SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN DISTRIBUIDA EN INTERNET. UNA REVISIÓN DE SU EVOLUCIÓN, SUS CARACTERÍSTICAS Y SUS PERPECTIVAS. SEGUNDA PARTE

6 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 aquél 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! representa los mejores ejemplos de este modelo, de sus limitaciones y virtudes.

A este modelo se achacan muchas limitaciones (28). En primer lugar, los sistemas sólo cubren una mínima fracción de los recursos disponibles. Yahoo!, el de mayor cobertura, sólo alcanzaba en fecha reciente 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 propio sistema.

Muchos de los recursos incluídos 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 facilitan 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.

7 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 altas exhaustividad y 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. Tras diseñar una aplicación personal para encontrar información en Internet, elaboró la interfaz Web que permitiera su uso público. WebCrawler 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 (29).

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 (30).

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 u organización de sedes Web. OpenText se clausuró el 16 de marzo de 1998 como servicio general. Fue sustituído por Livelink Pinstripe como servicio especializado en información económica y financiera (31).

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 7.000 búsquedas semanales (32).

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 (33).

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 sector 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 incluídos en la cobertura de los sistemas es un tema recurrente con aportaciones siempre recientes (34).

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 período 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 persista 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)* (35).

Antes, sin embargo, de tratar de la comercialización como una de las tendencias detectables en el desarrollo de los sistemas de recuperación de información distribuida en Internet, es necesario atender a sus características operativas básicas, que inciden en sus resultados: nivel de cobertura real, nivel de indización, rendimiento de la recuperación, lenguaje de recuperación, relevancia y feedback y nivel de representación de los recursos recuperados.

7.1 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 albergados a uno o más ordenadores centrales. En éstos, 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.

El proceso es cíclico y pasa por las siguientes fases:

1. Creación de una cola o lista de páginas pendientes de exploración.
2. Elección de una de las páginas de la lista como punto de partida para la exploración.
3. Una vez localizada, extracción de todos sus enlaces. Cualquiera que no haya sido explorado se añadirá a la cola inicial.
4. Procesamiento del contenido de la página para extraer información de su título, cabecera, palabras clave y cualquier información adicional, que se almacena en una base de datos.
5. Elección de una nueva página, y vuelta al segundo punto.

El observatorio permanente mantenido por Danny Sullivan en el servicio Search Engine Watch, permite afirmar que 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 (36). 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:

1. 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.
2. Ciertos conjuntos de documentos, en los ámbitos científicos, contienen complejos diagramas y fórmulas, tampoco traducibles.
3. Algunas sedes no admiten examen por robots, porque son de pago *(The New York Times)* o porque distorsionarían sus cifras de audiencia (CNN).
4. 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 (37).

7.2 El mecanismo de indización y la función de relevancia

La exhaustividad de la indización de los recursos está lejos de ser completa, tal y como revelan las indicaciones acerca del número de páginas de cada sede rastreadas por los buscadores. Algunos de los diseñadores han defendido que las páginas de nivel inferior de una sede sólo suponen un contenido redundante. Sin embargo, la heterogeneidad de los documentos, antes aludida, difícilmente justifica que se midan por el mismo rasero las páginas personales de un investigador, que pueden ofrecer borradores o textos de sus trabajos, y las dedicadas al queso de Camembert (www.camembert-france.com), por ejemplo de comparación. En la asignación de valores a los términos de indización se basan tanto la indización probabilística, empleada por la práctica totalidad de los sistemas, como el cálculo de la relevancia y, por ende, la ordenación de los documentos recuperados.

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» (38). 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 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:

— Título.
— Encabezamientos y subencabezamientos (titulares y ladillos).
— Las 100 palabras de mayor «peso» (que asignan empleando pesos Tf*IDf de Salton).
— Las primeras 20 líneas.
— El tamaño en bytes.
— El número de palabras (39).

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». Los factores que afectan a dicho nivel son:

— Los términos de búsqueda se han hallado en el título o cerca del título del documento.
— El documento contiene una mayor frecuencia de los términos empleados en el perfil.
— El documento contiene términos de búsqueda con gran peso, es decir, con baja frecuencia absoluta en la base de datos (40).

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:

— La frecuencia con que aparecen los términos (en el documento).
— La proximidad entre los términos (en el caso de búsquedas con varios).
— La proximidad de los términos a la cabecera (head, title) del documento.

Los resultados extraños se deben a que los algoritmos valoran más las palabras únicas o poco frecuentes (en la totalidad de la colección) (41).

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 (el sistema operativo de NEXT). 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 mediante la siguiente función:

— Frecuencia en el documento/Frecuencia en el dominio de referencia.

Las palabras con mayor numerador y menor denominador se ponderan más. Aquéllas con bajas frecuencias, menos. Este tipo de ponderación recibe el nombre de ponderación de particularidad (particularity weighting) (42).

HotBot también emplea la frecuencia de los términos en el documento, la extensión del documento 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) proporcionada por los creadores de las páginas (43).

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 (44). 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 (45). 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 (46). Otros trabajos ofrecen resultados totalmente comparables (47).

8 Bibliografía

28. 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,* 26 de septiembre de 1996, <http://www.ub2.lu.se/desire/radar/reports/D3.11/>.
29. *A brief Story of WebCrawler,* 7 de septiembre de 1997, <http://voyeur.mckinley.com/WebCrawler/Help/AboutWC/WCStory.html>.
30. MAULDIN, M. L. *Lycos: Design Choices of an Internet Search Service,* IEEE Expert Online, 7 de julio de 1997, <http://www.computer.org/pubs/expert/1997/trends/x1008/mauldin.htm>.
31. SULLIVAN, D. Open Text, *Search Engine Report* (17), 31 de marzo de 1998, <http://searchenginewatch.com/sereport/>.
32. SELBERG, E.; ETZIONI, O. *Multi-Service Search and Comparison Using the MetaCrawler,* 9 de octubre de 1995, <http://www.w3.org/pub/Cobferences/www4/papers/169/>.
33. CHU, H.; ROSENTHAL, M. *Search Engines for the World Wide Web: A comparative Study and Evaluation Methodology,* 24 de octubre de 1996, <http://www.ais.org/annual96/ElectronicProceedings/chu.html>.
34. NOTESS, G. R. Measuring the Size of Internet Databases, *Database,* october 1997, <http://www.onlineinc.com/database/OctDB97/net10.html>.
35. SHAW, R. Crawlers, Spiders and Worms. *Web Week,* 1(2), 1 de julio de 1995.
36. SULLIVAN, D. *Search Engine EKG.* <http://searchenginewatch.com/egk.htm>.

37. BRAKE, D. Lost in Cyberspace. *New Scientist,* 28 de junio de 1997, <http://www.newscientist.com/keysites/networld/lost.html>.
38. SULLIVAN, D. How Search Engines Rank Web Pages. *Search Engine Watch,* 30 de junio de 1997, <http://searchenginewatch.com/rank.html>.
39. MAULDIN, M. L.; LEAVITT, J. R. R. Web Agent Related Research at the Center for Machine Translation. *SIGNIDR Meeting,* 15 de julio de 1994, <http://fuzine.mt.cs.cmu.edu/mlm/signidr94.htm>.
40. Infoseek Corporation: *Document Retrieval Over Networks Wherein Ranking and Relevance Scores are Computed at the Client for Multiple Database Documents,* 8 de septiembre de 1997, <http://software.infoseek.com/patents/dist_search/patents.html>.
41. JANSEN, J. Using an Intelligent Agent to enhace Search Engine Perfomance, *First Monday,* 2(3), 3 de marzo de 1997, <http://www.firstmonday.dk/issues/issue2_3/jansen/index.html>.
42. PINKERTON, B. *Finding What People Want: Experiences with the WebCrawler,* 1994, <http://info.webcrawler.com/bp/www94.html>.
43. HOCK, R. E. Sizing Up HotBot. Evaluating one Web Search Engine's Capabilities, *Online,* noviembre 1997, 28 de mayo de 1998, <http://www.onlineinc.com/onlinemag(NovOL97/hock11.html>.
44. HERREN, R. Cómo nos engañan los buscadores: Las trampas que hacen los motores de búsqueda y por qué ofrecen cada vez peores servicios. *Tiempo* (855): 98, 21 de septiembre de 1998.
45. POLLOCK, A.; HOCKLEY, A. What's Wrong with Internet Searching. *D-Lib Magazine,* marzo 1997, 17 de septiembre de 1997, <http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/march97/bt/03pollock.html>.
46. AIMC. Macroencuesta a usuarios de Internet, 20 de septiembre de 1998, <http://www.aimc.es/aimc/html/inter/02resul0.html>.
47. LÓPEZ ALONSO, M. A.; 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, p. 489-493.