Recuperación en Internet: Cuatro modelos complementarios y una agenda para su integración

Carlos Benito Amat

Introducción

La recuperación de información en Internet no ha recibido hasta ahora tratamiento adecuado en la literatura española en información y documentación. López Alonso y Mares Marín han aportado un trabajo original donde evalúan el rendimiento de 6 sistemas [1]. Sánchez Montero ha tratado superficialmente el problema en el marco de su trabajo sobre diseño de contenidos para intranets corporativas [2]. Más recientemente, Baró ha aportado una mera relación, con descripciones muy genéricas de algunos sistemas de búsqueda en el entorno World Wide Web [3]. Más esquemática aún es la pretendida revisión de Senso [4], y la aportación al tema de Marcos Mora resulta más descriptiva de los servicios asociados que de los propios sistemas [5]. Una de las últimas aportaciones destaca por el número de sistemas analizados y por incluir entre ellos a 10 sistemas nacionales [6]. Sin embargo, la evaluación que su título anuncia se reduce al análisis de presencia o ausencia de una serie de características de representación de los documentos, posibilidades y opciones de búsqueda y elementos de recuperación. Los sistemas de recuperación de información distribuida en Internet también se han tratado de pasada en el marco de trabajos generales. En contraste, la literatura de divulgación informática trata muy a menudo el tema, y pueden mencionarse ejemplos recientes tanto a nivel internacional[7] como nacional [8], aunque la mayoría de estas publicaciones son meras descripciones o apenas sobrepasan el nivel de artículos de opinión.

El presente trabajo contrasta documentos y sistemas tradicionales con los de Internet, describe 4 sistemas de acceso y esboza un panorama y una agenda de acciones.

Los documentos tradicionales y los sistemas de recuperación de información estructurada

La expansión de la literatura científica y técnica del periodo de la inmediata postguerra y la guerra fría, tanto en términos meramente cuantitativos como en lo tocante a la aparición de nuevas disciplinas, subdisciplinas y especialidades, vino a coincidir con los primeros intentos de aplicar los computadores a tareas distintas del cálculo numérico. Al mismo tiempo, puso en evidencia las limitaciones de los grandes esquemas clasificatorios generalistas y abrió el campo para la investigación de sistemas postcoordinados de representación y recuperación de documentos. A este panorama vino a sumarse la necesidad de difusión y coordinación de servicios de las grandes instalaciones bibliotecarias [9].

Considerados globalmente, el conjunto de distribuidores y bases de datos accesibles online, sucesores directos del esquema esbozado, han dominado durante casi 30 años la recuperación de documentos en campos aparentemente tan dispares como la información científica y técnica, la documentación legislativa y jurídica y la información de actualidad. Una serie de características son comunes a los referidos servicios, a pesar de su contenido heterogéneo, y también delatan algunas de sus limitaciones.

Los sistemas mencionados han trabajado inicialmente con representaciones estructuradas de documentos impresos. Sólo el aumento en la capacidad de los dispositivos de almacenamiento permitió la distribución de bases de datos a texto completo, especialmente en los ámbitos legislativo y jurídico y en los de información de actualidad. Por lo que respecta al sector de la información científica y técnica, sigue estando constituído mayoritariamente por bases de datos referenciales o servicios de resúmenes. Diversas agencias se ocupan del acceso efectivo al documento original o su reproducción.

Los documentos representados en las bases de datos son, en su inmensa mayoría, de tipo textual. Sólo poco a poco productores y distribuidores comienzan poner en marcha dispositivos que permitan la representación, el almacenamiento y la recuperación de los elementos gráficos que contienen.

Todos los documentos son resultado de sucesivos procesos de edición o evaluación previos a su publicación. Uno de los resultados de estos procesos es la gran homogeneidad de los documentos científicos y técnicos, legales y jurídicos y periodísticos en sus respectivos grupos. También sus representaciones presentan un alto grado de homogeneidad. No podía ser menos, puesto que factores de tipo técnico y económico han propiciado el centralismo en la distribución de información, en manos de grandes grupos de comunicación sometidos a un proceso de concentración muy acusado.

A pesar de lo afirmado en los párrafos anteriores, subsiste el esquema que sitúa en fases sucesivas a productores y distribuidores de bases de datos.

Por otra parte, se ha observado tradicionalmente una tendencia hacia la especialización a nivel de la producción de bases de datos documentales. MEDLARS, Psycinfo, Compendex, Lexis, BIOSIS son sólo algunos ejemplos de sistemas dominantes en campos de conocimiento o de información especializados.

Finalmente, todos los sistemas mencionados cuentan con sistemas de recuperación de ajuste exacto (exact matching) apoyados en la lógica booleana. Esta característica, la necesidad de programas cliente y las restantes complejidades de sus lenguajes de consulta se han traducido en la necesidad de intermediarios especializados

Documentos en Internet y espacios informativos

En su discusión sobre la recuperación de información distribuida [10], Clifford Lynch tipifica la variada gama de objetos, servicios y flujos de información digitales distribuidos en Internet en dos grupos principales. En primer lugar, distingue los innumerables ficheros almacenados como archivos en muchos servidores, accesibles a través de protocolos como FTP y HTTP y que representan textos, imágenes, audio o vídeo digital o programas ejecutables. Cabría añadir a ellos recursos de un segundo tipo: los grupos de noticias y las listas de discusión, soportados por protocolos de transporte más especializados como NNTP (Network News Transfer Protocol) o IRC (Internet Relay Chat). Las páginas Web y los programas son simples colecciones de bits interpretables como texto ASCII o objetos binarios, al igual que los mensajes de correo, las noticias o las listas. Recientemente, un conjunto de normas de etiquetado (Multipurpose Internet Mail Extensions) ha permitido la transmisión, junto con los mensajes, de objetos digitales. A medida que se desarrolla el lenguaje HTML, también se enriquecen las páginas Web para adoptar la categoría de documentos compuestos.

El segundo grupo de recursos está representado por servicios interactivos accesibles a través de protocolos de emulación de terminal normalizados (Telnet, X Windows) o programas cliente especializados que emplean protocolos propietarios (como los clientes de servicios online). Gracias al empleo de pasarelas o "traductores", es posible incluir en este grupo a los servicios interactivos basados en SQL y a aquellos que emplean la norma Z39.50.

Por último, existen flujos de información efímera, que adoptan la forma de videos, audios, conferencias y otros recursos procedentes de diversas fuentes. Estas fuentes, análogas a emisiones de radio o televisión, se transmiten a través de Mbone (Multicast Backbone). Los usuarios de Internet pueden estar interesados bien en la existencia de flujos de información como un recurso activo y actual o bien en la simple localización de alguna parte específica de ese flujo de información, almacenada en algún archivo histórico relacionado.

Los espacios

Mauldin [11] ha formulado una distribución de contenidos informativos en un espacio global del que Internet es sólo una parte. El espacio Web, representado por la información accesible a través del Protocolo de Transferencia de Hipertexto (HTTP), como el espacio gopher, el que corresponde al protocolo de transferencia de mensajes de correo (SIMP) y el de transferencia de ficheros (FTP) se

encontrarían dentro de los llamados servicios propios de Internet. En contraste, el espacio de la información estructurada en bases de datos (SRIe), de las que una parte son accesibles a través de servidores WAIS, sustrae sus contenidos a la Red, por muchas pasarelas que se hayan diseñado para su consulta. Los grupos de discusión o grupos USENET tampoco están totalmente integrados en ese espacio informativo público que representa Internet, en la medida en que muchos de ellos se circunscriben a servicios online o redes privadas restringidos. El concepto de una red distribuida como un espacio informativo, que Lynch considera todavía en desarrollo16, resulta de interés en el sentido en que desplaza el énfasis hacia los contenidos. En efecto, cada uno de los "servicios" accesibles a través de la Red se caracteriza por un tipo determinado de contenido informativo accesible a través del programa de conexión desarrollado al efecto. El hecho de que los actuales programas clientes permitan el acceso a más de un espacio no obsta para considerarlos por separado, en función de los objetos informativos o recursos que contienen.

El primer elemento de contraste con el entorno de los documentos tradicionales y sus sistemas de acceso se basa en la naturaleza de los documentos. Frente a la abrumadora mayoría de las representaciones estructuradas de documentos textuales en el espacio de las bases de datos, Internet contiene documentos digitales íntegros codificados en una gran variedad de formatos. Los textos en diferentes juegos de caracteres ASCII, los ficheros audibles en formatos MIDI o WAV, las imágenes fijas GIF, JPEG, NEGF, las imágenes en movimiento AVI, MOV, MPEG o Quicktime son sólo algunos de los ejemplos más recurridos. Por lo que respecta a los textos, los formatos PDFy PostScript conviven con documentos preparados con diversos programas de procesamiento de texto.

Además, los documentos de Internet son compuestos. La información en ella contenida se combina en su preparación y en su visualización con diferentes códigos. Apenas pueden encontrarse páginas de Internet donde no coincidan caracteres textuales y representaciones gráficas. A medida que los lenguajes de marcas evolucionan, es más probable hallar combinaciones de elementos gráficos y textuales con sonoros.

La figura del productor de información es indistinguible de la figura del distribuidor de esa misma información en la Red. La preparación de contenidos y su publicación están a muy poca distancia. Esta simplificación de la secuencia tradicional de distribución de información conlleva otra característica que diferencia el ámbito de Internet: no existen procesos de edición y evaluación de la información y los documentos previa a su publicación.

La flexibilidad en el tratamiento de los datos y la multiplicidad de productores determinan otra característica diferencial de interés. Se refiere a la multiplicidad y redundancia de la información distribuida. Así, mientras en los ámbitos tradicionales se considera anomalía, fraude o excepción la publicación repetida de un documento, la redundancia es casi norma en Internet. El hecho de que sólo en España existan no menos de 5 destinos de Internet que contienen directorios electrónicos de bibliotecas españolas y extranjeras es una muestra de la "democratización" de la Red y sólo uno de los múltiples ejemplos de redundancia de contenidos.

La inexistencia de tradición y de concentración de distribuidores de información a través de Internet, junto con la gran diversidad de los soportes han causado una gran heterogeneidad en la estructura de los documentos. Esta heterogeneidad contrasta sobremanera con la poca variabilidad de los documentos y sus representaciones en el ámbito de las bases de datos. Es cierto que algunos documentos electrónicos, especialmente los de tipo transaccional basados en los protocolos SMTP y NNTP, cuentan con elementos obligados: el autor de un mensaje o un artículo, el receptor... Pero la gran mayoría de los documentos generados mediante HTML, abrumadores en número en la Red, no contienen metainformación de forma sistemática y ni siquiera cuentan con las mismas marcas en todos los casos. Véase, si no, la cantidad de páginas "untitled" que se recuperan tras cualquier interrogación simple.

La "democratización" que caracteriza el uso y los contenidos de la Red supone también una diferencia con el entorno de la información estructurada en bases de datos. Es posible que muchos de los productores de información distribuida en Internet sean especialistas en áreas de conocimiento

determinadas, pero no ocurre los mismo con los usuarios reales o potenciales de esa información, que es universal desde el punto de vista temático. Los usuarios, como público general, no están cualificados.

Además, frente a la especialización de las bases de datos en el entorno científico y técnico, se verá más adelante que los sistemas de recuperación de información distribuida en Internet pretender ofrecer un mismo nivel de cobertura a un mismo universo de objetos informativos. Claro está que existen servicios y sistemas especializados, pero tal especialización se basa en un tipo de documentos y de información determinados, concentrados en un sólo espacio informativo.

Por último, frente a la relativa estabilidad de los documentos, mayoritariamente impresos, representados en bases de datos estructuradas, y a la perdurabilidad de la información que contienen, la información en Internet está caracterizada por el dinamismo y la volatilidad. El dinamismo se refiere a los continuos cambios de contenido de muchos de los documentos de Internet. La volatilidad, a los cambios de destino de un mismo documento.

Los documentos

Los documentos del espacio Web son compuestos, altamente dinámicos, de moderado tamaño, de muy baja estructuración interna y, como es propio de este espacio, altamente interrelacionados.

En Noviembre de 1995, Open Text realizó un censo de los documentos Web existentes. Sobre una muestra de 1,524 millones de objetos, se halló que el 50% contienen al menos un enlace a una imagen y el 15% contienen exactamente una imagen. Las páginas que contenían un gran componente gráfico, lo hacían a costa de los típicos "bolarroja.gif" y similares [12]. En España, un reciente análisis de las sedes Web de 8 bibliotecas universitarias y 11 de otras instituciones catalanas reveló una proporción media general de textos e imágenes de 45 a 55% [13].

Por lo que respecta al dinamismo de este espacio, se pueden aportar muchas evidencias de su alta tasa de variabilidad. Así, tras el examen de dos conjuntos de documentos Web recopilados con 1 mes de diferencia (1,3 millones en Octubre y 2,6 millones en Noviembre de 1995), se observó empíricamente que muchos de los más populares URLs del primer conjunto ya no existían en el segundo [14]. En otro trabajo se muestrearon periódicamente 4.600 objetos HTTP distribuidos en 2.000 sedes diferentes durante un periodo de 3 meses. La vida de los objetos fue de 44 días como promedio. Para los objetos textuales el valor fue de 75 días y para las imágenes de 107. Otros documentos persistieron durante 27 días. El 28% de los objetos se actualizó como mínimo cada 10 días y un 1% se actualizó dinámicamente [15]. Según uno de los ingenieros de Infoseek, John Nauman, el 10% de las páginas indizadas en su base de datos ya no existen [16]. Una última evidencia, aunque se podrían aportar muchas más. La búsqueda por los mismos unitérmino y frase (4 palabras) arrojó diferentes resultados en 8 buscadores cuando se realizó en Febrero, Mayo y Noviembre de 1996.Los resultados de la búsqueda del unitérmino se decuplicaron (se multiplicaron por 10) en los pases extremos en Excite, Infoseek Guide, Lycos y WebCrawler. En AltaVista aumentaron de 20.000 a 30.000 y en OpenText de 1.026 a 3.758 [17].

Sobre el tamaño de los documentos Web también se han hecho diversas estimaciones. Por ejemplo, en Noviembre de 1995, el censo de Open Text reveló que una página ocupaba por términos medio 6518 bytes, con una mediana de 2021 y desviación típica de 31678 [18]. Del conjunto de 2,6 millones de documentos HTML recopilados por Inktomi en las mismas fechas, cada documento HTML ocupaba una media de 4,4 Kb (de =2 Kb). La extensión máxima fue de 1,6 Mb20. Las sedes recientemente analizadas por Térmens contienen un total de 4.277 páginas que ocupan un total de 18.635 Kilobytes. El número medio de páginas es de 225, pero las sedes de bibliotecas universitarias tienen en promedio 481 páginas [19].

Como índice del grado de interrelación de los documentos Web se suele tomar el mero recuento de enlaces que contienen. Térmens ha hallado una media de 743 enlaces en las páginas de las bibliotecas universitarias que ha analizado y 43 en el resto [19]. El análisis de Inktomi reveló que, por término medio, cada documento HTML contiene un total de 71 etiquetas. De ellas, 11 son únicas [20].

Bray estimó que el 80% de las páginas analizadas estaban enlazadas por otras en cantidad variable entre 1 y 10 [18]. También halló que el 80% de las sedes no contenían enlaces externos (!).

Las características distintivas antes mencionadas imposibilitan un tratamiento tradicional de los recursos informativos distribuidos a través de Internet. Los datos que describen cualquier recurso se han diferenciado en datos intrínsecos, derivados del examen de las páginas en cuestión, y datos extrínsecos, que sólo un observador externo que compara unos y otros objetos puede identificar. Tanto unos como otros presentan muchos problemas a un enfoque analítico tradicional.

Se ha mencionado ya que no existen valores "fijos" en los documentos, a pesar de la existencia de etiquetas que podrían albergarlos. Por ora parte, los recuentos estadísticos de términos de los documentos, a partir de los cuales se alimentan las bases de datos de muchos sistemas de recuperación, ofrecen un bajo rendimiento porque los términos no se presentan contextualizados (excepción hecha de WebCrawler y algunos desarrollos recientes de otros sistemas). La falta de estabilidad de los documentos y la acelerada dinámica del medio imposibilitan un tratamiento con la suficiente exhaustividad desde un enfoque extrínseco (por ejemplo, una clasificación temática). A pesar de todo ello, existen los sistemas y es posible distribuirlos en varios modelos. Los siguientes apartados tratan separadamente las listas y directorios de confección manual, los sistemas basados en recopilaciones automáticas y, en fin, aquellos basados en el empleo de elementos de inteligencia artificial.

Listas y directorios

El modelo más simple para la recopilación de bases de datos que describan y posibiliten el acceso a los recursos distribuidos en Internet es aquel que emplea descripciones manuales externas de los recursos, como si se estuviera recopilando una base de datos o un catálogo tradicionales.

Cuando las dimensiones de Internet y del espacio Web eran manejables, se produjeron muchos directorios impresos. Aún hoy, se siguen publicando series como las "Guías de Navegación" (Madrid, Anaya Multimedia). Muchas revistas especializadas incorporan trabajos e incluso secciones fijas que contienen listas y directorios. En la propia Web es extremadamente frecuente la existencia de páginas de enlaces (links) recopilados manualmente. Por último, Yahoo! y LookSmart representan los mejores ejemplos de este modelo, de sus limitaciones y virtudes

A este modelo se achacan muchas limitaciones [18]. En primer lugar, los sistemas sólo cubren una mínima fracción de los recursos disponibles. Yahoo!, el de mayor cobertura, sólo alcanzaba recientemente los 200.000. Las estructuras de navegación que ofrecen no constituyen sistemas controlados, extensibles y generalmente reconocibles de estructuración del conocimiento como podrían ser algunos de los sistemas clasificatorios más aceptados. La falta de coherencia y fiabilidad para indizadores y usuarios es la consecuencia. Existen deficiencias en su lógica, sus jerarquías, su desglose de categorías, la exhaustividad de la terminología, la forma en que se relacionan las diferentes clases y la capacidad de polijerarquía.

La selección de los epígrafes clasificatorios y de los elementos de descripción del recurso se deja en manos del usuario que incluye su documento en el sistema o en indizadores al servicio del mismo.

Muchos de los recursos incluidos en listas y directorios pierden utilidad pronto, ya que no existen mecanismos suficientemente ágiles para realizar un seguimiento de los cambios de dirección o contenido.

Por último, agregación y granularidad afectan en gran medida la cobertura de tales sistemas y la especificidad de sus contenidos.

Pero los índices y directorios de recopilación manual presentan grandes ventajas, tan grandes que siguen ocupando los primeros puestos en las listas de destinos más conectados y, además, han forzado al resto de los sistemas a adoptar elementos comunes con los directorios, cuando no claras alianzas y combinaciones. A causa de su limitación y de la intermediación manual en la descripción de recursos, los directorios presentan una selección implícita, que supone en definitiva una evaluación. Por otra parte, la estructuración jerarquizada y el hecho de que los contenidos de sus índices hayan sido elaborados manualmente facilita que los términos se interroguen dentro de un contexto más o menos definido. La existencia, por otra parte, de epígrafes clasificatorios que agrupan los destinos representa una ayuda adicional para los usuarios que deben expresar su necesidad de información.

Bases de datos de recopilación automática

Frente a los llamados directorios y a las listas, las bases de datos de recopilación automática proporcionan una mayor cobertura, mayores exhaustividad en la indización y nivel de representación de los documentos distribuidos en Internet, un grado muy elevado de actualización y una altísima especificidad en la indización. Como se verá a continuación, no todas estas características se han mantenido a lo largo del tiempo.

Los sistemas de recuperación basados en programas de recopilación automática hicieron su aparición en 1994. Una cronología simplificada podría ser la siguiente:

A principios de 1994, estudiantes del Department of Computer Science and Engineering de la Universidad de Washington se reunieron en un seminario informal para tratar la popularidad de Internet y la World Wide Web. Del seminario surgieron algunos proyectos y el de Brian Pinkerton fue WebCrawler. Este sistema se lanzó el 20 de Abril de 1994, con documentos procedentes de unas 6.000 sedes. En Octubre de ese año la media diaria de consultas era de 15.000 [19].

El trabajo en Lycos comenzó en Mayo de 1994, usando el programa LongLegs de John Leavitt como punto de partida. En Junio de 1994, John Mauldin añadió el programa de recuperación Pursuit para posibilitar la búsqueda de las páginas recopiladas. Pursuit estaba basado en la experiencia del Tipster Text Program de ARPA, que trataba de la recuperación en bases de datos textuales muy grandes. El 20 de Julio de 1994 se lanzó al público Lycos con una cobertura de 54.000 documentos. En Agosto había identificado 394.000 documentos [20].

El primer robot de recopilación de Open Text Index se lanzó el 14 de Febrero de 1995. El esfuerzo de recopilación tenía intenciones comerciales: el desarrollo y venta de productos como Livelink Search y Livelink Spider, dirigidos al mercado de los productos para intranets o organización de sedes Web20. OpenText se clausuró el 16 de Marzo de 1998 como servicio general. Fue sustituido por Livelink Pinstripe como servicio especializado en información económica y financiera [21].

El lanzamiento al público de MetaCrawler se realizó el 7 de Julio de 1995. Comprendía el acceso combinado a 5 servicios: Galaxy, InfoSeek, Lycos, WebCrawler y Yahoo, a los que luego añadiría OpenText. Por estas fechas procesaba más de 7000 búsquedas semanales [22].

Alta Vista comenzó su desarrollo en el verano de 1995 en los laboratorios de investigación de Digital Equipment Corporation en Palo Alto y comenzó a distribuirse oficialmente el 15 de Diciembre de 1995 [23].

Esta breve relación cronológica revela, una vez más, el origen académico de muchos de los sistemas (a veces surgidos de proyectos escolares) y la evolución posterior de estos hacia el mundo comercial, donde se han originado los de más reciente creación. Las correspondientes fuentes muestran también un énfasis especial en el tamaño de los índices recopilados. La discusión sobre el número de páginas, destinos o documentos incluidos en la cobertura de los sistemas es un tema recurrente con aportaciones siempre recientes [24].

A medida que cada base de datos ha ido creciendo, se ha hecho necesario contar con mayor poder de procesamiento para su mantenimiento. Esta necesidad ha favorecido el movimiento hacia el

patrocinio comercial o la adquisición de los servicios por empresas. A su vez, la exigencia de que los mensajes comerciales se difundan a un número cada vez mayor de potenciales consumidores ha acentuado el énfasis en la exhaustividad de la recuperación. En un periodo de tiempo muy reducido, se han venido cumpliendo una a una las previsiones de Shaw: "Los días de los ingenios de búsqueda completamente gratuitos, actualizados en cualquier departamento universitario de informática por un pequeño equipo de estudiantes de ojos enrojecidos hartos de cafeína están llegando a su fin. Es posible que persistan un puñado de buscadores de origen académico para acceso exclusivo desde el campus, pero en los próximos uno o dos años, habrá que pagar directamente (mediante suscripción o por referencia) o indirectamente (a través del aumento de precio de los productos cuyas compañías invierten en los servicios de recuperación en Internet)" [25].

Cobertura de los sistemas

A pesar de los millones de URLs barajados por los propios productores, que también esgrimen frecuencias de actualización sorprendentemente cortas, se ha demostrado que la cobertura de los sistemas no es completa y que sus ciclos de actualización se alargan indeseablemente.

El mecanismo básico de recopilación de los llamados "buscadores" pasa por introducir sus programas de recopilación en los ordenadores conectados a Internet y transmitir el contenido de los archivos (las páginas) allí albergadas a uno o más ordenadores centrales. En estos, un programa complementario indiza el texto contenido en los archivos y elabora una base de datos que permite el acceso a los registros y, desde ellos, la conexión con los recursos distribuidos. Pero la descarga de las páginas no es casi nunca completa y que los cambios de las páginas originales no se corresponden con conexiones y actualizaciones de las bases de datos de cada sistema [26]. Por otra parte, se han formulado acertadas críticas a la recopilación automática. Así, Brake anota que en los inicios del Web, las páginas contenían simples caracteres alfanuméricos y el acceso era totalmente indiscriminado. En la actualidad, se están produciendo algunos cambios:

Los formatos de los documentos no siempre son legibles por un programa de visualización ni almacenables como simple texto: los ficheros PDF y, sobre todo, PostScript no son fácilmente traducibles.

Ciertos conjuntos de documentos, en los ámbitos científicos, contienen complejos diagramas y fórmulas, tampoco traducibles

Algunas sedes no admiten examen por robots, porque son de pago (The New York Times) o porque distorsionarían sus cifras de audiencia (CNN)

Algunas sedes sólo revelan su contenido en respuesta a peticiones específicas de los usuarios. Es típicamente el caso de los catálogos bibliográficos y otras bases de datos [27].

El mecanismo de indización y la función de relevancia

El algoritmo empleado por la mayoría de los sistemas para la indización de los textos rastreados y para el cálculo de relevancia en respuesta a la búsqueda es conocido por "localización-frecuencia" [28]. En líneas generales, a la presencia de un término en un documento Web (o e un artículo USENET, o en el cuerpo de un mensaje de correo electrónico) se asigna un valor relacionado directamente con su frecuencia en el texto del documento en cuestión e inversamente relacionado con su frecuencia total en el índice global de la base de datos. Existe un factor de corrección que depende de la posición que el término ocupa. Este factor es más favorable si el término aparece en el título o la cabecera del documento o si su posición es próxima al inicio de la página. Cada sistema interpreta de forma particular esta expresión general y también presenta un grado diferente de exhaustividad en la indización.

Así, Lycos recupera documentos que se han indizado por el título, el subtítulo, los encabezamientos y subencabezamientos y los enlaces. Más las 100 palabras de mayor peso (determinado mediante la

función Tf*IDf) más las 20 primeras líneas. Además, "emplea un esquema de reducción de datos (representación de los documentos) para reducir la información almacenada de cada documento [29]. Infoseek ordena los resultados de búsqueda en función de su ajuste a la petición formulada y los presenta en orden inverso por su "nivel de confianza" [30]. AltaVista presenta los documentos en respuesta a una petición situando los más relevantes en la cabecera de la lista. La ordenación se basa en la inclusión de todos los términos del perfil en los documentos hallados y en una combinación de otros criterios [31].

La base de datos de WebCrawler tiene 2 componentes: un índice a texto completo y una representación del Web en forma de grafo. El índice a texto completo se basa en la actualidad en el IndexingKit de NEXSTEP. Emplea un modelo de espacio vectorial para afrontar las peticiones. Para preparar un documento para su indización, un analizador lexicográfico lo segmenta en una lista de palabras que incluye tokens del título y el cuerpo del documento. Las palabras se filtran a través de una stop list (lista de palabras vacías) y son ponderadas. Las palabras con mayor numerador y menor denominador se ponderan más. Aquellas con bajas frecuencias, menos. Este tipo de ponderación recibe el nombre de ponderación de particularidad (particularity weighting) [32].

HotBot también emplea la frecuencia de los términos en el documento, la extensión y la frecuencia en la base de datos. Las combina con la presencia de los términos de búsqueda en los títulos de los documentos y en la lista de palabras clave (etiqueta META) proporcionadas por los creadores de las páginas [33]

La experiencia de los usuarios finales y también la de los documentalistas o recuperadores profesionales no parecen muy favorables. Los medios de información general no se han quedado atrás en sus acusaciones [34]. La encuesta de Pollock y Hockley, a pesar de su reducida muestra, es ilustrativa de la insatisfacción de usuarios legos en Internet y sus posibilidades [35]. En otra encuesta, la segunda de las organizadas entre usuarios españoles por la Asociación de Investigación de Medios de Comunicación, el directorio Yahoo sigue apareciendo como el principal destino, mientras el primer servicio de búsqueda, AltaVista, sólo aparece en el quinto lugar por número de conexiones [36]. Otros trabajos ofrecen resultados totalmente comparables [37]

Sistemas distribuidos con indización asistida
Desde el punto de vista arquitectónico, directorios y bases de datos de recopilación automática plantean el problema de la multiplicidad de accesos simultáneos a servidores centralizados. Desde el punto de vista del análisis y la recuperación del contenido, representan los extremos opuestos de un continuum que separa la representación indiscriminada de contenidos con falta de atención a las estructuras de los documentos y, por otra parte, los sistemas que estructuran en registros sui generis el contenido de los recursos distribuidos. Al énfasis en la exhaustividad, la cobertura y la rapidez de actualización de los sistemas de recopilación automática se contraponen la selectividad, la búsqueda de precisión y la "burocratización" de los directorios y esquemas clasificatorios. La entrega a procedimientos automáticos frente al conservadurismo del análisis humano.

La búsqueda del equilibrio entre ambos extremos parece estar representada por un tercer modelo que combina, por un lado, arquitecturas distribuidas y, por otra parte, aprovecha descripciones normalizadas de los recursos producidas no por el propio sistema, sino por los mismos distribuidores de los documentos.

Los sistemas archie y VERONICA, asociados inicialmente a los protocolos FTP y Gopher, y Harvest y ALIWEB (Archie Like Indexing of the WEB), basados en los primeros y en el sistema WAIS, fueron sólo el preludio de la tendencia creciente hacia el control de contenidos de los recursos distribuidos a través de Internet, actualmente potenciada gracias a la Iniciativa sobre Metadatos.

Las descripciones de Harvest [38] y ALIWEB [39]parten criticando la innecesaria sobrecarga en servidores y conexiones producida por los sistemas de recopilación automática. También el excesivo tráfico originado por la propia popularidad de estos sistemas, así como su dificultad de tratar con formatos informativos heterogéneos. Su alternativa para la recuperación de recursos se basa en la indización en los servidores de origen, la existencia de descripciones normalizadas de cada recurso, una arquitectura distribuida de las bases de datos recopiladas y el empleo de programas cliente de

consulta y recuperación. Estas características están directamente inspiradas por el concepto de WAIS (Wide Area Information Server), desarrollado a finales de los años 80 por Brewster Kahle.

WAIS era un sistema en que múltiples bases de datos especializadas se distribuían en servidores dispersos controlados por un directorio, y cuyos contenidos eran accesibles y recuperables mediante el empleo de programas cliente. Los usuarios obtenían una lista de las bases de datos y, en respuesta a una expresión de búsqueda dirigida a una base de datos seleccionada, se accedía a los servidores que la contenían. Como resultado, se obtenía una descripción de los textos y la posibilidad de obtener completos los documentos.

Aunque el propio Kahle se refería a WAIS como una "herramienta de Internet para la búsqueda de información" [40], declaraba que consideraba Internet como un medio de distribución de información y se refería a los organismos públicos, los editores, las bibliotecas y las empresas dispersas como sus principales mercados. Por otra parte, ninguna de las características que en su opinión diferenciaban WAIS de los sistemas de recuperación tradicionales se han mantenido: la interfaz amigable que acepta expresiones de búsqueda en lenguaje natural, las posibilidades de búsqueda booleana y limitación por campos destinadas al usuario avanzado y el feedback de relevancia se han ido incorporando en mayor o menor medida a los sistemas de recopilación automática y aún a los directorios.

Archie, que siguió cronológicamente al sistema WAIS a principios de 1992, permitía la recuperación por palabras clave de ficheros informáticos, de los cuales controlaba 2 millones en 1994. En Noviembre de 1992, se anunció VERONICA, desarrollado en la Universidad de Reno, y que en respuesta a una búsqueda sobre menús Gopher, ofrecía otro menú de resultados. En Enero de 1995, VERONICA indizaba 5.057 servidores y dos meses antes su índice incluía unos 15 millones de items [41]. En ambos casos existía un esquema para la descripción de los recursos, aunque a veces se tratara de una simple línea [42]

Formatos normalizados y metadatos

El entronque de WAIS con Harvest, ALIWEB y otros sistemas, se basa en el hecho de emplear una representación estructurada de los documentos y de las transacciones. En el caso de WAIS, el esquema correspondía a extensiones de la norma Z39.50 (Information Retrieval Service Definition and Protocol Specification for Library Applications) de NISO [43]. Por su parte, Harvest, distribuido a partir del 9 de Noviembre de 1994 y cancelado el 20 de Agosto de 1998, empleaba el Summary Object Interchange Format (SOIF) [44]. El objeto de este formato, a modo de tabla atributo-valor, era tanto unificar la representación de la información recopilada como comprimirla de forma que su transmisión resultara más eficiente. Otro tanto hacía ALIWEB con su empleo del IAFA (Internet Anonymous FTP Archives), definido por el correspondiente grupo de trabajo de la Internet Engineering Task Force. Por su parte, el Proyecto de Biblioteca Digital de la Stanford University emplea también SOIF en su propuesta sobre recuperación. Más concretamente, para corregir los desajustes existentes en los sistemas de búsqueda múltiple (metabuscadores) con el cálculo de relevancia que combina los diferentes algoritmos de ordenación de los sistemas individuales [45].

El esquema de funcionamiento de todos los sistemas de este grupo pasaba por el consenso entre los servicios de recuperación y los proveedores de información acerca del modo de representar el contenido de los documentos. Y es precisamente la existencia de este consenso la que permite delimitar una trayectoría coherente que enlaza WAIS, un servicio no estrictamente originado en Internet, con los iniciales archie y VERONICA y los sistemas basados en HTTP como Harvest y ALIWEB. Este enfoque ha sido ampliamente revisado por Rachel Heery [46]. Se debe tener en cuenta, sin embargo, que todos estos sistemas exigían el acuerdo previo entre proveedores y servicios de recuperación. Algo impensable si lo que se pretende es un control de contenidos de todos los documentos distribuidos a través de Internet. De hecho, las exigencias de mantenimiento y actualización de los sistemas hasta ahora mencionados han provocado su progresivo estancamiento o su desaparición.

La continuación de esta línea está representada por el empleo de metadatos para la descripción coherente de los contenidos de los documentos distribuidos. Pero atribuídos por los creadores de los propios documentos.

"Dado que Internet contiene más información de la que catalogadores, indizadores y elaboradores de resúmenes pueden gestionar con los medios y sistemas disponibles, se llegó al acuerdo de que una alternativa razonable para obtener metadatos de los recursos electrónicos sería proporcionar a los autores y proveedores de información medios para que describieran los recursos por sí mismos" [47].

Esta afirmación de Stuart Weibel figura en el documento resultante del primer encuentro sobre el control de contenidos de recursos distribuidos en Internet mediante metadatos, celebrado en Dublin (Ohio) del 3 al 5 de Marzo de 1995. Desde entonces, la Dublin Core Metadata Inititative se ha ido configurando como la plataforma y una de las tendencias más prometedoras para una recuperación eficiente de información distribuida en Internet.

En el mismo documento, Weibel emplea la metáfora del continuum y propone

"como solución alternativa que equilibre ambos extremos la creación de registros que resulten más informativos que los índices pero sean menos completos que un registro catalográfico formal. Si sólo se requiere un pequeño esfuerzo para crear tales registros, se podrían describir más objetos, especialmente si los autores de los recursos se animan a crear la descripción. Si, por otra parte, dicha descripción sigue una norma establecida, sólo la creación del registro requeriría intervención humana, mientras que el descubrimiento y la recopilación de recursos se podrían efectuar mediante herramientas automatizadas".

Aunque proliferan las definiciones del concepto, cabe describir los metadatos como valores que se presentan asociados a su carga semántica, expresada por la unión entre un elemento estructural (autor, título, fecha...) y las correspondientes variables. El conjunto inicial se limitaba a identificar el significado de un grupo de elementos descriptivos con objeto de mejorar la detección de recursos en el espacio Web que se pudieran considerar objetos similares a documentos (Document-Like-Objects). El resultado del segundo seminario fue la adopción del Esquema de Warwick (Warwick Framework), un modelo conceptual de una arquitectura de contenido para paquetes de metadatos de diversos tipos. En el tercer seminario se extendió el esquema para la descripción de imágenes y, poco después, el conjunto inicial de 13 elementos se extendió a 15, de los que existe traducción española [48].

Internet no cuenta, como ya se ha visto, con un universo de proveedores de información cualificados y expertos en campos temáticos concretos. Así que, si sólo se produjera una normalización de elementos y ciertas instrucciones sintácticas, la situación sería tan caótica como la que se derivaría de sistemas de indización tradicionales que se basaran únicamente en las palabras clave asignadas por autores noveles o remedaría la problemática de los sistemas de recopilación automática. Afortunadamente, el seminario celebrado en Camberra (3 a 5 de Marzo de 1997) vino a incorporar al Dublin Core el calificador esquema y el subelemento tipo [49]. Mediante su aplicación es posible despejar ciertas ambigüedades que se podrían plantear a los autores de los documentos y que, de hecho, son habituales en la catalogación e indización por profesionales.

De hecho, los sucesivos desarrollos del esquema inicial suponen un grado de normalización progresivamente mayor. Así, a la lista consensuada de elementos de descripción viene a añadirse la normalización de valores, que aprovechan esquemas preestablecidos. Además, la ambigüedad queda también reducida por el empleo de subelementos tipo.

Los requisitos del modelo de indización asistida
Elénfasis en la necesidad de consenso, no es casual. Para garantizar un funcionamiento eficiente de este modelo es necesario, en primer lugar, el acuerdo de los grandes proveedores de información, por ejemplo los grandes editores del sector electrónico. En segundo lugar, que se produzca de forma generalizada la utilización del conjunto de datos, sus calificadores y su sintaxis. En tercer lugar, se requiere que las sucesivas versiones de los lenguajes de marcas acojan el esquema vigente y sus

desarrollos. Por último, los programas de recopilación automática y los sistemas de descripción de los directorios y esquemas deben de reconocer y aceptar el valor de los metadatos.

El primer requisito no sólo se cumple, sino que los grandes proveedores se cuentan entre los primeros impulsores de la idea de utilizar un formato normalizado como mecanismo simple para representar los metadatos: así, el comunicado del 8 de Septiembre de 1997 que anunciaba el apoyo de NetScape a la iniciativa Resource Description Framework, mencionaba entre los impulsores a nombres de peso en el sector de la edición electrónica. Entre ellos, CNN, CBS, Time Inc o Knight-Ridder eran sólo algunos [50]. Por otra parte, el 3 de Octubre siguiente se produjo el anuncio del primer borrador público de esta iniciativa, con el apoyo de nombres no menos relevantes [51], que fue presentado oficialmente en la quinta reunión de la Dublin Core Initiative una semana más tarde.

La adopción generalizada del esquema por parte de los productores iniciales de información depende de que los redactores de páginas en lenguaje de marcas o los usuarios de programas para su confección cuenten con facilidades para la inclusión de metainformación en ellas. Por otra parte, no se puede descartar el empleo de procedimientos que, tras el análisis automático del código de cada documento, puedan generar una lista normalizada de valores [52]. Ya existen, por añadidura, proyectos y sistemas automatizados que transforman los diversos formatos normalizados de descripción y los valores de catalogación al esquema de metadatos [53], [54]. Cabe destacar en esta línea, por su elegancia y simplicidad, el procedimiento propuesto por Massimo Marchiori, basado en la "propagación" de los valores de los metadatos de unos a otros documentos en función del grado de interconexión entre ellos, que sirve de base para un cálculo borroso (fuzzy) de su similitud de contenido [55].

Las reservas sobre el empleo de cualificadores y del subelemento tipo manifestadas en la reciente reunión del grupo de indización de RedIRIS [56] ilustran la necesidad de ajuste entre los lenguajes de marca y la expresión de los metadatos. HTML, XML o cualquier subconjunto del SGML deben posibilitar la inclusión de los conjuntos de metadatos en los documentos y su reconocimiento. Afortunadamente, la simbiosis entre RDF y XML permiten despejar dudas sobre la capacidad para albergar datos normalizados según el Dublin Core en la redacción de nuevas páginas. Persiste, sin embargo, la duda sobre los millones de documentos ya distribuidos en el espacio Web y su redacción.

Los sistemas de recopilación automática más popularizados (Alta Vista, Excite, HotBot, Infoseek, Lycos, WebCrawler y Northern Light) aceptan etiquetas META en la indización de páginas Web. Sin embargo, Lycos y Northern Light no los emplean en la representación de los resultados de búsqueda y sólo HotBot e Infoseek los emplean en el cálculo de relevancia [57]. De entre los sistemas españoles, Olé menciona la posibilidad de buscar entre las palabras clave pero no hace referencia explícita, al contrario que Trovator, al empleo de metadatos en la recopilación, la indización o el cálculo de relevancia.

No obstante todo lo anterior, trabajos recientes ponen el dedo en la llega del coste de la incorporación de metadatos a las páginas generadas en entornos académicos, comerciales u oficiales. Sus razonamientos se basan en tres hechos: la necesidad de intervención humana (y el consiguiente gasto) en la elaboración de las etiquetas, el esfuerzo añadido que supone la incorporación de metadatos a la generación del contenido que, en definitiva, suponen las páginas y el hecho de que, en el sector comercial, el retorno de la inversión realizada, en forma de número de accesos tras la localización de páginas, no parece justificar la inversión requerida [58]. No resulta extraño, por ello, que se empleen los metadatos más descriptivos para caracterizar las sedes de los documentos, en lugar de emplearlos para evidenciar el contenido de cada página y que la mayor parte se genere de forma automática a partir de los programas de edición HTML [59]

El concepto de delegación y el empleo de agentes
Inicialmente, el descubrimiento y recuperación de recursos distribuídos en Internet quedó a la exclusiva iniciativa de los usuarios. En una segunda fase, las listas, índices, directorios y bases de datos de recursos han representado soluciones aportadas desde el extremo de los proveedores. Este esquema en dos capas resulta problemático: los usuarios se han visto incapaces de localizar

recursos por sí mismos, los sistemas se han visto desbordados en su misión de organizarlos para proporcionar un acceso efectivo y, además, unos y otros se han venido comportando como extraños: la práctica totalidad de los sistemas han desconocido el estado de conocimiento de los usuarios quienes, a su vez, sólo de forma aproximada han alcanzado a comprender las condiciones de operación de los diversos servicios. Los conceptos de mediación y delegación pueden proporcionar un marco adecuado para la mejora de la recuperación de información distribuída.

Deficiencias de los actuales sistemas

Los servicios de recuperación muestran claras deficiencias que serán más aparentes en el futuro. Entre otras, se han anotado las siguientes [60], [61]:

El hecho de que se realice la recuperación basada en uno o más términos de búsqueda a expensas del usuario presupone un conocimiento del vocabulario y los sistemas que, con frecuencia, sólo conduce a la existencia de ruído.

La confección de índices se realiza mediante la recopilación y el transporte de documentos. Este método provoca congestión en las conexiones y no es eficiente porque no existe cooperación entre los diversos servicios.

La cobertura se limita a algunos espacios informativos. Otros, como las bases de datos tradicionales, escapan a la recopilación y, por tanto, a la recuperación.

Los sistemas no siempre son accesibles.

La indización se produce de forma indiscriminada, como una simple recopilación de términos que se ordenan como entradas individuales en los índices sin atender al contexto del documento del que provienen.

Los sistemas de recopilación automática no pueden seguir con el ritmo adecuado la dinámica y falta de estabilidad de los documentos.

Los sistemas actuales no posibilitan el intercambio de "experiencia" entre los usuarios con intereses afines ni el ajuste entre diversos episodios de recuperación de un mismo usuario y los cambios en el estado de conocimiento del mismo.[62]

El concepto de delegación

Bjorn Hermans ha definido el concepto de "Agency", que cabe traducir por "Delegación", como "el conjunto de medios (técnicas, conceptos, aplicaciones y otros) para personalizar, elaborar, delegar y catalizar procesos en el entorno online" [63].

Este esquema, que interpone una mediación a los extremos representados por los productores y distribuidores de información en un lado y a los usuarios demandantes de información, en el otro, es perfectamente traducible al modelo en 3 capas popularizado en muchos trabajos sobre delegación y agentes, y avanzado hace tiempo en el marco del diseño de sistemas de información [64]. El propio Wiederhold enumera las funciones que la capa mediante debe realizar:

Localización y recuperación de datos relevantes procedentes de múltiples fuentes heterogéneas

Condensación y transformación de los datos recuperados hasta representarlos mediante formatos y semántica comunes

Integración de los datos homogeneizados en función de las claves de selección

Reducción de los datos integrados por abstracción para aumentar la densidad informativa en el resultado a transmitir [65].

Haverkamp y Gauch [66], Jansen [67] y muchos otros autores ofrecen una panorámica de los sistemas de agentes múltiples, su organización y sus características operativas. Además, proporcionan algunos ejemplos de sistemas en operación o en experimentación (algo anticuados en el caso de las autoras estadounidenses, cuyo original data de Noviembre de 1996). En el contexto del presente trabajo interesa, más que abundar en inventarios ya existentes, situar la variada gama de asistentes, robots, agentes y otros dispositivos mediadores en el entorno que representa el modelo en 3 capas.

Mecanismos de delegación en la producción y provisión de información
El conjunto de operaciones englobadas en la localización y selección de recursos (selección, filtrado y procesamiento previo de los documentos) ha sido de los primeros en beneficiarse de la aplicación de técnicas y dispositivos de delegación. Los propios robots o programas que emplean los sistemas de recopilación e indización automáticas constituyen el ejemplo más evidente. Además, buena parte de los procedimientos de filtrado asociados a la tecnología "push" y a los canales de distribución, así como los mecanismos de organización y filtrado de mensajes de correo electrónico representan mecanismos delegados.

Más avanzados resultan los sistemas de procesamiento previo que permiten la asociación de documentos en función de su contenido. La tecnología de agentes para la localización de recursos uniformes (URA) [68] y productos como ReferralWeb, que emplea procedimientos similares a los de la indización por citas para establecer asociaciones entre recursos y presentarlos gráficamente agrupados, son sólo algunos ejemplos. Lo que resulta distintivo de esta segunda gama de productos es su elaboración de un modelo de recursos. En este sentido, los sistemas para la visualización de los resultados de búsqueda desarrollados por el Xerox PARC y otros centros, que combinan el análisis de textos con la presentación gráfica de los recursos [69] son só lo el preludio de aquellos que se apoyan en la interactividad y la respuesta de los usuarios ante los resultados de búsqueda, retroalimentando sus sistemas mediante recuentos de conexiones (DirectHits), el procesamiento de los contenidos (AskJeeves, aunque la asistencia en este caso tenga participación humana [70]) o, de forma más habitual, permitiendo el refinamiento de los resultados a través del procesamiento estadístico de los términos de los documentos hallados e, incluso, mediante la búsqueda mediante patrones de documentos (query by example). Acaso las soluciones más avanzadas procedan de Google y Clever. En el caso de Google, no sólo existe un recuento de "popularidad" de los recursos recuperados, sino un procesamiento previo en función de los enlaces que los documentos contienen. En Clever, aún no operativo, el recuento de enlaces y la definición de recursos centrales (authorities) o de recursos concentradores (hubs), es resultado de un procesamiento más sofisticado [71]

Operaciones delegadas por los usuarios
Al igual que en el extremo de los proveedores, es posible ordenar la gama de dispositivos de delegación puestos al servicio de los usuarios demandantes de información desde los meros asistentes hasta los sistemas basados en conocimiento. El asistente de búsqueda múltiple Sherlock, incorporado a una de las últimas versiones del Mac OS, ha venido a añadirse a una plétora de programas que traducen expresiones en lenguaje natural, envían los perfiles resultantes a varios servicios y compactan y ordenan los resultados [72]. Algo más avanzadas son las funciones de Alexa, que caracteriza una sede a través del número de accesos que recibe y, además, asocia unas y otras en función de sus enlaces y de los destinos comunes de los usuarios [73]. La integración de estas capacidades con la versión 5 del programa de navegación Internet Explorer se ha producido recientemente [74].

La noción de inteligencia
El hecho de que los programas y dispositivos desarrollados por los sistemas de recuperación tengan en cuenta algunas acciones y respuestas de los usuarios y el que algunos asistentes "de sobremesa" puedan apreciar cambios en el entorno de los recursos, caracteriza a unas y oras aplicaciones como dispositivos que intermedian entre ambos extremos en el proceso de acceso a la información distribuída en Internet. Sin embargo, otros elementos resultan más claramente centrados en el esquema descrito.

Autonomía, fiabilidad, capacidad de iniciativa, reactividad y habilidad social son algunas de las muchas propiedades que generalmente se atribuyen a los agentes. Pedro Hípola y Benjamín Vargas proponen la definición siguiente:

Un agente inteligente se define como una entidad de software que, basándose en su propio conocimiento, realiza un conjunto de operaciones destinadas a satisfacer las necesidades de un usuario u otro programa, bien por iniciativa propia o porque alguno de ellos lo requiere [75]

No es posible la inteligencia sin una base de conocimiento cambiante y no es posible esa base sin capacidad de extracción de datos y elaboración de modelos que caractericen entornos. Gracias a la habilidad social, los agentes se comunican con otros agentes y con personas y, a través de su reactividad, son capaces de captar cambios en un entorno determinado, cambios ante los que reaccionan sin necesidad de instrucciones coyunturales.

Algunos agentes se sitúan junto a los datos o los recursos que observan, otros generan interfaces para facilitar la comunicación con los usuarios, otros conectan los primeros con los segundos. Se han descrito agentes móviles capaces de transmitirse a través de redes y examinar, procesar y comunicar descripciones de recursos. Todos ellos, sin embargo, están dotados de modelos que les permiten ajustar oferta y demanda de información de forma dinámica, mediante un continuo aprendizaje.

Remembrance Agent [76] y SiteHelper [77] ejemplifican a la perfección estos requisitos. El primero realiza continuas asociaciones entre el entorno de los documentos y el estado de conocimiento del usuario mediante una observación continua de sus acciones. SiteHelper propone un modelo del entorno informativo de cada sede Web y también monitoriza a los usarios para distribuir selectivamente nuevos documentos e información en respuesta a cambios en cualquiera de ambos modelos.

Una agenda para un panorama próximo
Del observatorio de buscadores repetidamente mencionado se pueden extraer noticias significativas. En los últimos 18 meses se han producido estos hechos:

AltaVista y HotBot han añadido el directorio LookSmart a sus servicios. NorthernLight ha comenzado a emplear editores humanos para la clasificación de recursos, al igual que HotBot. Lycos ha decidido convertirse en un directorio y emplea un equipo de 10.000 voluntarios para la clasificación humana de los recursos. Ha aparecido UK Max, un servicio regional para el Reino Unido basado en Inktomi. Esta empresa acaba de ampliarlo al resto de Europa. El Gobierno estadounidense ha instaurado su propio servicio de recuperación basado en Northern Light. LookSmart ha lanzado directorios para las 65 mayores áreas urbanas de los Estados Unidos y, mucho antes, Yahoo! y otros habían iniciado el lanzamiento de servicios nacionales. AltaVista se ha asociado con AskJeeves para ofrecer servicios de pregunta respuesta. HotBot emplea Direct Hits para mejorar sus resultados, mientras éste último servicio ha iniciado la personalización de los resultados de búsqueda a través de filtros con base demográfica. AltaVista acaba de introducir el cobro por posiciones en sus resultados de búsqueda.

Es relativamente simple apreciar en ellos y otros una serie de tendencias:

Los servicios se personalizan, tanto en su oferta como en sus procesos de mantenimiento y respuesta a las demandas

Existe un proceso de regionalización, de especialización de contenidos y de aparición de servicios especializados ("niche search engines")

La intervención humana en la descripción de recursos es cada vez mayor y la frontera entre los sistemas de recopilación automática y los servicios de directorio se difumina

La comercialización de los servicios y la lucha por las estadísticas de conexiones continúan con igual ímpetu.

Se asiste a una unificación o, al menos, aproximación de los espacios informativos. Especialmente notable en el caso de Northern Light y su "Special Colletion", aunque no sea el único

La lectura global de estos datos resume ingentes esfuerzos por ajustar los niveles de servicios y las cuotas de mercado a un entorno que evoluciona a pasos agigantados. Sin embargo, la última de las tendencias permite vislumbrar un panorama próximo, que bien se puede ejemplificar empleando la experiencia de las bibliotecas.

En la actualidad, cualquier biblioteca de un sistema que se precie se plantea ofrecer a sus usuarios acceso a recursos distribuídos en Internet previamente seleccionados. Por otra parte, los catálogos de cualquier biblioteca de un sistema que se precie son accesibles a través de Internet. El tercer actor en este panorama está representado por el sector de la edición científica y técnica, que sólo tímidamente comienza a valorar las posibilidades de distribución y negocio que la Red ofrece.

Este entorno, que algunos han dado en llamar "electronic marketplace", puede estar dominado por las actividades de evaluación de recursos y servicios previas al acceso a esos mismos recursos mediante alguna forma de pago. Evaluación y coste implican cuanto menos las siguientes necesidades:

Definición y estructuración completa de los documentos a través de lenguajes que sobrepasen el mero nivel descriptivo de las etiquetas y marcas de visualización.

Representación de los documentos y sus relaciones a través de los adecuados modelos de datos y de relaciones entre unos y otros recursos. No mediante una simple recopilación y organización de términos.

Definición de modelos de usuarios y elaboración de bases de conocimiento dinámicas, que se ajusten mediante procesos de aprendizaje al perfil informativo de cada usuario y a sus cambios.

Desarrollo de infraestructuras de comunicaciones eficientes en el marco de una "cultura de red".

Son varios los grupos llamados a colaborar en las diferentes tareas de esta agenda: los especialistas en comunicaciones y en proceso de datos, los investigadores en inteligencia artificial, los documentalistas y los restantes especialistas en información pueden ser algunos de ellos. Sus iniciativas y posibles soluciones dependen de un cuarto agente: los productores de información que, desde editoriales comerciales o desde sedes académicas impongan una sintaxis en la elaboración de documentos que permita su recuperación eficiente a pesar de los costes de acceso. El reciente llamamiento de los organizadores del Congreso Semestral de la American Society for Information Science resulta muy revelador en este sentido[78].

Referencias

1. López Alonso, MA y Mares Marín, J: El futuro de la identificación de la información en Internet. Quintas Jornadas Españolas de Documentación Automatizada. 1996, 17 a 19 de Octubre, Cáceres. vol 1 pags 513-518.

2. Sánchez Montero,JA: Hacia una optimización de los recursos de Internet en la empresa. Revista Española de Documentación Científica, 20(1): 52-60, 1997.

3. Baró i Queralt, Jaume: Cerca i recuperació d'informació al World Wide Web: una aproximació a les eines disponibles. Sisenes Jornades Catalanes de Documentació. 1997, 23-25 d'Octubre. Barcelona, pags 469-479

4. Senso, JA: Herramientas para realizar búsquedas en Internet: una revisión. El Profesional de la Información, 7 (1-2): 24-25, 1998

5. Marcos Mora, MC: Motores de recuperación de información: un análisis comparativo (parte 1). El Profesional de la Información, 7 (1-2): 18-22, 1998.

6. Maldonado Martínez, A; Fernández Sánchez, E: Evaluación de los principales "buscadores" desde un punto de vista documental: recogida, análisis y recuperación de recursos de información. Sextas Jornadas Españolas de Documentación Automatizada. Valencia, 29-31 de Octubre de 1998. Vol 2 Pags 529-551.

7. Lidsky, D; Sirapyan, N; Dawes, TA; Friedland, NE; Macicack, S; Miller, MJ et al: Your Complete Guide to Searching the Net. PC Magazine Online, December 2, 1997 http://www.zdnet.com/pcmag/features/websearch/open.htm

8. Díez Ferreira, MA: Buscar más allá de la Web. iWorld (10): 38-49, Diciembre, 1997. http://www.idg.es/iworld/199711/articulos/20buscadores.html

9. Lilley, DB; Trice, RW: A History of Information Science 1945-1985. San Diego, Academic Press, 1985.

10. Lynch, CA: Networked Information Resource Discovery: An Overview of Current Issues. IEEE Journal on Selected Areas in Communications, 13(8), October1995. http://portal.research.bell-labs.com/jsac/prot/jsac13.8/lycnh/lycnh.html. Documento de acceso restringido

11. Mauldin, ML: Measuring the Web with Lycos. Third International WWW Conference, April 11, 1995. http://fuzine.vperson.com/mlm/lycos-websize.html

12. Bray, T: Measuring the Web. Fifth International World Wide Web Conference. 6 de Mayo, 1996

13. Térmens i Graells, M: Les Webs de les biblioteques de Catalunya: Estructura interna i enllaços. Sisenes Jornades Catalanes de Documentació. Barcelona, 23-25 Oct. 1997

14. Woodruff, A; Aoki, PM; Brewer, E; Gauthier, P; Rowe, LA: An investigation of documents from the World Wide Web. Fifth International World Wide Web Conference. 6 de Mayo, 1996. http://www5conf.inria.fr/fich_html/papers/P7/Overview.html

15. Chankhunthod, A; Danzig, PB; Neerdaels, C; Schwartz, MF; Worrel, KJ: A Hierarchical Internet Object Cache. http://excalibur.usc.edu/cache-html/cache.html. 5 de Septiembre, 1997

16. Brake, D: Lost in Cyberspace. New Scientist, 28 de Junio, 1997 http://www.newscientist.com/keysites/networld/lost.html

17. Peterson, RE: Eight Internet Search Services Compared. First Monday, 2(2) 1 de Enero, 1997 http://www.firstmonday.dk/issues/issue2_2/peterson/index.html

18. Koch, T; Ardö, A; Brümmer, A; Lundberg, S: The building and maintenance of robot based internet search services: A review of current indexing and data collection methods. http://www.ub2.lu.se/desire/radar/reports/D3.11/

19. A brief Story of WebCrawler http://voyeur.mckinley.com/WebCrawler/Help/AboutWC/WCStory.html. 7 de Septiembre de 1997

20. Mauldin, ML: Lycos: Design Choices of an Internet Search Service. IEEE Expert Online. http://www.computer.org/pubs/expert/1997/trends/x1008/mauldin.htm. 7 de Julio de 1997.

21. Sullivan, D: Open Text. Seacrh Engine Report [17]. http://searchenginewatch.com/sereport/ 31 de Marzo de 1998.

22. Selberg, E; Etzioni, O: Multi-Service Search and Comparison Using the MetaCrawler. http://www.w3.org/pub/Conferences/WWW4/Papers/169/. 9 de Octubre de 1995

23. Chu, H ; Rosenthal, M: Search Engines for the World Wide Web: A comparative Study and Evaluation Methodology. http://www.asis.org/annual-96/ElectronicProceedings/chu.html. 24 de Octubre de 1996

24. Notess, GR: Measuring the Size of Internet Databases. Database, October, 1997. http://www.onlineinc.com/database/OctDB97/net10.html

25. Shaw, R: Crawlers, Spider s and Worms. Web Week, 1(2), 1 de Julio de 1995

26. Sullivan, D: Search Engine EKG. http://searchenginewatch.com/reports/ekgs/index.html

27. Brake, D: Lost in Cyberspace. New Scientist. http://www.newscientist.com/keysites/networld/lost.html 28 de Junio de 1997.

28. Sullivan D: How Search Engines Rank Web Pages. Search Engine Watch. http://searchenginewatch.com/rank.htm 30 de Junio de 1997

29. Mauldin, ML; Leavitt, JRR: Web Agent Related Research at the Center for Machine Translation. SIGNIDR Meeting. http://fuzine.mt.cs.cmu.edu/mlm/signidr94.htm 15 de Julio, 1994

30. Infoseek Corporation: Document Retrieval Over Networks Wherein Ranking and Relevance Scores are Computed at the Client for Multiple Database Documents. http://software.infoseek.com/patents/dist_search/patents.html. 8 de Septiembre de 1997

31. Jansen, J: Using an Intelligent Agent to enhace Search Engine Perfomance. First Monday 2(3). http://www.firstmonday.dk/issues/issue2_3/jansen/index.html. 3 de Marzo de 1997.

32. Pinkerton, B: Finding What People Want: Experiences with the WebCrawler. http://info.webcrawler.com/bp/www94.html,1994

33. Hock, RE: Sizing Up HotBot. Evaluating one Web Search Engine's Capabilities. Online, November, 1997. http://www.onlineinc.com/onlinemag/NovOL97/hock11.html 28 de Mayo de 1998.

34. Herren, R: Cómo nos engañan los buscadores: Las trampas que hacen los motores de búsqueda y porqué ofrecen cada vez peores servicios. Tiempo, (855): 98, 21 de Septiembre de 1998.

35. Pollock, A; Hockley, A: What's Wrong with Internet Searching. D-Lib Magazine, March, 1997. http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/march97/bt/03pollock.html 17 de Septiembre de 1997.

36. AIMC: Macroencuesta a usuarios de Internet. http://www.aimc.es/aimc/html/inter/02resul0.html. 20 de Septiembre de 1998.

37. López Alonso, MA ; Mares Marín, J: La organización del conocimiento contenido en la información hipertextual de Internet. Sextas Jornadas Españolas de Documentación Automatizada, Valencia, 29-31 de Octubre de 1998. Vol 2 Pags 489-493.

38. Technical Discussion of the Harvest System. http://harvest.transarc.com/public/trg/Harvest/technical.html 9 de Septiembre de 1998

39. History of ALIWEB. http://aliweb.emnet.co.uk/ 9 de Septiembre de 1998

40. Ubois, J: Wide Area View: An interview with WAIS Inc CEO Brewster Kahle. Internet World, 6(9), September, 1995. http://www.internetworld.com/print/monthly/1995/09/feat74.htm 19 de Octubre de 1998.

41. McMurdo, George: How Internet was indexed. Journal of Information Science Electric Writting, 21(6). http://www.qmced.ac.uk/cis/staff/cimmu/jisew/ewv21n6/default.htm 3 de Julio de 1997.

42. Mañas. JA: Búsqueda y recuperación de información en Internet. Novática (110): 75-81, Julio Agosto de 1994

43. Kahle, B: Wide Area Information Server Concepts. http://www.alexa.com/~brewster/essays/wais-concepts.htm 19 de Octubre de 1998.

44. The Harvest Summary Object Interchange Format (SOIF) http://harvest.transarc.com/Harvest/brokers/soifhelp.html 2 de Noviembre de 1998

45. Gravano, L; Chang, K; García-Molina, H; Lagoze, C; Paepcke, A: STARTS, Stanford Protocol Proposal for Internet Retrieval and Search. http://www-db.stanford.edu/~gravano/starts.html 19 de Enero de 1997.

46. Heery, R: Review of Metadata Formats. Program, 30(4):345-373, October, 1996. http://www.ukoln.ac.uk/metadata/review.html 4 de Noviembre de 1998.

47. Weibel, S: Metadata: The Foundations of Resource Description. D-Lib Magazine, July 1995 http://www.cnri.reston.va.us/home/dlib/July95/07weibel.Html 2 de Noviembre de 1998.

48. Massa, Javier: Elementos del conjunto de metadatos de Dublin Core: Descripción de Referencia. http://www.rediris.es/metadata/dublin_core_elements.es.html. Creado el 11 de Diciembre de 1997.13 de Mayo de 1998.

49. Weibel, S; Iannella, R; Cathro, W: The 4th Dublin Core Metadata Workshop Report. D-Lib Magazine, June 1997. http://www.dlib.org/dlib/june97/metadata/06weibel.html 2 de Noviembre de 1998.

50. NetScape Communications: Netscape works with W3C and leading content providers to drive new specification for organizing, describing and navigating information on Internet, intranets and desktops. http://home.netscape.com/flash1/newsref/pr/newsrelease488.html 8 de Septiembre de 1997.

51. World Wide Web Consortiun Publishes Public Draft of Resource Description Framework (RDF). http://www.w3.org/Press/RDF 3 de octubre de 1997

52. López, DR ; Massa, J: Dando forma al envase y, con ello, al contenido: Webber. Boletín de RedIRIS, (45):15-21, Octubre de 1998. http://www.rediris.es/rediris/boletin/45/enfoque1.html

53. San Segundo Manuel, R: Organización del conocimiento en Internet: Metadatos bibliotecarios Dublin Core. Sextas Jornadas Españolas de Documentación Automatizada. Valencia, 29 a 31 de Octubre de 1998. Vol 2: 805-818.

54. Waugh, A: Specifying metadata standards for metadata tool configuration. Seventh International World Wide Web Conference. Brisbane, 14 a 18 de Abril de 1998. http://www7.scu.edu.au/programme/fullpapers/1913/com1913.htm. 27 de Noviembre de 1998.

55. Marchiori, M: The limits of Web metadata, and beyond. Seventh International World Wide Web Conference. Brisbane, 14 a 18 de Abril de 1998. http://www7.scu.edu.au/programme/fullpapers/1896/com1896.htm. 27 de Noviembre de 1998.

56. Massa, J: Resumen de la reunión IRIS.SEARCH. http://rediris.es/si/iris-index/coord/gt5/resumen.txt. 7 de Mayo de 1998.

57. Sullivan, D: Search Engine Watch: Search Engines Features. http://searchenginewatch.com/webmasters/features.html 1 de Diciembre de 1998.

58. Thomas, CF; Griffin, LS: Who will create the metadata for the Internet?. First Monday, 2(12), December 1998. http://www.firstmonday.dk/issues/issue3_12/thomas/index.html 21 de Mayo 1999

59. O'Neill, ET; Lavoie, BF; McClain, PD: Web Characterization Project An Analysis of Metadata Usage on the Web. Annual Review of OCLC Research, 1998. http://www.oclc.org/oclc/research/publications/review98/metadata.htm.

60. Hermans. B: Intelligent Software Agents on the Internet. Tesis doctoral, Univ de Tilburg http://www.broadcatch.com/agent_thesis/ 25 de Julio de 1996

61. Hermans, B: Intelligent Software Agents on the Internet. An inventory of current offered functionality in the information society and a prediction of (near)future developments. First Monday, (2-3), 1997. http://www.firstmonday.dk/issue2_3/ch_123/index.html

62. Newell, SC: Improved Internet information retrieval through the use of user models, filtering agents and a knowledge-based system. Tesis doctoral, Department of Electrical Engineering, McGill University, Montreal, 1997.

63. Hermans, B: Desperately Seeking: Helping Hands and Human Touch. First Monday, 3(11), November 1998. http://www.firstmonday.dk/issues/issue3_11/hermans/index.html. 19 de Mayo 1999.

64. Wiederhold, G: Mediation in the architecture of future information systems. IEEE Computer, 26(3), 38-49 March 1992

65. Wiederhold, G; Genesereth, M: The Conceptual Basis for Mediation Services. IEEE Expert, Intelligent Systems and their Application, 12(5). September October 1997.

66. Haverkamp, DS; Gauch, S: Intelligent Information Agents: Review and Challenges for Distributed Information Sources. Journal of the American Society for Information Science, 49(4): 304-311, 1998.

67. Jansen J: Using and Intelligent Agent to enhace search engine perfomance. First Monday, 2(3), 1996. http://www.firstmonday.dk/issues/issue2_3/jansen/index.html

68. Daigle, LD; Newll, S: Intelligent Agents and the Internet Information Infrastructure. http://www.isoc.org/isoc/whatis/conferences/inet/96/proceedings/a4/a4_2.htm

69. Hearst, MA: Interfaces for Searching de Web. Scientific American (3), 1997. http://www.sciam.com/0397issue/0307hearst.html

70. Ask Jeeves: Asking questions to give you answers. Search Engine Report, November, 1998. http://www.searchenginewatch.com/sereport/9811-askjeeves.html

71. Members of the Clever Project: Hypersearching the Web. Scientific American, (6), June 1999. http://www.sciam.com/1999/0699issue/0699raghavan.html

72. Ford, R: Mac OS 8.5 Special Report: Sherlock. MacIntouch Special Reports, 1998. http://www.macintouch.com/m85_sherlock.html

73. Alexa: Searching Serendipity and more. Search Engine Report. January, 1998. http://www.searchenginewatch.com/sereport/9801-alexa.html

74. Internet Explorer 5 makes search easier. Search Engine Report. April, 1999. http://www.searchenginewatch.com/sereport/99/04-ie5.html

75. Hípola, P; Vargas Quesada, B: Agentes inteligentes, definición y tipología. Los agentes de información. El Profesional de la Información, 8(4): 13-21, 1999.

76. Rhodes, BJ ; Starner, T: Remembrance Agent: a continous running automated information retrieval system. First International Conference on The Practical Application Of Intelligent Agents and Multi Agent Technology: 487-495, 1996. http://rhodes.www.media.mit.edu/people/rhodes/Papers/remembrance.html

77. Weng Ngu, DS; Wu, X: SiteHelper: A Localized Agent that Helps Incremental Exploration of the World Wide Web. Sixth International World Wide Web Conference. 1997.

78. McClure, C; Bertot, JC; Hert, CA: Expanding Our Knowledge of Evaluating Information Services and Resources: Prelude to the Mid-Year Meeting. Bulletin of the American Society for Information Science, 25(4), April May 1999. http://www.asis.org/Bulletin/Apr-99/expanding_our_knowledge__.html

---

Carlos Benito Amat
(benito [at] rtvv [dot] es)