámbito por el Servicio Retiarius del Servidor Web Jabato (Tramullas, 1996).

5.3. Por agente

Las limitaciones presentes en la gestión de las respuestas a los requerimientos de los usuarios han sido determinantes en el desarrollo de aplicaciones capaces de interrogar, sin utilizar un cliente Web, varios motores de búsqueda, ofreciendo además la posibilidad de procesar las respuestas para refinar su contenido, evitar duplicaciones o redundancias, y construir bases de datos con los documentos resultantes en el ordenador del usuario final. Para ello utilizan agentes (v. *supra*; Eichmann, 1994; Lawrence, 1995), que lanzan las consultas contra los motores, usando las listas de los mismos que pueden crearse o manipularse por el usuario final. Las respuestas recibidas pasan a formar parte de una base de datos local, en la que pueden incluirse los datos recibidos del motor de búsqueda, o bien pueden utilizarse para obtener el documento original, que se visualiza utilizando un cliente Web. Por el momento no incluyen mecanismos que faciliten la indización automática de los contenidos de un servidor Web, tomando como punto de partida una de sus páginas Web. Esta base de datos puede consultarse mediante los mecanismos clásicos (búsqueda booleana), pueden establecerse periodos de actualización automática de las consultas, aplicar mecanismos de eliminación de duplicados, pueden generarse nuevas páginas Web locales con los resultados de las búsquedas. La más conocida de estas aplicaciones que utilizan agentes de interrogación es Quaterdeck WebCompass,[8] actualmente en la segunda beta de su versión 2.0. Aunque menos conocida, sin embargo, es competitiva con ésta la aplicación WebSeeker,[9] desarrollada por ForeFront., actualmente en su versión 2.2. Deben citarse, aunque sus funcionalidades son inferiores, dos aplicaciones *shareware*, NetSearch 1.00,[10] similar a las dos anteriores, pero más limitada en lo que se refiere a la gestión de las respuestas, y WebSeek 1.0,[11] que muestra una barra en el entorno de Windows 95, mediante la cual se pueden enviar consultas a siete motores de búsqueda de forma independiente. Un agente de navegación, que actúa como herramienta de apoyo para evitar el desbordamiento cognitivo es una de las nuevas apuestas de IBM como complemento a los clientes Web.[12]

5.4. Por robot personal

La aproximación más compleja es la que utiliza un robot personal (situado en el ordenador del usuario final) para acceder a servidores Web, analizar todos sus contenidos, y generar una base de datos en la que se recogen los mismos (consultable utilizando lógica booleana, como mínimo), así como la capacidad de generar representaciones gráficas de la disposición de los recursos de información presentes en ellos. A esta exploración debe añadir la capacidad de navegar interactivamente de un servidor a otro, utilizando las referencias hipermedia externas presentes en los mismos (Eito, 1996). No son numerosas las aplicaciones de este tipo disponibles en la actualidad, ya que muchas de las que puedan citarse corresponden a desarrollos de proyectos de investigación en curso, todavía no disponibles ni como *free/shareware* ni como productos comerciales. Si es necesario recordar,

---

7. All4one Search Machine <http://all4one.com>, MetaCrawler <http://www.cs.washington.edu/research/projects>, All-in-One Search Page <http://www.albany.net/allinone>, Savvy Search <http://www.cs.colostate.edu/dreiling/smartform.html>.
8. *WebCompass es un producto de Quaterdeck <http://www.quaterdeck.com>.*
9. WebSeeker es un producto de ForeFront Group <http://www.ffg.com>.
10. Más información sobre NetSearch enviando un mensaje a <tagraham@alpha.wcoil.com>.
11. *Más información sobre WebSeek enviando un mensaje a <jeffhu@umich.edu>.*
12. *IBM Web Browser Intelligence (WBI)-Agent Software* <http://www.raleigh.ibm.com/wbi/wbisoft.htm>.

para máquinas UNIX, la existencia de TkWWW, aunque es necesario conocer el lenguaje tcl/tk para aprovechar toda su potencialidad, y de Fisch-Search, para clientes X-Mosaic, que también utiliza la navegación automática y las expresiones *regexp* de UNIX como criterios.[13] Más conocida es la aplicación CyberPilot Pro,[14] en su versión 2.0, para plataformas Windows 95/NT. Ésta permite la creación de mapas y de índices de contenidos de los servidores Web, así como una representación gráfica de la estructura de los mismos. Para disponer de los documentos o páginas Web originales, sigue necesitando ejecutar un cliente Web, al que enviará el URL del objeto a *obtener.*

Cuestión aparte son la gran cantidad de robots que pueden encontrarse, incluyendo su código fuente, en Internet.[15] Muchos de ellos son herramientas de indización, quedando a cargo del usuario la creación y consulta de las bases de datos derivadas, así como de sus interfaces de interrogación. Su compilación, instalación y aprovechamiento suelen requerir conocimientos de UNIX, por lo que suelen estar lejos de las capacidades de un usuario final común.

6. A modo de conclusión y perspectivas de desarrollo

La revisión de los mecanismos existentes en Internet para la *recuperación de información, en el ámbito* World Wide Web, pone de manifiesto que los problemas planteados por la *Information Retrieval* desde la década de 1960 siguen siendo permanente actualidad. Si bien han sido detenidamente estudiados en situaciones de entornos cerrados (como en una base de datos en la que se utilizan mecanismos de control terminológico), todavía está por delinear una teoría sobre la recuperación de información en documentos hipermedia distribuidos, en la línea de la reflexión realizada por Ingwersen: «Information retrieval is concerned with the processes involved in the representation, storage, searching and finding of information which is relevant to a requirement for information desired by a human user» (Ingwersen, 1992). En este marco se proponen varias líneas de actuación para una teoría y una praxis sobre la recuperación de información en el World Wide Web.

6.1. Actuación sobre la creación de documentos

En primer lugar, es necesario aplicar ciertos criterios comunes en la identificación de documentos. Aceptada ya de *facto* la notación que indica la localización del recurso de información que se trate, mediante su URL, es necesario utilizar elementos que permitan representar adecuadamente el contenido del documento. La utilización de las etiquetas <META> de HTML para incluir descriptores y resúmenes, que puedan utilizarse en la indización que llevan a cabo los robots, debería ser común. Cuestión más compleja es la adopción de lenguajes documentales, casi imposible, por lo que debería iniciarse alguna acción, por parte de la Internet Society, para establecer unos principios comunes en la utilización de descriptores. Además de las acciones sobre los documentos originales, deben citarse numerosas iniciativas tendentes a categorizar, de forma automática (mediante agentes) los recursos informativos existentes en World Wide Web (véasen los reunidos en *Project Aristotle(sm): Automated Categorization of Web Resources*).[16]

6.2. Actuación sobre desarrollo de herramientas

Las *herramientas existentes para la recuperación de información utilizan principios técnicos desarrollados en* las dos últimas décadas, y que pueden encontrarse en cualquier sistema de IR *clásico*. A pesar de la interactividad que se supone a la navegación en WWW, en recuperación de información el usuario sigue siendo pasivo, dependiendo sobremanera de las respuestas recibidas a sus ecuaciones o requerimientos. No todos los motores ofrecen la misma cobertura, ni utilizan los mismos criterios de indización. No todos los agentes personales incluyen mecanismos de exploración automática, ni los escasos robots personales ofrecen representaciones gráficas convincentes de las infoestructuras hipertextuales de los repositorios de información. En este ámbito, la capacidad de explorar automáticamente los servidores que contienen documentos pertinentes, de generar representaciones gráficas comprensibles de los mismos y ser capaces de discriminar aquellos documentos que no sean adecuados, según perfiles definidos por el usuario u obtenidos en el mismo proceso de exploración, deben ser comunes en la próxima generación de herramientas. Éstas deberán configurarse como verdaderos sistemas de identificación, recuperación y selección de información.

---

13. Más información sobre TkWWW en <http://fang.cs.sunyit.edu/robots/spetka.html> sobre Fish Search, Bra, P.M.E. de y Post, R.D.J. «Information Retrieval in the World-WideWeb: Making Client-based searching feasible» <http://www.win.tue.nl/win/cs/is/debra/wwwf94/article.html>.

14. CyberPilot Pro es un producto de NetCarta Corporation <http://www.netcarta.com>.

15. Véase la base de datos compilada por M. Koster, *The Web Robots Database*, disponible en <http://info.webcrawler.com/mak/projects/robots/active.html>.

16. Disponible en <http://www.public.iastate.edu/~CYBERSTACKS/Aristotle.htm>.

6.3. Actuación sobre formación de usuario

Uno de los mayores retos es la formación de los usuarios. Cualquier política de acceso a la información debe iniciarse por los mismos. Uno de los mayores problemas en la recuperación de información en el WWW en la actualidad no son los medios técnicos disponibles, sino la escasa formación que han recibido los usuarios en lo concerniente a la creación, organización y recuperación de recursos de información hipermedia distribuidos. Por lo tanto, las acciones a desarrollar deben ir encaminadas a: 1) hacer comprensibles las estructuras lógicas y físicas subyacentes a los documentos Web; 2) aprender el funcionamiento de las herramientas de recuperación; 3) establecer métodos y criterios de recuperación que combinen búsqueda (*searching*) con exploración visual (*browsing*); 4) utilizar mecanismos de asociación y predicción, y 5) desarrollar métodos de colaboración entre equipos y sistemas de recuperación. Los usuarios deben cambiar su enfoque tradicional, basado en la interrogación y comprobación, por una actitud más dinámica, en la que se complemente con navegación, evaluación y decisión (Tillman, 1996).

La combinación de estas líneas debe llevar a una mejora y especialización de las interfaces de las aplicaciones de recuperación, y a la progresiva integración de técnicas de tratamiento del lenguaje natural y de reconocimiento de relaciones entre conceptos, expresados con términos. En este campo deberán integrarse de forma consistente técnicas de inteligencia artificial (Teenor, 1995). Ello sin olvidar la aparición de nuevos agentes encargados, no ya de buscar la información, sino de filtrar (Takkinen, 1996) la que recibe el usuario, eliminando aquella que sea inútil o redundante. La recepción pasiva de información, por ejemplo a través de correo electrónico, demandará la implementación de agentes que actúen como barrera frente a un desbordamiento informativo. El panorama que se dibuja en un futuro cercano corresponde a agentes especializados que buscan, filtran y presentan información que responda a perfiles de usuario previamente definidos (Belkin y Croft, 1992), liberando el trabajo del usuario del tiempo que dedica a estos menesteres, para emplearlo en la evaluación, análisis y utilización de la información recibida.

Bibliografía

BELKIN, N; CROFT, W.B. (1992). « Information filtering and information retrieval: Two Sides of the Same Coin?» *Communication of the ACM*. Vol. 35, nº12, p.29-38.

BLAIR, D.C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier.

BOCHER, B.; IHLENFELDT, K. (1996). *A Higher signal-to-noise ratio: effective use of Web search Engines*. <http://www.state.wi.us/agencies/dpi/www/search.html>.

BRA, P.M.E. de; POST, R.D.J . (1994). «Information Retrieval in the World-WideWeb: Making Client-based searching feasible» En: *Second International WWW Conference*. <http://www.win.tue.nl/win/cs/is/debra/wwwf94/article.html>.

CHEONG, F.C. (1996). *Internet agents: spiders, wanderers, brokers and Bots*. New Riders.

CHORAFAS, D.N. (1994). *Intelligent multimedia databases*. Englewood Cliffs: Prentice Hall.

CODINA, L. (1996). «Cómo descubrir información en Internet y cómo conseguir que nos descubran a nosotros» *Net Conexión*. Nº 13, p. 52-67.

DÍAZ, P.; CATENAZZI, N.; AEDO, I. (1996). *De la multimedia a la hipermedia*. Madrid: Ra-Ma.

EICHMANN, D. (1994). *Ethical Web Agents*. <http://rbse.jsc.nasa.gov/eichmann/www-f94/ethics/ethics.html>.

EITO BRUN, R. (1996). «Una nueva forma de recuperar información: los robots personales» *Information World en Español*. Nº 46, p.15-19.

ELLIS, D. (1990). *New Horizons in information retrieval*. London: The Library Association Pub.

FARIÑA, F.; ARCE, R. (1993). *Ansiedad ante los ordenadores*. Madrid: Eudema.

GARCÍA MARCO, F.J. (1995). «Paradigmas científicos en recuperación de información» En: García Marco, F.J. *Organización del Conocimiento en Sistemas de Información y Documentación*. Zaragoza: ISKO, p. 99-112.

GARCÍA MARCO, F.J.; TRAMULLAS SAZ, J. (1996). *World Wide Web: fundamentos, navegación y lenguajes de la red mundial de información*. Madrid: Ra-Ma.

GILSTER, P. (1996). *Finding It on the Internet. The Internet navigator's guide to search tools and techniques.* New York: John Wiley & Sons.

INGWERSEN, P. (1992). *Information retrieval interaction*. London: Taylor Graham.

KOSTER, M. (1996) *WWW. Robot Frequently Asked Questions*. <http://info.webcrawler.com/mak/projects/robots/faq.html>.

LANDOW, G.P. (1995). *Hipertexto: la convergencia de la teoría crítica contemporánea y la tecnología*. Barcelona: Paidós.

LAWRENCE, A. (1995). «Computing by proxy. (Intelligent agents)». *Computer Business Review*. Vol. 3, nº 4, p. 5-8.

MARCHIONINI, G.; BARLOW, D.; HILL, L. (1994). «Extending retrieval strategies to networked environments: old ways, new ways, and a critical look at WAIS». *JASIS*. Vol. 45, nº8, p. 561-564.

MARTIN, J. (1990). *Hypedocuments and how to create them*. Englewood Cliffs: Prentice Hall.

MIGUEL, A. de; PIATTINI, M. (1995). «Visión general de los repositorios y diccionarios». En Piattini, M. y Daryanani, S. *Elementos y herramientas en el desarrollo de sistemas de información*. Madrid: Ra-Ma. p. 229-239.

RIJSBERGEN, C. J. VAN. (1979). Information Retrieval. London: Butterworths. <http://dcs.glasgow.ac.uk/Keith/Preface.html>.

SALTOR, F. (1992). «Acceso integrado a bases de datos heterogéneas» En: *Encuentro sobre bases de datos en las administraciones públicas*. Madrid: MAP. p. 195-200.

TAKKINEN, J. (1996). *Information Retrieval and Information Filtering (IRIF)*. <http://www.ida.liu.se/labs/iislab/courses/IRIF/IRIF'litteraturlista.html>.

TEENOR, K. (1995). *Communications 515, Chapters 13-17.* <http://www5.fullerton.edu/viscom/COM515.html>.

TILLMAN, H. N. (1996). *Evaluating Quality on the Net*. <http://www.tiac.net/users/hope/findqual.html>.

TRAMULLAS SAZ, J. (1995). «Una introducción a la informática documental» En: *Apuntes CCUZ*. Nº 6, Zaragoza: Centro de Cálculo de la Universidad de Zaragoza. p. 6-10.

TRAMULLAS SAZ, J. (1996a). *Apuntes de Informática Documental*. Zaragoza: Kronos.

TRAMULLAS SAZ, J. (1996b). «Servidor Web Jabato: el servicio retiarius» En *Actas de las V Jornadas Españolas de Documentación Automatizada Documat 96*. Cáceres: Servicio de Publicaciones de la Universidad de Extremadura, p.295-303.

TRAMULLAS SAZ, J. (1997). *Documática: conceptos básicos*. Zaragoza, (en prensa).

TYNER, R. (1996). *Sink or swim: Internet search tools & technique*. <http://www.sci.ouc.bc.ca/libr/connect96/search.htm>.

UBIETO, A.P. (1995). *Documentación automatizada: manual de uso de la red Internet*. Zaragoza: Anubar.

UNIVERSITY OF MICHIGAN SCHOOL OF INFORMATION. (1996). *The Matrix Of Internet Catalogs and Search Engines*. < http://www.sils.umich.edu/~fprefect/matrix/matrix.shtml>.

VAQUERO PULIDO, R.; GARCÍA FIGUEROLA, C. (1996). Motores de búsqueda en Internet» En: *Actas de las V Jornadas Españolas de Documentación Automatizada Documat 96*. Cáceres: Servicio de Publicaciones de la Universidad de Extremadura. p. 621-629.

ZORN, P. *et al.* (1996). «Advanced Web searching: tricks of the trade». *Online*. Vol. 20, nº 3, p.14-28. <http://www.onlineinc.com/onlinemag/MayOL/zorn5.html>.