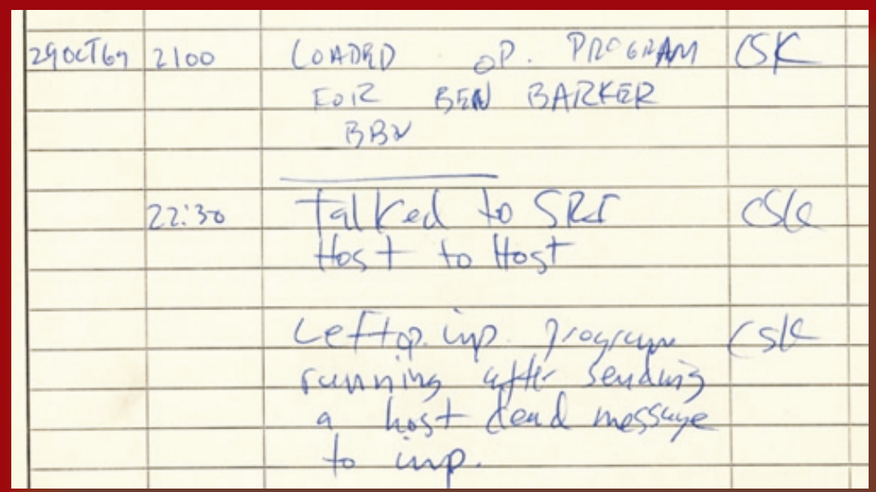
UNIVERSITAT DE VALÈNCIA

Facultat de Medicina

Evaluación de sistemas españoles de recuperación de información distribuida en Internet



Carlos Benito Amat • Tesis Doctoral • 2004

Tesis Doctoral
Carlos Benito Amat
Valencia, 2004

Indice

Página

Figuras, tablas y anexos

Página

1. Introducción

El conjunto de espacios informativos que, colectivamente, se denomina Internet, plantea serios desafíos desde el punto de vista de la documentación y la recuperación de información. Parece conveniente introducir este conjunto de problemas con una revisión de la evolución de Internet que, más que centrarse en los desarrollos técnicos, atienda a la progresiva configuración de su contenido informativo. Desde este punto de vista, Internet parece haber evolucionado en sentido centrífugo desde un estado de homogeneidad temática hasta un universo de gran heterogeneidad. Este acercamiento permite caracterizar de forma conveniente el universo documental que alberga y sus propiedades, que lo diferencian mucho del universo documental tradicional, alrededor de documentos y fuentes de información estructurados. Tras esta revisión, se examinan los sistemas para la recuperación de la información distribuida desarrollados en cada uno de los espacios que han venido integrándose en Internet y, especialmente, los del espacio Web. Más que disponerlos en orden cronológico, se propone una clasificación funcional de estos sistemas y se atiende a las ventajas e inconvenientes de cada modelo. Por último, se revisan los trabajos que han intentado evaluar los sistemas de recuperación de información distribuida como paso previo a establecer un plan de trabajo que permita evaluar los sistemas españoles de recuperación de información en Internet.

El examen de la evolución de Internet, el análisis de las características de la información y los documentos que contiene, el establecimiento de una taxonomía de sistemas para su recuperación y los métodos de evaluación de estos mismos sistemas se basan en

una revisión de la literatura amplia, pero especialmente centrada en las aportaciones más recientes y procedentes con frecuencia de campos no estrictamente relacionados con la documentación tradicional.

1.1 Algunas definiciones

El 24 de Octubre de 1995, el Federal Networking Council publicó una resolución que definía Internet en los siguientes términos:

> "Internet" se refiere a un sistema de información global que
> (i) está lógicamente interconectado por un espacio universal de direcciones basado en el protocolo internet (IP) y sus extensiones y desarrollos subsiguientes;
> (ii) es capaz de soportar comunicaciones que usen el conjunto Transmission Control Protocol/Internet Protocol (TCP/IP) sus futuras extensiones y desarrollos y otros protocolos compatibles con el IP;
> (iii) proporciona, usa o hace accesible, de forma pública o privada, servicios de alto nivel basados en la infraestructura de comunicaciones y otras infraestructuras relacionadas aquí descritas."
> (Federal Networking Council, 1995).

Esta definición, que resume el impresionante desarrollo iniciado a partir de las ideas sobre una "red galáctica" de JCR Licklider y W Clark (Licklider & Clark, 1962), se circunscribe por razones obvias a aspectos técnicos, dejando de lado el impacto que la interconexión de redes ha supuesto en muchos órdenes de la actividad humana. Una aproximación a este impacto se puede encontrar en la descripción que introduce el relato sobre el desarrollo de Internet ofrecido por la Internet Society:

> "Internet es a un tiempo una posibilidad de transmisión mundial, un mecanismo para la diseminación de información y un medio de colaboración e interacción entre individuos y sus ordenadores independiente de su localización física."
> (Leiner *et al.*, 1997; Leiner *et al.*, 2000).

Al igual que en la definición de Internet que éste mismo organismo proporciona:

> "Internet es una red de redes global que permite a ordenadores de cualquier tipo comunicarse directa y transparentemente y compartir servicios a través de la mayor parte del mundo. Además de su gran valor, que potencia las capacidades de tantas personas y organizaciones,

> constituye un recurso de información y de conocimiento y un medio de colaboración y cooperación entre incontables comunidades diversas”.
>
> (Internet Society, 2002).

Difusión, diseminación de información, colaboración e interacción personal son términos familiares a la actividad de los especialistas de la información, puesto que es la propia información el concepto subyacente a todos ellos. Para abordarlos adecuadamente en el contexto de Internet, se debe prestar atención a aspectos evolutivos y de desarrollo no necesariamente técnicos. Además, el examen de esta evolución puede permitir la identificación de las características más relevantes para las actividades de información y documentación sobre Internet.

Existen varias fuentes sobre los orígenes y el desarrollo de Internet. Algunas son directas, elaboradas por los propios protagonistas de los hallazgos (Leiner *et al.*, 2000). También se dispone de cronologías esquemáticas, aunque completas y suficientemente detalladas (Spicer, Bell, Zimmerman, Boas, & Boas, 2002; Zakon, 2003) (Davila, 2000) y de trabajos muy tempranos pero, acaso por ello, de gran valor para el estudio de Internet (Hardy, 1993) . Unas y otras fuentes se han revisado añadiendo una perspectiva nacional (Sanz, 1998) desglosando las diversas fases de su desarrollo (Nogales Flores, 1999) o relacionando las fases de su evolución técnica con los intentos paralelos de organización y control de la información distribuida a través de este medio (Griffiths, 2002; Amat, 2003) . Los siguientes apartados describen las fases sucesivas, que cabría interpretar como círculos concéntricos de progresiva extensión.

1.2 El periodo inicial: experimentación y uso compartido de recursos técnicos en el seno de ARPAnet.

El término “internet” aparece por primera vez a finales de 1974, como contracción de “internetwork” (Cerf, Dalal, & Sunshine, 1974). Valga esto como recordatorio de que Internet supone una interconexión de redes y sobrepasa, por tanto, el ámbito de las redes

aisladas. De hecho, el procesamiento online de transacciones se había inaugurado con el sistema SABRE, que IBM desarrolló para el sistema de reservas de American Airlines en 1964 pero, por mucho que supusiera la conexión de 2000 equipos remotos con dos ordenadores centrales (IBM 7090) con tiempos de respuesta de 3 segundos, en ningún momento se sobrepasaban los límites de una única red. Por el contrario, en 1965 Lawrence Roberts y Thomas Merrill consiguieron la conexión experimental de los ordenadores TX-2 del MIT y Q-32 de la Universidad de California en Los Angeles, creando así la primera red de ordenadores de área amplia. El resultado de este experimento fue "el reconocimiento de que los ordenadores de proceso compartido podían trabajar juntos, ejecutando programas y recuperando datos remotamente" (Leiner *et al.*, 2000). Sin embargo, fueron la confluencia de los avances en ingeniería de computación, los nuevos conceptos en la programación de interfaces y protocolos y, sobre todo, un proyecto multimillonario que acogiera esos y otros desarrollos lo que propició la aparición de ARPAnet: el 29 de Octubre de 1969, tras un primer intento fallido, se produjo la primera conexión con éxito entre el Stanford Research Institute y la Universidad de California en Los Angeles. De hecho, la elección de los 4 primeros nodos de la red estuvo dictada por las aportaciones de las instituciones participantes en el proyecto: 1) Los Interface Message Processors fueron un desarrollo de la empresa Bolt Benarek y Newman Ics. ; 2) el Network Control Protocol fue diseñado por el equipo de UCLA; 3) El Network Information Center dependía del Stanford Research Institute; 4) los métodos de representación de la información, especialmente los relacionados con funciones matemáticas, se habían desarrollado en la Universidad de California en Santa Bárbara (Culled-Fried interactive mathematics) y en la Universidad de Utah (gráficos), por último, 5) fue ARPA quien, en 1966, confió a Robert Taylor un presupuesto de un millón de dólares para dedicarlos a la investigación de una "red cooperativa de ordenadores de proceso compartido". La adquisición de ILLIAC-IV, el mayor supercomputador de la época, que Burroughs había desarrollado por contrato de la NASA, estaba destinada a posibilitar a los científicos el acceso remoto a sus recursos.

Una cultura sirve de base a éste y otros proyectos de la época: la cultura de los recursos compartidos. No se debe olvidar que ya en

1963 se había lanzado Syncom, el primer satélite de comunicaciones asíncronas y se había desarrollado el código ASCII para posibilitar el intercambio de información. ARPAnet se desarrolló con varios objetivos: el uso directo de servicios de hardware distribuidos, la recuperación de datos desde bases remotas y el uso compartido de subrutinas y programas, no disponibles en los ordenadores de los usuarios por la incompatibilidad entre ordenadores o entre lenguajes (Computer Museum History Center, 2002).

1.3 La extensión al mundo académico: comunidades homogéneas

Mientras el trabajo técnico sobre programas, protocolos y normas proseguía y el número de redes interconectadas aumentaba, dos nuevos factores inauguraron una nueva fase en la extensión del concepto de Internet. La aparición de las redes académicas y la constitución de comunidades especializadas.

El número de instituciones y organismos conectadas a ARPAnet no cesó de aumentar. Por lo general, se trataba de instalaciones industriales, empresas de consultoría o agencias gubernamentales directa o indirectamente implicados en el proyecto: BBN Inc, Xerox, NASA, ANSI y otros. Del auge de la incipiente red puede dar idea su volumen de transacciones, que excedía los 3 millones de “paquetes” diarios en 1974 (Spicer *et al.*, 2002).

Sin embargo, los hechos más relevantes en el progreso hacia Internet los protagonizó la National Science Foundation. De hecho, se ha llegado a considerar la participación de la NSF como el hito que separa el internet experimental del internet operativo. Las razones son sobradas.

En primer lugar, por su activo apoyo a las actividades de computación e interconexión de casi 120 universidades estadounidenses, cuyos científicos empleaban un protocolo especial para acceder al superordenador CDC 7600 del National Center for Atmospheric Research. En segundo lugar por su apoyo al proyecto de la red de investigación en ciencia de la computación, que se concretó en la aprobación de un plan quinquenal dotado con 5

millones de dólares en 1980. A principios del siguiente año, más de 200 ordenadores en docenas de instituciones se habían conectado a CSnet (Computer Science Network) (Spicer *et al.*, 2002). Justo en la primavera de 1981, se constituyó BITnet entre la City University of New York y Yale University. Esta red llegó a conectar más de 140 organismos en 49 países.

El objetivo de las redes académicas era el intercambio electrónico no comercial de información para la investigación y la educación. Se trataba de redes verdaderamente cooperativas, donde cada organismo participante aportaba líneas de comunicación, almacenamiento intermedio y capacidad de procesamiento suficiente para el funcionamiento de sus conexiones. Además, desde el punto de vista de los contenidos, proporcionaban cuentas de correo electrónico y listas de distribución (llegaron a constituirse 3000 foros especializados en BITnet). Se empleaban para la transferencia de programas y datos y para la rápida transmisión de mensajes interactivos (Corporation for Research and Educational Networking, 1997). A finales de los años 70 y durante toda la década de los 80, se puede afirmar que proliferaron redes de inspiración académica o bien de orientación temática muy específica: EARN (European Academic and Research Network) en Europa (1983), Janet (Joint Academic Network) en Gran Bretaña (1984) o RedIRIS en España son ejemplos procedentes del mundo educativo; USEnet (Universidades de Duke y Carolina del Norte inicialmente) y Eunet, constituídas alrededor del sistema UNIX, HEPnet y MFEnet (física de altas energías y magnetismo respectivamente, auspiciadas por el US Department of Energy) y SPAN (comunidad de los físicos espaciales de la NASA) ilustran las redes centradas en grupos especializados (Sanz, 1998; Leiner *et al.*, 2000).

La confluencia entre estas iniciativas y la red inicial no se hizo esperar. En 1983 se estableció una pasarela entre Csnet y ARPAnet. En Agosto de 1989 nació la Corporation for Research and Education Networking como resultado de la fusión de CSnet y BITnet y, por encima de todo, se debe entender como el factor decisivo de desarrollo el abandono de protocolos de comunicación propios y la adopción generalizada del protocolo de control de transacciones (TCP), que en 1982 sustituye al anterior protocolo de control de red (NCP). Este, al fin, es el protocolo adoptado por las redes europeas e internacionales que, gradualmente, fueron confluyendo en Internet.

Así,en España, RedIRIS abandonó OSI como apoyo de sus servicios para adoptar IP en 1991 (Sanz, 1998).

En líneas generales, la fase de expansión de la interconexión de redes al ámbito académico siguió un esquema similar al de la fase anterior: un organismo o agencia disponía de un programa que, con su apoyo presupuestario, posibilitaba la infraestructura principal y, de forma cooperativa, instituciones o grupos especializados se iban sumando a la iniciativa de forma directa o tras modificar sus programas y protocolos iniciales. El papel “troncal” lo representaban NSFnet en Estados Unidos y IXI (luego EMPB, European Multiprotocol Backbone y EBone) en Europa, auspiciados por el programa COSINE. Los grupos que constituyen nodos y esta es una característica distintiva de esta fase, se beneficiaban de la interconexión de redes sobre todo para “los servicios de correo electrónico, la transferencia de ficheros, el terminal virtual y la introducción remota de trabajos”. En esta fase y en círculos académicos “la existencia de una red de investigación, en estrecha colaboración con otras redes similares que por esas fechas iban apareciendo en otros países europeos, se consideraba como un instrumento indispensable para el progreso de las diversas disciplinas científicas y tecnológicas” (Sanz, 1998).

Si este esquema se reduce al ámbito nacional, los términos son los siguientes: el Plan Nacional de Investigación y Desarrollo de 1988 incluyó como programa horizontal el Programa de Interconexión de Recursos Informáticos (IRIS) que, tras el empleo de Iberpac bajo X.25 como infraestructura, crea ARTIX (ARTeria Iris X.25) en 1990. En Agosto de ese mismo año se produjo la primera conexión experimental con Internet, operativa en marzo de 1991 a través de SIDERAL y con la infraestructura de la red europea IXI. La proliferación de instalaciones bajo Unix, la aparición de la informática distribuida, el parque creciente de ordenadores personales y el abandono de la plataforma OSI a favor de TCP/IP son factores que contribuyeronn a la progresiva ampliación de la red y a la integración de las redes especializadas con protocolos propietarios en esta plataforma.

1.4 La diversificación de usuarios y la interrelación de contenidos

> "Existe el convencimiento creciente entre autores y académicos de que la Red no es un mero ensamblaje de hardware conectado por cables y operado por programas. Quizá el aspecto más interesante de la Red sea la nueva cultura que se está desarrollando sobre esa base. En la Red, la gente actúa de forma diferente a como actuaría cara a cara. De hecho, la Red no contiene una única cultura humana, si no muchas. Redes diferentes basadas en tecnologías diferentes han evolucionado dando lugar a subculturas...".
>
> (Hardy, 1993)

Valga esta prolongada acotación para introducir la serie de acontecimientos más importante y, sin embargo, menos estudiada en la literatura, en el desarrollo de Internet. El fenómeno de diversificación de los contenidos, central en la problemática de organización y recuperación de la información distribuida en Internet, parece ser fruto de tres series de factores: 1) la existencia de una "cultura de la colaboración" previa o paralela al desarrollo de la propia Internet; 2) la disponibilidad de herramientas simples de creación y de interrelación de los documentos y de información contenida en ellos, acompañadas de otras que permitían su visualización y 3) la consideración de las conexiones como oportunidades de progreso social y de ventaja económica.

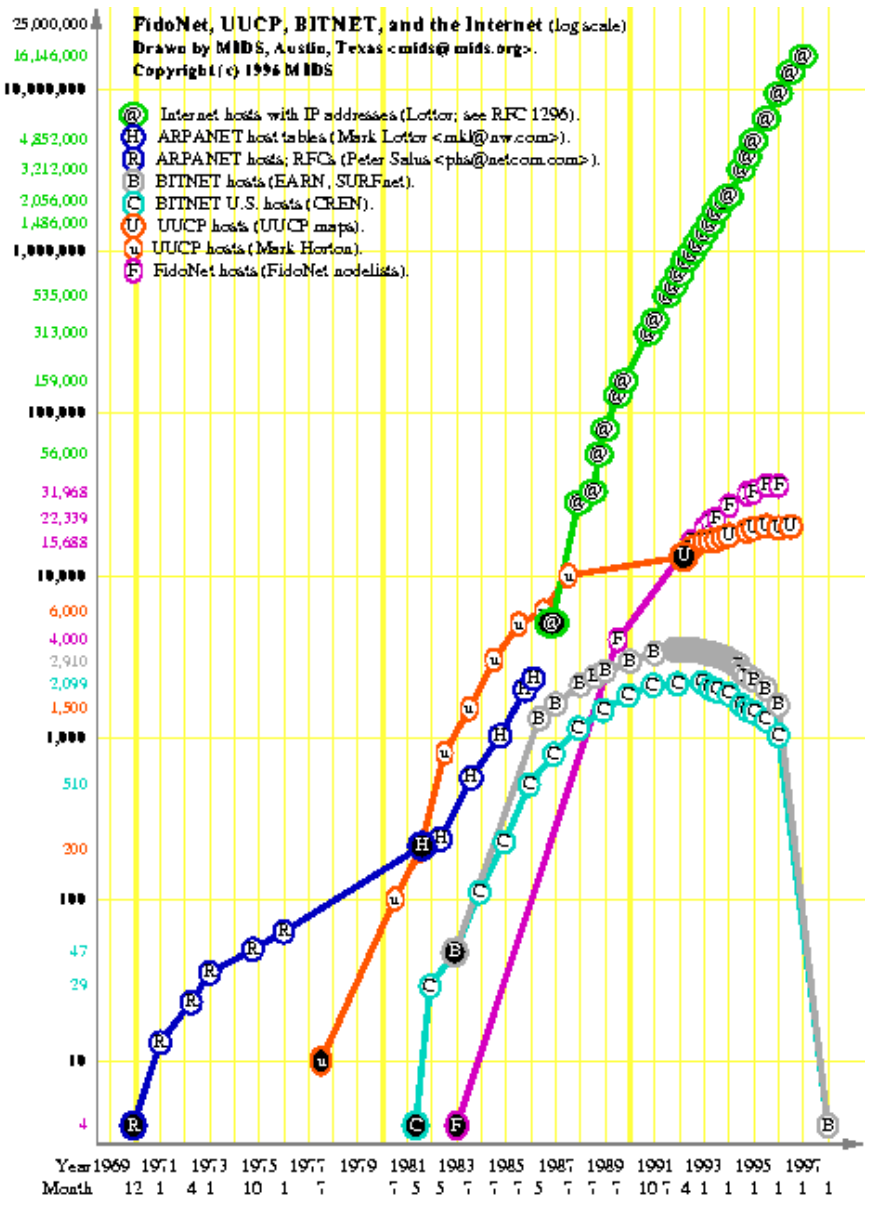


Figura 1.1: Evolución comparativa del volumen informativo de las redes de almacenamiento y distribución, ARPAnet y número de servidores IP entre 1969 y 1997 (Quaterman, 1996).

La tabla 1.1 (página 14) presenta en orden cronológico algunos acontecimientos relevantes en esta fase. Se puede comprobar en ella que no es posterior a las anteriores, sino que se imbrica en ellas. De hecho, en fecha tan temprana como 1974 la empresa Bolt, Benarek y Newmann Inc., participante en el desarrollo de los Interface Message Processors iniciales, ofrecía su sistema Telenet como servicio **público** de comunicaciones a través de la reciente tecnología de conmutación de paquetes. Bajo la dirección de Lawrence Roberts, Telenet conectaba clientes en 7 ciudades y representó la primera red de valor añadido (Computer Museum History Center, 2002). Muestras de la "cultura de la colaboración" aludida en el párrafo anterior son las redes de almacenamiento y distribución ("store-and-forward networks"). Tanto USEnet (University of North Carolina, 1979-) como BITnet (City University of New York, 1981-) o FIDOnet (Tom Jennings, San Francisco, 1984-) comparten una misma arquitectura: sus usuarios conectan a un servidor que, a su vez, les redirige a otro ordenador donde se organizan los envíos (postings, news) en grupos temáticos (newsgroups) jerarquizados. La creación de foros y el intercambio de información entre grupos e individuos marca el inicio de la divulgación o popularización de Internet. La figura 1.1 muestra, justamente, el despegue de las redes de almacenamiento y distribución. En algunas estimaciones, el número de usuarios de USEnet se cifraba en 265.000, los de BITnet se calculaban en 50.000 y a Compuserve se adjudicaban 370.000 (Quaterman, 1996).

Sin embargo, no son los aspectos cuantitativos los más destacables en esta fase sino, como se expondrá a continuación, la ruptura del esquema habitual de desarrollo.

Tras 5 años de colaboración en el marco del proyecto MULTICS (Multiplexed Information and Computing Service), los laboratorios Bell de ATyT se retiraron para iniciar la investigación y el desarrollo de lo que, en 1969, apareció como UNIX. Es relevante la diferencia de orientación entre ambos proyectos: MULTICS era otro proyecto más para el desarrollo de un sistema de procesamiento compartido; UNIX se orientó hacia un sistema abierto, colaborativo. Esta orientación (y su gratuidad) facilitaba su extensión, posibilitaba

que sus diseñadores contaran con la experiencia de los usuarios y permitía a académicos y estudiantes el diseño de aplicaciones propias. La aparición de una amplia comunidad de usuarios se revela como una consecuencia natural. Poco importa que fuera la afición compartida al ajedrez la que impulsó a Tom Truscott y Jim Ellis primero a aprovechar las facilidades de conferencia del sistema y luego a la adaptación del Unix-to-Unix Copy Protocol (UUCP, Mike Lesk en ATyT Bell Labs, 1976) para el transporte de “news”. Es más significativa el desarrollo, a través de USEnet, de comunidades de usuarios. En este sentido, este entorno también difería del correspondiente a ARPAnet: éste último se caracterizaba por la existencia de *listas*, mientras que USEnet se caracterizaba por la existencia de *foros* (Griffiths, 2002).

USEnet se presentó oficialmente en el encuentro del Academic UNIX Users Group de enero de 1980. A mediados de ese año, a las universidades de Duke y North Carolina se habían conectado otros centros. De ellos, resultan especialmente relevantes los situados en los laboratorios Bell, que financiaban los costes de las comunicaciones y la Universidad de Berkeley, que también participaba en ARPAnet. A finales de 1981, la red conectaba 150 ordenadores y 400 un año más tarde. Las previsiones sobre el volumen de las comunicaciones resultaron demasiado modestas: de una estimación de 2 ó 3 mensajes diarios se había pasado a un número real de 50 por las mismas fechas. En 1986, existían 241 grupos temáticos o foros extendidos por 2500 puntos de conexión que intercambiaban información a un ritmo de 500 mensajes o artículos diarios (Hardy, 1993; Griffiths, 2002).

Se ha mencionado anteriormente la participación de los laboratorios Bell sufragando los costes de comunicaciones de USEnet. Su papel es homólogo al de otras agencias e instituciones en fases anteriores aunque los costes de las comunicaciones se repartían entre los diversos “nodos” de la red y sus instituciones patrocinadoras. Lo que aparta esta fase del esquema de las anteriores es, justamente, el conflicto que se estableció entre los administradores de la red y sus usuarios. Son identificables dos componentes en este conflicto. La primera se refiere al uso de la red, la segunda al contenido de los mensajes.

En cuanto al uso, la existencia de encadenamientos de mensajes (chain letters) el envío inmoderado de mensajes insultantes “flaming” y otras prácticas obligaron a la adopción de unas reglas (“netiquette”, marzo de 1986) que limitaran el volumen de tráfico generado. En cuanto al contenido, ha de tenerse en cuenta que los 3 foros troncales se habían diversificado hasta abarcar grupos de dudosa seriedad: net.jokes, net.rumour o net.bizarre se pueden citar como ejemplos.

El grupo de administradores de sedes (conocidos como el conciliábulo o USEnet Cabal) se enfrentaban con el continuo aumento de los costes de transmisión, especialmente aquellos correspondientes a las líneas que unían los centros estadounidenses con los de Gran Bretaña, Europa continental (la comunicación con el Centro de Matemáticas de la Universidad de Ámsterdam costaba 6 dólares por minuto) y Australia. Para racionalizar los flujos de información y sus costes, se crearon nuevas categorías entre julio de 1986 y marzo de 1987 y se restringió la comunicación de los contenidos de algunas (talk o soc, chismes y ligoteo) entre centros estadounidenses y, especialmente, con Europa. Este proceso de reorganización (Great Renaming) suscitó de inmediato acusaciones de censura y una ruptura en el seno del Cabal. Puede resultar anecdótico que las dificultades para crear foros sobre recetas, sexo, drogas o rock and roll impulsaran a algunos de los administradores de sedes a la creación de la categoría alt. Pero, más allá de las orientaciones de los grupos, lo que marca la gran diferencia con las fases anteriores es el fin de la especialización y la apertura de las infraestructuras de comunicaciones a comunidades diversas imbuídas de la cultura del intercambio de contenidos (Hauben & Hauben, 1996). En efecto, a la consideración de USEnet como un sistema técnicamente abierto y de contenidos heterogéneos hay que sumar otra característica distintiva: la falta de una autoridad de control. Los esfuerzos del grupo de administradores de sedes (Cabal) se vieron minimizados en primer lugar por la transición desde el protocolo UUCP (una serie de programas diseñados para el “shell” de los sistemas UNIX) hasta el NNTP (Network News Transfer Protocol) que permitía el almacenamiento e intercambio de news o artículos entre sistemas y plataformas diferentes (Kantor & Lapsley, 1986). Además, la aparición de la categoría alt. llevó aparejado el empleo de plataformas igualmente alternativas: acaso

el mejor ejemplo fue la iniciativa de Rick Adams (Center for Seismic Studies) que condujo a UUnet (Bumgarner, 2002), un servicio que terminaría por ser tan comercial como Compuserve o AOL (America Online) (Griffiths, 2002) otros dos grandes protagonistas del espacio de artículos (news) organizados y participados por foros (groups). La reciente adquisición del archivo de Dejanews y la recopilación de artículos procedentes de USEnet arrojó la cifra de 149.808.000 news (Google Groups Team, 2001).

Con todo, el hecho fundamental en esta fase fue la aparición de un nuevo espacio informativo que, si cabe, magnificaba la heterogeneidad de sus contenidos y, andando el tiempo, también su volumen. El 2 de Agosto de 1991, se sometía al grupo alt.hypertext una pregunta:

> "Alguno de los lectores de este grupo está al tanto de trabajos de investigación o desarrollo sobre (...) la posibilidad de emplear los hiperenlaces para permitir la recuperación a partir de múltiples fuentes heterogéneas de información?..."
>
> (Kannan, 1991)

Tras responder de forma sumaria el 6 de Agosto, Tim Berners-Lee, desde el Organismo Europeo de Investigación Nuclear (CERN) ofreció un resumen de acciones sobre el proyecto World Wide Web que se iniciaba con las siguientes frases:

> "El proyecto WWW combina las técnicas de recuperación de información e hipertexto para configurar un sistema de información global simple, pero poderoso. El proyecto se inició con la filosofía de que la mayor parte de la información académica debe estar libremente disponible para todo el mundo. Se propone permitir el intercambio de información entre equipos dispersos a escala internacional y la difusión de información por grupos de apoyo".
>
> (Berners-Lee, 1991).

Resulta significativo que en el origen de la web se mencione el problema de la recuperación de información. Sobre todo teniendo en cuenta que en la propuesta original de Berners-Lee a la dirección del CERN, que data de marzo de 1989, se compara los sistemas de

información jerarquizados con los basados en palabras clave y se proponen los sistemas de información basados en enlaces como ventajosos frente a los demás modelos (Berners-Lee, 1989).

Sin embargo, interesa por el momento destacar la aparición de este concepto y de las herramientas que siguieron, como los originarios de otro gran espacio informativo en Internet: el espacio Web. De alguna manera, los trabajos de programación en el entorno NeXTStep tanto del editor como del programa de visualización, la segregación de HTML como subconjunto del lenguaje de marcas normalizado SGML o los seminarios internos en el CERN se ajustan a un esquema comparable a la fase más técnica del desarrollo de ARPAnet. El ofrecimiento de los programas a través de FTP, en agosto y la presentación pública en el encuentro Hypertext'91, en diciembre de ese año, precedieron a la aparición de los primeros programas cliente de visualización de documentos HTML: Erwise (29 de Abril) y Viola (15 de Mayo de 1992). Es notable el hecho de que, fuera del círculo más técnico, la primera presentación de WorldWideWeb se dirigió a la comunidad de físicos de altas energías en el congreso de septiembre de 1992. El hecho más destacable de aquellos días y aquel que mayor repercusión alcanzará, fue la declaración del Consejo de Directores del CERN del carácter público y abierto de la tecnología WWW, efectuada el 30 de Abril de 1993 (Caillou, 2002). En septiembre de ese mismo año, los trabajos de Marc Andreesen y colaboradores en el National Center for Supercomputing Applications (NCSA) produjeron el programa de visualización Xmosaic (Andreesen, 1993) y el servidor httpd, posteriormente comercializados como NetScape y Apache respectivamente.

Las herramientas de programación, visualización en modo gráfico y montaje de servidores Web compartían el carácter "abierto" que había posibilitado al entorno UNIX acoger el espacio USEnet. La conveniencia de este carácter abierto se reconocía de forma explícita entonces y años después, así:

"...se definió el Hypertext Markup Language como el formato de datos que se transmitieran por escrito. Dada la previsible dificultad de animar al mundo a usar el nuevo sistema de información global, HTML se eligió por su similitud con algunos sistemas basados en SGML para facilitar su adopción por la comunidad

de la documentación, en cuyo seno SGML era la única sintaxis considerada como normalizada.”.

(Berners-Lee, 1996)

A nadie puede extrañar que, gracias a este carácter abierto, a la hospitalidad del protocolo http y a una incansable labor de promoción, el número de servidores Web (26 en noviembre de 1992) se hubiera duplicado en enero de 1993 y que, entre marzo y septiembre de ese mismo año, las comunicaciones a través del protocolo http pasaran de suponer el 0,1% al 1% del tráfico en la red troncal de la NSF (Caillou, 2002) .

La tabla 1.1 incluye, a título meramente ilustrativo, determinados hitos de la incorporación al Web, como espacio de distribución informativa y de transacción, de algunas instituciones y empresas españolas (Peiró, 1996; Adell, 2002). Pero en esa misma tabla destaca con fuerza la incorporación de grandes comunidades de usuarios de servicios telemáticos implantados, por lo general, entre el final de los años 70 y el principio de los 80: Minitel en Francia ((Kessler, 1995; Johnstone & Carlson, 2002) Captain en Japón, Telidon en Canadá, Prestel en Gran Bretaña (Johnstone *et al.*, 2002) y, naturalmente, Ibertex en España (Fernández Beobide & González Obiol, 1992). Se trataba en todos los casos de sistemas unidireccionales, con comunidades de usuarios que representaban “demanda” de información, mientras que al otro lado se situaban grandes instituciones, agencias o empresas proveedoras de servicios informativos.

En la misma tabla se ordenan cronológicamente los “desplazamientos” de las empresas y organismos proveedores y la integración paulatina en el espacio Web. Así, de la misma forma que el acuerdo establecido con Goya Servicios Telemáticos permitió la migración de los usuarios de Ibertex al Web, el acuerdo entre IBM y France Telecom del Otoño de 1998 se dirigía al desarrollo del software para la integración de ambas plataformas (Bergonneau, 2002).

Tabla 1.1 Algunos hitos del desarrollo temprano de Internet y su progresiva popularización

1974	Bolt, Benarek y Newman Inc crea Telenet, el primer servicio público con tecnología de conmutación de paquetes
1977-78	Ward Christianson inventa el primer Bulletin Board System para R-CPM
1979	British Telecom. Inaugura el servicio PRESTEL Compuserve se convierte en el primer servicio que ofrece cuentas de correo electrónico y apoyo técnico a usuarios de ordenadores personales Se establece la Unix Users Network (USEnet) entre las universidades de Duke y North Carolina
1980	Compuserve es el primer proveedor que ofrece servicio online de chat en tiempo real
1981	Se despliega en Francia el servicio MINITEL Las universidades de Yale y City University of New York establecen BITnet
1982	Compuserve transforma su Network Services Division para proporcionar capacidad de interconexión en áreas amplias a clientes corporativos
1986	La CTNE presenta públicamente el servicio Ibertex Goya Servicios Telemáticos ofrece servicios de correo electrónico a empresas
Marzo de 1989	Propuesta sobre gestión de la información dirigida al CERN por Tim Berners Lee
1990	World Comes Online comienza a ofrecer conexión a Internet a sus abonados
1991	Goya Servicios Telemáticos comercializa, desde su nodo español de Eunet, el acceso a Internet a empresas y particulares
1993	Un equipo de la Universitat Jaume I de Castellón registra en el CERN la primera sede web española
4 de Julio de 1994	Un acuerdo entre Eunet (Goya) y Cesel posibilita el acceso a Internet de más de 400.000 usuarios españoles de Ibertex
1 de Abril de 1995	Se inaugura la versión electrónica del diario Avui
12 de Mayo de 1995	Banesto es la primera entidad financiera española en ofrecer servicios a través de Internet
Julio de 1995	Se inaugura en Madrid la “Ciberteca”
Enero de 1996	Entra en operación Infovia, un servicio de conexión a Internet a coste de llamada local de la CTNE

Se ha señalado el carácter “unidireccional” de los servicios de videotex. Acaso el cambio más radical que supuso su integración en la web no se cifró en la potenciación de la oferta de servicios y contenido, sino en la posibilidad de que amplias comunidades de usuarios emplearan la sintaxis de HTML para la elaboración de documentos que posteriormente serían distribuidos. Aunque inicialmente se desarrolló un editor en modo gráfico para la confección de páginas en lenguaje de marcas y luego se crearon aplicaciones profesionales de alta gama para la confección y montaje de páginas y sedes Web, la aportación de mayor trascendencia en esta línea la ofreció a finales de 1995 la empresa californiana Vermeer Technologies, que diseñó los programas FrontPage y Personal Web Server para el entorno Windows 95. Justo en el momento en que ATyT incluía estos programas en su Easy World Wide Web Service, como medios para “desarrollar y mantener sedes web de calidad” en su sistema de alojamiento de sedes, Microsoft adquirió Vermeer Technologies y sus productos, que posteriormente incluyó dentro del conjunto de programas MS-Office desde la versión 98 de su sistema Windows. No habían transcurrido dos año desde la creación de Vermeer Technologies en abril de 1994 ni dos meses desde el lanzamiento de los programas, dato que ilustra a las claras el acelerado ritmo de desarrollo que el espacio Web estaba adquiriendo (Milne, 1995; AT&T, 1995; Microsoft, 1996).

Contemporáneos en su aparición con el WorldWideWeb fueron dos espacios claramente diferenciados de éste: gopher y WAIS. Además, una importante iniciativa contribuyó al crecimiento y la heterogeneidad de Internet como espacio informativo: la formación de la Commercial Internet eXchange (CIX) Association, Inc. por General Atomics (CERFnet), Performance Systems International, Inc. (PSInet) y UUNET Technologies, Inc. (AlterNet), que hubo de esperar a que la National Science Fundation levantara la prohibición para el uso comercial del tráfico de Internet a través de su red troncal en 1991. (Zakon, 2003).

El gopher se inició como un servicio de distribución de documentos de la Universidad de Minnesota (Lindner, 1991) en la Primavera de 1991. El usuario del “gopherespacio” combinaba el recorrido por menús jearquizados, la visualización de documentos

seleccionados y la recuperación a texto completo. Originalmente se diseñó como un sistema de información de campus, aunque en noviembre de 1994 se estimaba que los 8000 servidores gopher existentes se había extendido más allá de los centros universitarios. La segunda generación del sistema, Gopher+, añadía metadatos a los recursos jearquizados para facilitar su recuperación (McCahill & Anklesaria, 1995).

WAIS (Wide Area Information Servers) distribuía en múltiples servidores diversas bases de datos especializadas, con un índice central. A través de un programa cliente, los usuarios podían consultar los textos completos de los documentos incluídos. WAIS utilizaba un protocolo propio, extensión de la norma Z39.50 (`Information Retrieval Service Definition and Protocol Specification for Library Applications, de ANSI) (Kahle & Medlar, 1991).

Tanto en uno como en otro caso, eran los administradores de los sistemas y servidores quienes, en última instancia, publicaban sus menús, documentos y bases de datos. De ahí el contraste con el espacio Web, totalmente “democrático” y abierto a la incorporación indiscriminada de contenidos.

En Abril de 1994, el coordinador del sistema de información del campus de la Universidad de Michigan formulaba la siguiente pregunta:

> “A juzgar por las cifras Z39.50 de NSFnet, el crecimiento en la cantidad de datos servidos por WAIS parece idéntico al de las cantidades servidas por Gopher y Web. Incluso si se contempla el hecho de que las pasarelas Gopher enmascara muchas de las transacciones WAIS y se tiene en cuenta que la cantidad de datos transmitidos no es necesariamente proporcional al valor de la base de datos, se podría concluir de estas cifras que, de momento, los medios de visualización son “vencedores . ¿ Es correcta esta conclusión?”.
> (Wiggins, 1994)

En un intercambio de artículos de marzo de ese mismo año, a propósito de la comparación estadística de Gopher y WWW, se esgrimía un aumento del tráfico anual en el espacio Gopher del

997%, mientras que el incremento de intercambios en el espacio Web crecía anualmente un 341.634% (Walton, 1994). Precisamente Jordi Adell intervino para mencionar las estadísticas de Merit Network como fuente fiable de los datos (Adell, 1994). No parece descabellado que el equipo que estaba desarrollando el World Wide Web lo hubiera calificado como "**el** universo de información", integrador de las restantes tecnologías:

> "La combinación de las técnicas de hipertexto, recuperación de información y de redes de área amplia produce el modelo W3"

De hecho, el concepto de "espacio informativo" que desde un primer momento se asoció con la web, no sólo estaba apoyado en el esquema de denominación de documentos. El hecho de que su propia denominación lo calificara de "global" se apoyaba en el concepto de enlace hipertextual y en la integración del protocolo http con WAIS, los sistemas jerarquizados (incluyendo gopher), FTP, el sistema normalizado de denominaciones X.500 y otros (Berners-Lee, Caillou, Groff, & Pollermann, 1992).

1.5 Espacios informativos en Internet: características y accesibilidad

Hata aquí se ha empleado la metáfora de los círculos concéntricos para caracterizar la evolución de Internet. En realidad, una metáfora más apropiada puede ser la empleada por Mauldin, de quien se reproduce la figura 1.2. La comunicación que presentó en la tercera Conferencia Mundial sobre la web (11 de Abril de 1995) contenía dos elementos de interés. En primer lugar, su definición del espacio Web:

> "Definimos la web como [el espacio] que incluye cualquier documento en los espacios FTP, Gopher o http" (Mauldin, 1995).

En segundo lugar, el propio diagrama de la figura 1.2 (a continuación):

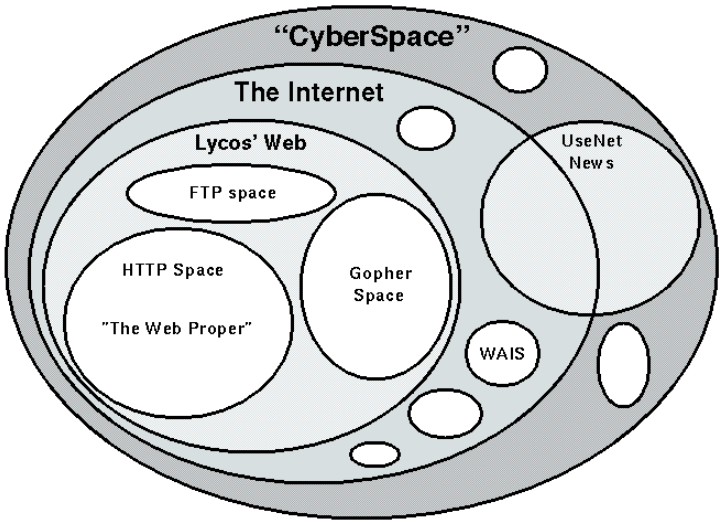


Figura 1.2:Distribución de los espacios informativos que configuraban el universo del "Ciberespacio" (Mauldin, 1995)

En él se contemplan diversos conjuntos intersectantes o disjuntos. Estos conjuntos se diferencian no sólo desde un punto de vista técnico, por la diversidad de protocolos y formatos que cada uno emplea. Las diferencias más notables radican en sus contenidos y, especialmente, en los productores de tales contenidos. Ni que decir tiene que la voluntad de integración del W3 acabó proporcionando una interface común a todos ellos; es decir, se alcanzó en fecha temprana cierta transparencia desde el punto de vista técnico a través de la creación de pasarelas entre protocolos. Lo que persiste aún hoy en día es la heterogeneidad en los formatos de los documentos y su organización en cada espacio.

El énfasis en el W3 con que se inicia esta sección se justifica precisamente en función de la preponderancia que este espacio informativo alcanzó al poco tiempo de su lanzamiento y conserva en la actualidad. Un vistazo a su evolución cuantitativa (Zakon, 2003) lo confirma. Y, aunque no es posible hacer sinónimos Internet y World Wide Web, también es necesario reconocer que, en los poco más de 10 años transcurridos desde su aparición, ha concentrado la mayor parte y las mejores aportaciones y desarrollos.

1.5.1 Volumen y crecimiento

Hay pocas cifras tan provisionales y relativas como las estimaciones del tamaño del espacio W3. Sin embargo, ha existido y existe práctica unanimidad en considerar este espacio como ingente y de tamaño masivo: en abril de 1995, Mauldin, empleando los programas de recopilación del sistema Lycos, estimaba la existencia de 23.550 servidores Web y 4051 millones de URLs (Mauldin, 1995); en noviembre del mismo año, Tim Bray elevaba a 89.271 el número de sedes Web y reducía a 11.366.121 el número de URLs únicas; el equipo del cento de investigación de NEC aportó en 1998 una primera estimación (Lawrence & Giles, 1998) y luego una segunda que corregía y completaba la anterior (Lawrence & Giles, 1999). 16 millones de servidores; de ellos 2.8 millones eran públicamente accesibles,(ver más abajo) y contenían 800 millones de páginas. La estimación más reciente, realizada por el Web Characterization Project de la oficina de investigación de OCLC en Junio de 2002, arroja un número de 3.080.000 sedes y 140 mil millones de páginas en el espacio público del W3 (O'Neill, Lavoie, & Bennett, 2003). Los datos del registro delegado en España (ES-NIC) muestran la evolución desde 7.219 sedes (dominios de segundo nivel) en 1997 hasta 48.933 en Junio de 2003 (2003). Si se admite el promedio de páginas por sede estimado por Amat, el número total de páginas en el dominio .es sería actualmente de 4.423.543 (Amat, 2003). Según Baeza, estas cifras representan la tercera parte de las sedes españolas (Baeza-Yates, 2002).

El estudio de los investigadores de OCLC contiene un sorprendente hallazgo:

> "... las evidencias sugieren que el crecimiento de la Web Pública, medido por el número de sedes Web, ha alcanzado una meseta. La tasa anual de crecimiento se ha ralentizado durante el quinquenio que abarcan las encuestas del Web Characterization Project; en el último año, la web pública contrajo ligeramente su tamaño."
> (O'Neill *et al.*, 2003)

El hallazgo no parece coherente con toda la serie de estudios contemporáneos que coinciden en atribuir un modelo de crecimiento

exponencial a la Web, partiendo del trabajo de Michalis, Petros y Christos Faloutsos:

> "A pesar del aparente desorden de Internet, hemos descubierto algunas leyes exponenciales sorprendentemente simples en su topología. Estas leyes proceden de tres instantáneas de Internet, tomadas entre noviembre de 1997 y diciembre de 1998 y muestran un crecimiento del 45% en el periodo. Estas leyes se ajustan muy bien a los datos reales, resultando en coeficientes de correlación iguales o superiores al 96%". (Faloutsos, Faloutsos, & Faloutsos, 1999).

Sin embargo, la contradicción desaparece si se tiene en cuenta que el modelo exponencial de crecimiento no sólo se aplica al número de sedes o páginas: también a su "popularidad" (número de visitas a determinada sede) (Adamic & Huberman, 2001) y, especialmente, al número de enlaces internos y externos que se establecen entre sedes y documentos. Se trata, en definitiva, de un cambio no sólo de volumen, cuantitativo, sino también de "densidad", estructural (Huberman & Adamic, 1999; Menczer.F, 2002; Yok, Hawoong J, & Barabasi, 2002).

1.5.2 Estabilidad y dinamismo

Además del ingente volumen, la segunda característica significativa de Internet como espacio informativo es la gran inestabilidad de los documentos que contiene y distribuye. La dinámica es bastante compleja y admite diversos indicadores:

> "Los cambios en la topografía de la web se pueden expresar al menos de cuatro formas: (1) más sedes en más servidores de más lugares, (2) más páginas y objetos añadidos a las páginas y sedes existentes, (3) cambios en el tráfico y (4) modificaciones en los textos, los gráficos y otros objetos Web existentes". (Koehler, 2002)

Tal inestabilidad ha suscitado gran número de trabajos por su implicación en la arquitectura de los sistemas (en el almacenamiento intermedio o "caching" de páginas) y, naturalmente, en la tasa de refresco o ritmo de actualización de índices de los sistemas. 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 (Woodruff, Aoki, Brewer, Gauthier, & Rowe, 1996). 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 (Chankhunthod *et al.*, 1996). En estudios no directamente relacionados con sistemas de recuperación operativos y de gran rigor estadístico, se han alcanzado similares estimaciones (Douglis, Feldmann, Krishnamurthy, & Mogul, 1997; Cho & García-Molina, 1999; Brewington & Cybenko, 2000; Fetterly, Manasse, Najork, & Wiener, 2003). Especialmente valiosos son los conceptos manejados por Koehler y sus resultados. En su tesis, identifica y cuantifica no sólo cambios en el contenido de las páginas, sino también en la estructura y en la densidad de éstas y de las sedes que las albergan (Koehler, 1999). Tras 4 años de seguimiento, éste mismo autor conluye que la vida media de una página web es de 2 años y que parece producirse una estabilización del contenido de las páginas tras periodos iniciales de cambio (Koehler, 2002). En otro trabajo, directamente relacionado con el rendimiento de los sistemas de recopilación, 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 (Peterson, 1997). Parecidos resultados se hallaron al estudiar los efectos del dinamismo del espacio Web sobre la cobertura de 7 grandes sistemas de recuperación a los largo de 5 meses de 1998 (Bar-Ilan, 1999).

1.5.3 Disparidad

Una tercera característica contrastada de los documentos en Internet y el espacio W3 es su heterogeneidad. Esta característica ya se ha adelantado como una tendencia en el repaso de las fases evolutivas. Como tal, se justifica en la progresiva “democratización”

de los contenidos, cuya elaboración queda al alcance de cualquiera capaz de elaborar un artículo destinado a un foro o una página o conjunto de páginas albergadas en un servidor http. Los documentos Web son dispares en función de su orientación temática, de su formato, de su estructura y dimensiones y de su función.

La distribución temática de la información distribuida a través de la web también muestra gran variabilidad en repetidos muestreos tanto globales (Lawrence *et al.*, 1999) como a nivel más reducido (Martínez de Lejarza Esparducer, 1999). Los dominios funcionales (org, net, etc. en oposición a los geográficos) son sólo una pálida indicación de la diversidad temática. Una idea más ajustada a la heterogeneidad del espacio Web la proporciona el seguimiento del Open Directory Project (ODP) una iniciativa colectiva de clasificación de sedes puesta en marcha originalmente el 5 de junio de 1998 con 2.000 categorías que organizaban 27.000 sedes. El 2 de julio, el número de categorías era ya de 3.900. Cuando, el 16 de abril de 1999, se anunció la integración del ODP y el sistema Lycos, 43.000 sedes se distribuían en 65.000 categorías. En julio de 2003, 3.800.000 sedes se distribuyen en 460.000 categorías temáticas diferentes (Wikipedia, 2003).

Pero, además de la heterogeneidad de los contenidos, también hay que contar con la disparidad de formatos que pueden albergar los documentos, de naturaleza compuesta casi siempre. Amat, por ejemplo, identifica 5 tipos diferentes de elementos textuales (Amat, 2003). Mucho más esclarecedor es el trabajo de Stephani Hass y Erika Grams, quienes realizan una calsificación funcional de las sedes Web, las páginas HTML y los enlaces. Las autoras concluyen su trabajo expresando la necesidad de un nivel de análisis superior al sintáctico para determinar la temática de las páginas (Haas & Grams, 2000).

Sobre el tamaño de los documentos Web 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 (Bray, 1996). 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 (d.e. =2 Kb). La extensión

máxima fue de 1,6 Mb20 (Woodruff *et al.*, 1996). 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 (Térmens Graells, Ribera Turró, & Sulé Duesa, 2003). Sobre una muestra aleatoria del espacio Web correspondiente al dominio .es, también se halló gran variabilidad en los tamaños de las páginas, en la composición relativa de elementos textuales y gráficos y en otros elementos (Amat, 2003).

1.5.4 Accesibilidad

Se han acuñado muchos términos para referirse a los diversos grados de accesibilidad del espacio Web y algunos de sus componentes: “deep web”, “public web”, “invisible web”, “hidden web”, “internet invisible”, “infranet” son sólo algunas de las expresiones empleadas. Todas ellas hacen referencia a la posibilidad de reconocimiento y posterior recuperación de los contenidos que alberga.

Acaso el estudio de mayor envergadura sobre la accesibilidad de la información en el espacio Web sea el de Michael K Bergman, basado en una recopilación de datos realizada a lo largo de una semana del mes de marzo de 2000. Algunos de sus hallazgos resultan demoledores:

> “La información pública en la web profunda es actualmente entre 400 y 550 veces más grande que en lo que se define normalmente como World Wide Web (...) Más de la mitad del contenido de la web profunda reside en bases de datos especializadas (...) Hasta un 90% de la web profunda es accesible públicamente, sin estar sujeta a pagos o suscripciones”.
>
> (Bergman, 2001).

Isidro Aguilló define “internet invisible o infranet” como el “conjunto de recursos accesibles únicamente a través de algún tipo de pasarela o formulario Web y que, por tanto, no pueden ser indizados de forma estructural por los robots de los buscadores” (Aguilló, 2000). En realidad, lo que dificulta la localización y organización de esos contenidos es la necesidad de una transacción o interacción entre sistemas o entre usuarios y sistemas. Se pueden

distinguir tres casos. Los espacios corporativos restringidos (normalmente intranets institucionales o empresariales) cuya navegación requiere el reconocimiento de la dirección del equipo cliente constituyen un primer ejemplo. Los servicios que exigen el registro (gratuito o venal) de sus usuarios con anterioridad a la transacción y aquellos otros cuyos documentos (páginas HTML) se generan en respuesta a una solicitud, de forma dinámica, completan la tipología. Ya no es cierto que determinados formatos dificulten la recopilación de los documentos que soportan, su análisis y su recuperación. En la actualidad es posible la recuperación, a través de una amplia variedad de sistemas, de documentos en formato de intercambio (PDF) y los generados con aplicaciones gráficas (TIFF, PostScript y otros). Tomando en cuenta las características de accesibilidad, es posible la siguiente representación del espacio Web:

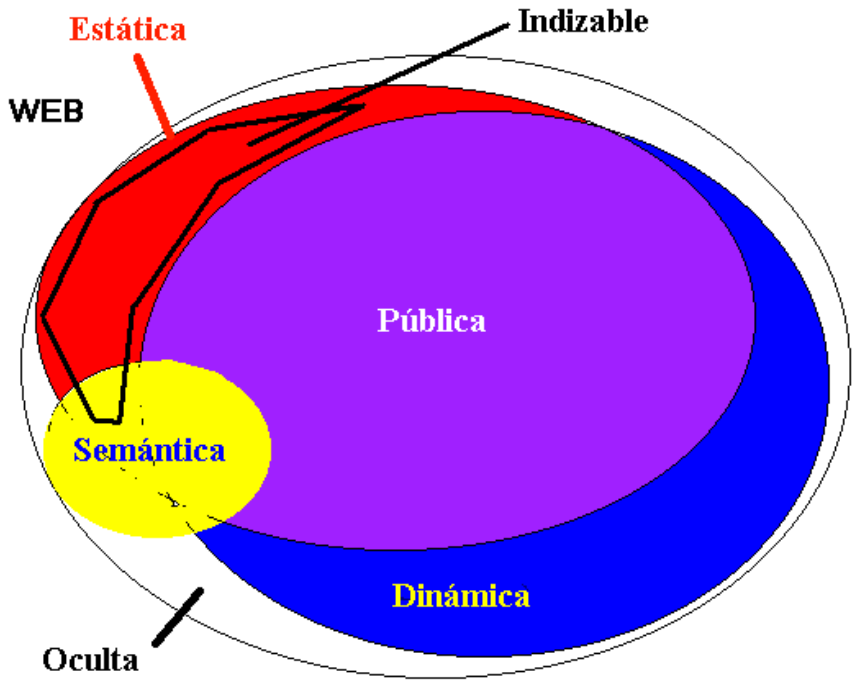


Figura 1.3: Distribución del espacio Web (la Web) según la accesibilidad de la información distribuida (Baeza-Yates, 2004)

1.5.5 El contraste con los sistemas tradicionales

En su discusión sobre las bases teóricas de las bibliotecas digitales, Edward Fox y Shalini Urs establecen una comparación útil entre bibliotecas tradicionales y digitales basada en informes de la iniciativa de la National Science Foundation. En realidad, su análisis se puede aplicar tanto a las bibliotecas en sí como a sus fondos. Según su esquema, las bibliotecas tradicionales son 1) estables, de lenta evolución; 2) de contenido mayoritariamente textual; 3) organizado de forma simple; 4) generado a través de un sistema de publicación reglado en que predomina la revisión por expertos... y 5) de unidades documentales no interrelacionadas dinámicamente por su contenido. En contraste, las bibliotecas digitales son 1) altamente dinámicas, con documentos efímeros y variables; 2) que tienen una naturaleza compuesta y variable en formatos y dimensiones; 3) cuyas estructuras guardan gran riqueza contextual; 4) su valor informativo sólo se determina a posteriori, mediante el uso y... 5) altamente interrelacionados (Fox & Urs, 2002). Aunque los citados autores no van más allá, cabría introducir otro elemento de contraste que resulta de la mayor importancia en el conexto de este trabajo: el contenido de las bibliotecas tradicionales recurre a sistemas de recuperación (catálogos manuales o automatizados y bases de datos bibliográficas) que organizan representaciones de los documentos, dada la relativa escasez de puntos de acceso físico a los propios documentos. Sin embargo, el espacio Web e Internet en general proporcionan acceso a documentos íntegros. Un comentario algo más pormenorizado comparando ambos medios (Amat, 1998; Amat, 1999) ofrece los siguientes puntos de contraste:

1) 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.

2) 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.

3) 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 diferenciadora de Internet: los documentos no son controlados antes de su publicación.

4) 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.

5) 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.

6) 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.

7) 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.

8) 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.

Además, las propias características de los documentos de Internet agraban los desafíos impuestos a los sistemas de recuperación (Baeza-Yates, 2003). A pesar de todo ello, ¿ Existen

sistemas capaces de abarcar la información distribuida en Internet, organizarla y posibilitar su recuperación de forma coherente?.

1.6 Respuesta de los sistemas de recuperación.

No ha existido espacio informativo distribuido en red que no cuente con uno o muchos sistemas de recuperación de los documentos que contiene. En palabras de uno de los adelantados de tales sistemas

> "Dado el número de servidores empleados hoy en día como sedes informativas, existen muchas dificultades para encontrar los programas que se desean en un entonrno distribuido. Uno puede estar seguro de que los programas están ahí, pero a veces es muy difícil encontrarlos".
>
> (Emtage, 1990).

El sistema archie, desarrollado en la McGill School of Computer Science es, en efecto, el primero de los dedicados a la recuperación de información distribuida en los diversos espacios de Internet. En concreto, el primer componente de archie mantenía una lista o índice de unas 600 sedes con programas informàticos accesibles mediante el protocolo FTP. Un segundo componente permitía a usuarios remotos la consulta del índice mediante cadenas de caracteres.

La tabla 1.2 ordena cronológicamente la aparición de los primeros sistemas de recuperación de información distribuida en Internet.

Tabla 1.2: Cronología de aparición de los primeros sistemas de recuperación en Internet*

Fecha	Sistema	Espacio
14 de noviembre de 1990	archie	FTP
17 de noviembre de 1992	veronica	gopher
1 de abril de 1993	World Wide Web Wanderer	World Wide Web
30 de noviembre de 1993	ALIWEB	World Wide Web
20 de enero de 1994	Galaxy	World Wide Web
1 de abril de 1994	Yahoo!	World Wide Web
20 de abril de 1994	WebCrawler	World Wide Web
7 de noviembre de 1994	Harvest	World Wide Web
12 de julio de 1995	MetaCrawler	World Wide Web
15 de diciembre de 1995	Altavista	World Wide Web

*Las fechas proceden de los anuncios comunicados a la lista de sistemas de información de USENET (comp.infosystems).

A Archie, presentado en noviembre de 1990 como sistema de recuperación del espacio FTP (Emtage, 1990) siguieron Veronica, centrado en el espacio gopher (Foster, 1992) y, un año después, ALIWEB (Koster, 1993) el primero de los muchos sistemas dedicados a la recuperación en ese espacio de integración de protocolos y contenidos que acabaría siendo el espacio Web. La cronología inluye el lanzamiento del World Wide Web Wanderer (Matthew Gray, MIT) en abril de 1993 porque, aunque tenía por objeto el seguimiento del crecimiento del espacio Web, resultó el primero de una serie de “robots” que circulaban por la red recopilando información sobre sus sedes. WebCrawler, por su parte, es el primer robot de un sistema de recopilación automática de documentos Web a texto íntegro (Pinkerton, 1994). Galaxy se adelantó en unos meses a Yahoo como directorio o sistema de clasificación jerárquica de sedes Web (Behlendorf, 1994). Por último, MetaCrawler apareció en julio de 1995 como un prototipo, luego consolidado como sistema de búsqueda simultánea y en paralelo de otros (Selberg, 1995).

Una cronología de la aparición o implantación en España de diversos sistemas de recuperación se ofrece, a título meramente ilustrativo, en la tabla 1.3.

Tabla 1.3: Cronología de aparición de algunos sistemas españoles de recuperación en Internet*

Fecha	Sistema
12 de mayo de 1999	Altavista
1996	apali !
28 de octubre de 1996	BIWE
1 de febrero de 2000	enlaweb
5 de mayo de 1999	Excite
12 de mayo de 1996	OLE
11 de mayo de 1996	Ozu
20 de noviembre de 2001	Salman
15 de agosto de 1997	Telépolis

*Las fechas corresponden al registro de los diversos sistemas en ES-NIC.

Es posible extrapolar el crecimiento de Internet y el espacio Web al sector de los sistemas de recuperación de la información distribuida: Ricardo Fornás había recopilado 3027 sistemas de búsqueda en 642 temas hasta el 20 de agosto de 2003 en su servicio buscopio (http://www.buscopio.net). Más adelante se comprobará que también la especial dinámica de Internet ha alcanzado a estos sistemas.

1.6.1 Clasificación de los sistemas

Se han propuesto diversos esquemas clasificatorios de los sistemas de recuperación de información distribuida en Internet. Ya en 1993, se ofrecía un recuento de los sistemas disponibles, apenas una mera enumeración con detalle de algunas características

(Danzig, Obraczka, & Li, 1993). En contraste, el equipo de diseño del sistema WAIS aportaba en fecha aún anterior un análisis de los sistemas existentes basado en un conjunto de criterios clasificatorios que mantienen su validez (Schwartz, Emtage, Kahle, & Neumann, 1992). Los criterios esgrimidos eran:

> Granularidad (Granularity)
> Sistema de almacenamiento (Distribution)
> Topología de interconexión (Interconnection Topology) y
> Esquema de integración de datos (Data Integration Scheme)

El concepto de granularidad es, en cierto modo, similar al de exhaustividad de la indización. Refleja el nivel de representación de los contenidos de los documentos y la unidad de recuperación y será discutido más adelante. Los criterios relativos al sistema de almacenamiento y a la topología son arquitectónicos. El esquema de integración de datos, en contraste, está directamente relacionado con las modalidades de recopilación de información de los sistemas e, igualmente, se tratará a continuación.

El NetLab de la Universidad de Lund realizó, en el marco del programa europeo Telematics, una taxonomía más comprensiva sobre criterios diferentes (Koch, Ardo, Brümer, & Lundberg, 1996). En su esquema se distinguen inicialmente tres grandes grupos de servicios:

1) los de amplia cobertura, que abarcan varios espacios informativos
2) los servicios dedicados a la recuperación de información en espacios limitados o sobre recursos específicos y
3) las recopilaciones temáticas de recursos.

El primer grupo distinguía sistemas de acceso público y programas comerciales empleados en sedes corporativas. Los primeros se distribuían en 3 grandes grupos: sistemas de recopilación automática (*robot based indexes*), sistemas manuales

(*template based indexes*) y sistemas de agentes (*"intelligent" agents*). El alcance geográfico y la orientación temática se combinaban para generar subgrupos. Por otro lado, se reservaba un apartado diferente para los sistemas de búsqueda múltiple (*combined, simultaneous and collected ind*exes).

Esta clasificación se ha tomado como base para, empleando una combinación de criterios funcionales y fases evolutivas, ofrecer 4 grandes grupos (Amat, 1998) a pesar de que difícilmente se puede hablar de sistemas que se encuadren únicamente en una de las categorías.

Antes de exponer esta taxonomía, conviene atender a dos aportaciones más recientes. La primera, contenida en el texto de Baeza y Ribeiro de 1999, diferencia 3 modos de recopilación y construcción de sistemas: el primero de ellos corresponde a los sistemas de recopilación automática, el segundo a las listas y directorios de recopilación manual. Previamente, estos autores han caracterizado el modelo de los sistemas disponibles como búsqueda "sintáctica" (Baeza-Yates & Ribeiro-Neto, 1999). Sin embargo, a la tercera modalidad de recuperación, empleada por sistemas que "explotan la estructura de hiperenlaces de la web" bien pudiera denominarse búsqueda "estructural o hipertextual".

La segunda aportación, más reciente, distribuye los sistemas de recuperación según una "taxonomía de las búsquedas en la web" (Broder, 2002). Sobre una muestra aleatoria de peticiones (*log analysis*) y cuestionarios, formula una tipología de las búsquedas y distingue las búsquedas directas (*navigational*), las informativas (*informational*) y las transaccionales (*transactional*).

Las búsquedas directas corresponden a las búsquedas de un item conocido en el contexto de la recuperación tradicional. Su propósito es llegar a una sede o elemento determinados que el usuario ya conoce, bien por previas visitas o por la suposición de su existencia. Las búsquedas informativas tratan de hallar información supuestamente disponible en el espacio Web en forma estática. Es decir, se basan en la supuesta existencia de uno o más documentos que contienen la información requerida. No requieren acciones

posteriores a la construcción del perfil y su ejecución. Las búsquedas transaccionales tratan de alcanzar un destino donde se produzca una interacción. Tal interacción constituye la transacción que define estas demandas. Un ejemplo simple es el interés por el horario de un trayecto ferroviario, que previamente requiere la búsqueda de la sede de la compañía.

A continuación, Broder describe la evolución de los sistemas de recuperación por generaciones. La primera, que sitúa entre 1995 y 1997, está muy próxima a la recuperación de información tradicional y se basa, además, en el hallazgo de documentos estáticos y en el contenido explícito de los mismos. Los sistemas de segunda generación, entre 1998 y 1999, posibilitan las búsquedas informativas, además de las directas (a hito) y emplean elementos ajenos al contenido meramente textual de los documentos, sobre todo los textos de los enlaces y los datos de navegación. La generación actual de sistemas, que aparece ahora, trata de superar las limitaciones del corpus textual existente en los documentos empleando análisis semántico, determinación de contextos, selección dinámica de bases de datos y otros recursos para posibilitar los 3 tipos de búsqueda mencionados.

A la luz de estas aportaciones, es posible revisar la distribución de los sistemas en listas y directorios, bases de recopilación automática, sistemas con indización asistida y empleo de sistemas de agentes.

1.6.2 Sistemas manuales: 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 de “guías de navegación”. 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 recopilados manualmente. Yahoo! representó en sus inicios el ejemplo más popular de este modelo. Open Directory lo sigue más estrictamente hasta la actualidad.

Básicamente, los directorios presentan una clasificación jerarquizada de sedes Web agrupadas temática y/o geográficamente. Cada entrada se puede acompañar de textos descriptivos o críticos sobre el contenido de la sede y de otros elementos. La definición de categorías, el hallazgo y la descripción de las sedes y su clasificación se realizan de forma manual. El manejo fundamental se basa en el descenso a través de la jerarquía partiendo de amplias cabeceras de serie hasta alcanzar la clase o grupo que más se ajusta al tema de búsqueda solicitado. En casi todos los casos es posible la búsqueda por términos de los títulos de las sedes, de los nombres de las categorías o de los comentarios.

A este modelo se achacan muchas limitaciones (Koch *et al.*, 1996). 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.

Se puede establecer una correspondencia con la distribución de Broder para concluir que los directorios proceden de la primera generación de sistemas. Son sintácticos, es decir, ofrecen mecanismos de recuperación textuales. La exhaustividad de su análisis es muy baja (baja granularidad) puesto que representan sedes, no documentos individuales. Están basados en el análisis humano (manual) de las mismas y, originariamente, se dirigían a espacios informativos determinados.

1.6.3 Los sistemas de recopilación automática y recuperación sintáctica

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. En conjunto, se puede afirmar que los sistemas de recopilación automática afrontan con mayores garantías la recuperación en un entorno tan dinámico y cambiante como el espacio Web y, en general, Internet (Risvik & Michelsen, 2002). Los sistemas de este grupo se basan en la ejecución de programas que, mediante conexiones http, localizan sedes y documentos, los rastrean y construyen índices centralizados del contenido que sus textos expresan.

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, convertido en "producto antes que en tesis" (Pinkerton, 2000). Este sistema se lanzó el 20 de Abril de 1994, con documentos procedentes de unas 6.000 sedes. El 14 de Abril de ese mismo año, alcanzó el millón de búsquedas. La arquitectura de WebCrawler, su sistema de agentes de recopilación (hasta 15) y otras características (Pinkerton, 1994; Pinkerton, 1994) han permanecido como modelo para la mayoría de los sistemas posteriores.

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 (Mauldin, 1997). Su proceso de recopilación (Mauldin, 1998) ha servido de modelo tanto teórico como real a muchos otros sistemas.

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 (Sullivan, 1998). El lanzamiento al público de MetaCrawler se realizó el 7 de Julio de 1995 (Selberg, 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 (Selberg & Etzioni, 1995). 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 (Digital Equipment Corporation, 1995).

Esta breve relación cronológica revela, una vez más, el origen académico de muchos de los sistemas (a veces surgidos de proyectos escolares o académicos) 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. Una de las razones se expone en el siguiente párrafo.

A medida que cada base de datos crecía, se hizo necesario contar con mayor poder de procesamiento para su mantenimiento. Esta necesidad favoreció el movimiento hacia el patrocinio comercial o la adquisición de los servicios por empresas. WebCrawler, por ejemplo, consiguió patrocinio comercial el 1 de diciembre de 1994. En octubre de 1995 se financiaba exclusivamente a través de la publicidad. En junio de ese mismo año fue adquirido por el proveedor America Online quien, el 1 de abril de 1997, lo vendió a Excite y, a su vez, lo vendió a InfoSpace en 2001 (Pinkerton, 2002).

La exigencia de que los mensajes comerciales se difundieran a un número cada vez mayor de potenciales consumidores acentuó 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)".

(Shaw, 1995).

1.6.3.1 El mecanismo de recuperación de los sistemas sintácticos

La oferta masiva de resultados, resultante de la búsqueda del mayor número de conexiones posibles y el énfasis en la exhaustividad de la recuperación, ha obligado a los diseñadores de sistemas al desarrollo de programas que no sólo recuperen documentos, sino que también los ordenen en función de su relevancia.

El algoritmo básico 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". 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, el tradicional "inverse document frequency weight" (Salton & McGill, 1983). 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. En la descripción del diseño de Lycos (Mauldin, 1997), este conjunto de variables se expresa mediante las siguientes cuestiones:

> ¿ Cuántos términos del perfil se incluyen en el documento ?

¿ Cuán próximos están los términos del perfil en el documento (proximidad) ?
¿ Dónde aparecen los términos del perfil en el documento (posición)?
¿ En qué medida se corresponden los téminos del perfil con palabras aisladas ?

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 (Mauldin & Leavitt, 1994). 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" (Kirsch, 1997). 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 (Burrows, 1998).

La base de datos de WebCrawler tiene 2 componentes: un índice a texto completo y una representación de la 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) (Pinkerton, 2000).

La experiencia de usuarios finales y también la de los documentalistas o recuperadores profesionales no parecen muy

favorables. Son continuas las denuncias sobre la excesiva comercialización de los servicios, sobre todo por el recurso de "pago por posición"[1]. 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 (Pollock & Hockley, 1997). Igualmente son indicativos los resultados de Bruce que, sobre una muestra de personal académico australiano, estimó entre 4.0 y 4.2 (sobre una escala de 1 a 6) el grado de satisfacción con los resultados de búsqueda en Internet con estos sistemas (Bruce, 1998). Otros estudios emplean grupos experimentales procedentes del mundo académico o educativo y llegan a conclusiones similares: 30 por ciento de resultados nulos en el caso de Wang *et al.* (Wang, Wawk, & Tenopir, 2000). En contraste, los estudios basados en el análisis de conexiones (log analysis) que reflejan el uso de los sistemas de búsqueda por la población general, no ofrecen indicaciones directas sobre la satisfacción con los resultados (Wolfram, Spink, Jansen, & Saracevic, 2001).

Marcia Bates y Trudi Bellardo han enfocado los sistemas de recuperación de información distribuida en Internet mediante la comparación con los sistemas tradicionales de bases de datos bibliográficas. Ninguno de los trabajos se basa en análisis propios, pero ambos concluyen una serie de carencias de los sistemas. En concreto, Bates esgrime las siguientes:

1. Uso de esquemas clasificatorios anticuados.
2. Sumisión a la falacia de la "ontología".
3. Ignorancia de la distribución de Bradford.
4. Ignorancia de la sensibilidad al volumen de la información de la recuperación de datos.
5. Permitir errores en el procesamiento por humanos
6. Ignorancia de la experiencia en el trabajo informativo.

[1] Procedimiento por el que se situa un documento o conjunto de documentos en la cabecera de la lista de resultados de búsqueda, a cambio de una compensación económica o patrocinio que los titulares de tales documentos realizan al servicio de recuperación

Y sugiere, además, el empleo de tesauros generales y especializados (Bates, 2002). Por su parte, Bellardo defiende el empleo de indicadores de evaluación tradicionales para mejorar el rendimiento de los sistemas, aunque su tratamiento es ciertamente inespecífico (Bellardo Hahn, 1998).

Ni estos trabajos ni otros similares representan una evaluación formal de los sistemas de este grupo.

A diferencia de los de primera generación, los sistemas de recopilación automática presentan una granularidad (exhaustividad en la indización) elevada. No están orientados a espacios determinados y, por propia naturaleza, afrontan con cierto éxito los requisitos de un entorno tan dinámico como Internet.

Sus programas de recopilación son hipertextuales y sólo en algún caso se ha empleado el procesamiento de datos contextuales para la recuperación (la “Intelligent Concept Extraction” de Excite, solicitud de patente retirada en fecha no especificada). Por otra parte, ciertos sistemas (AltaVista, HotBot) pronto posibilitaron el seguimiento de los enlaces de las páginas recuperadas. A pesar de esto, la naturaleza de estos sistemas es sintáctica: se basan en el hallazgo de los términos del perfil en los documentos recuperados que, por otra parte, se ordenan según cálculos aplicados ya en el contexto de la recuperación de información tradicional.

1.6.4 El análisis contextual y de enlaces de los sistemas de recuperación estructural

La idea de emplear elementos estructurales, como los hiperenlaces, para mejorar el rendimiento de la recuperación, es tan antigua como el propio espacio Web o aún más (Croft & Turtle, 1989). El periodo en que Broder sitúa los sistemas de segunda generación, sin embargo, marca las fechas de aplicación de esas ideas a los sistemas reales. El uso de criterios de popularidad, de técnicas métricas y, especialmente, de la información de y “alrededor de” los enlaces contenidos en documentos del espacio Web han contribuído muy significativamente a una nueva (y más exacta) consideración del problema de la recuperación en Internet.

DirectHit no constituye un sistema en sí mismo, sino un mecanismo de mejora de relevancia a través del “recuento de popularidad”. En palabras de su inventor:

“ [Es] un método de organización de la información en el cual se registra la actividad del usuario que busca, y se utilizan los datos de ese registro para organizar los documentos en búsquedas sucesivas del mismo o de otros usuarios”.

Originalmente, se basaba en una mecánica simple:

> Proporciónese un índice que sea capaz de almacenar términos clave y asociar cada documento con al menos uno de ellos, siendo además capaz de asociar puntuaciones para cada término en el documento en el momento del almacenamiento, de forma que sea posible asociar a cada término del índice las puntuaciones en cada documento ;
> acéptese una primera búsqueda de un primer usuario;
> identifíquense los términos clave que se ajustan a la primera petición;
> preséntense los documentos relacionados con la primera búsqueda al primer usuario;
> permítase que el primer usuario seleccione al menos uno de los documentos resultantes de la primera búsqueda y sea éste un documento seleccionado;
> modifíquese el índice, de forma que la puntuación del término clave asociado al documento seleccionado modifique su valor relativo respecto a las puntaciones de los términos restantes;
> (modifíquese el índice, de forma que las puntuaciones totales de los términos clave de al menos uno de los documentos seleccionados se alteren en relación con las puntuaciones totales de los restantes términos;
> acéptese una segunda búsqueda de un segundo usuario;
> identifíquense los términos clave que se ajustan a la segunda petición, que llamaremos el segundo conjunto de términos;
> preséntense los documentos resultantes de la segunda petición al segundo usuario, de forma que se organicen en orden decreciente de las puntuaciones de sus términos clave siempre que exista al menos un término coincidente entre el primer y el segundo conjunto, de forma que el documento seleccionado para el segundo usuario se ordenará por encima del lugar que ocupaba antes de que el primer usuario lo seleccionara”.
> (Culliss, 1999)

Más allá de la mecánica concreta que este sistema propone, lo importante es que representa el primer ejemplo de empleo de información contextual (las acciones de los usuarios tras la obtención de resultados de búsqueda) que, en términos tradicionales, cabría interpretar como un procedimiento de *relevance feedback* indirecto.

DirectHit no tardó en aplicarse a sistemas que, hasta ese momento, proponían una recuperación meramente sintáctica. Pero, al mismo tiempo, otros dos algoritmos, traducidos a sendos sistemas, hacían su aparición en el panorama de los sistemas de recuperación en el espacio Web: HITS y PageRank.

El sistema HITS (Hipertext-Induced Topic Search) fue desarrollado por Jon Kleinberg durante una estancia en el Centro de Investigación de IBM en Almaden (Kleinberg, 2000). Posteriormente, se incorporó al sistema Clever (Clever Project, 1999). En cuanto a PageRank, obra de estudiantes de Stanford (Page, 2001) se incorporó al sistema Google (Brin & Page, 1998) y continua siendo su esencia.

HITS no constituye en sí mismo un sistema sino un procedimiento de clasificación automática de documentos en el espacio Web que parte de resultados previos de búsqueda:

> "…Se selecciona un conjunto inicial de páginas, preferentemente realizando una petición convencional basada en palabras clave y luego se seleccionan las páginas que enlazan con las resultantes de la primera búsqueda o con las que las páginas resultantes conectan (...) Después, de forma repetitiva, se calculan valores de autoridad para las páginas del conjunto inicial, basándose en sus enlaces de partida y de llegada. Se definen una o más comunidades o "vecindarios" de páginas relacionadas en función de esos valores. Es probable que tales comunidades sean de interés y valor para el usuario interesado en la búsqueda por palabras clave de una página determinada".
>
> (Kleinberg, 2000).

Por el contrario, PageRank se emplea en el seno del sistema Google:

> "Un método asigna orden de importancia a nodos de una base de datos interrelacionados, como cualquiera que contenga citas, el world wide web o cualquier base de datos de hipermedios. El orden asignado a un documento se calcula a partir del correspondiente a los documentos que lo citan. Además, se hace intervenir en el cálculo una constante que representa la probabilidad de encontrar por azar un documento. El método es especialmente útil para mejorar el rendimiento de los resultados de sistemas de recuperación en bases de datos de hipermedios, como el world wide web, cuyos documentos ofrecen una calidad muy dispar" .
>
> (Page, 2001).

En ambas acotaciones, resultan significativos algunos términos: *authoritativeness*, *variation in quality*. Además, las listas de referencias de las correspondientes patentes y las comunicaciones contemporáneas no pueden ser más reveladoras: es habitual en estos y otros documentos referencias a los trabajos de Henry Small y Francis Narin. En efecto, indicios de la integración de la investigación de estos sistemas en la corriente del análisis de citas y la bibliometría de evaluación. Por otra parte, la divulgación, siquiera parcial, de estos algoritmos ha generado la integración de las investigaciones sobre sistemas de recuperación de información distribuida en el espacio Web con los sistemas avanzados de recuperación textual. Siguiendo esta línea, se ha llegado a estudiar la combinación, por ejemplo, de HITS con otros modelos de recuperación (Okapi, Cover Density Ranking, Three-Level Scoring Method o el modelo general de espacio vectorial) en busca de una mejora, que efectivamente se produjo, del rendimiento (Li, Shang, & Zhang, 2002).

Cabe añadir que existe en la actualidad un elemento contextual en estos mecanismos: el procesamiento de los textos de los documentos enlazados y también de los fragmentos del texto del propio documento que “rodean” a cada hiperenlace (Eiron & Mccurley, 2003).

Revisiones relativamente recientes revelan el modelo de sistemas hipertextuales o basados en elementos estructurales de búsqueda, como un campo extremadamente fructífero de investigación (Greco, Greco, & Zumpano, 2001; Picard & Savoy, 2003).

1.6.5 Los sistemas de indización asistida

A diferencia de los directorios, que sometían a los documentos de Internet a una representación estructurada en el momento de su incorporación al sistema, otros sistemas se han basado en que los propios creadores de los documentos describan su contenido en el momento de su producción. Los sistemas distribuidos de primera generación, las iniciativas de metadatos y el concepto de Web semántica comparten este modelo de funcionamiento o, mejor, estas expectativas.

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 (Bowman, Danzig, Hardy, Manber, & Schwartz, 1995) y ALIWEB (Koster, 1994) partían 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 basaba

1. en la indización en los servidores de origen,
2. la existencia de descripciones normalizadas de cada recurso,
3. una arquitectura distribuida de las bases de datos recopiladas y
4. 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" (Kahle *et al.*, 1991) 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 (MacMurdo, 1995). En ambos casos existía un esquema para la descripción de los recursos, aunque a veces se tratara de una simple línea (Mañas, 1994).

1.6.5.1 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 (Kahle, 1989). 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) (Hardy, Schwartz, & Wessels, 1996). El objeto de este formato, a modo de tabla atributo-valor, era doble: 1) unificar la representación de la información recopilada y 2) 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 empleaba 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 (Gravano, Chang, García Molina, Lagoze, & Paepcke, 1997).

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 fue revisado hace tiempo por Rachel Heery (Heery, 1996) que, sin embargo, ha concentrado su atención en otras iniciativas. Mucho más completa es la visión que, en su tesis doctoral y luego en la correspondiente monografía, aporta Eva Méndez (Méndez Rodríguez, 2002). Se debe tener en cuenta, sin embargo, que todos estos sistemas exigían el acuerdo previo entre proveedores y servicios de recuperación. Algo impensable (ver más abajo) 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. De entre las definiciones del concepto de metadatos que Sherry Vellucci discute (Vellucci, 1998), acaso la más simple sea la de mejor aplicación en el conexto del presente apartado:

> “Los Metadatos contienen información legible automáticamente para la web”. (World Wide Web Consortium, 2001).

Desde el punto de vista de la recuperación de información, más ilustrativo resulta el planteamiento de Dublin Core, la más extendida de las iniciativas sobre metadatos:

> "Puesto que Internet contiene más información de la que los autores de resúmenes, indizadores y catalogadores profesionales pueden tratar con los métodos existentes, se acordó como una alternativa razonable para obtener metadatos utilizables para recursos electrónicos proporcionar a los **autores y distribuidores de información medios para describir por sí mismos los recursos**. La tarea principal del Metadata Workshop fue identificar y definir un conjunto sencillo de elementos para la descripción de recursos electrónicos distribuidos. Para poder abordar esta tarea, se limitó en dos sentidos. Primero, sólo se tuvieron en cuenta aquellos elementos necesarios para el descubrimiento de cada recurso. Se consideró que el descubrimiento de recursos es la necesidad más acuciante y se debe satisfacer a despecho de la complicación temática o estructural de los objetos".
>
> (Weibel, 1995) (énfasis añadido).

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:

> "Una solución alternativa que promete mediar entre ambos extremos implica la creación de un registro más informativo que una simple entrada en un índice, pero menos completo que un registro catalográfico formal. Si sólo se nececitara un pequeño esfuerzo humano para crear tales registros, se podrían describir más recursos, especialmente si se pudiera animar al autor del documento a **crear la descripción**. Y si tal descripción siguiera un estándar establecido, sólo la redacción del registro requeriría intervención humana ; **herramientas automatizadas podrían detectar esas descripciones y recopilarlas**" (énfasis añadido).

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 (Massa, 2003).

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 (Weibel, Ianella, & Cathro, 1997). 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 (Dublin Core Metadata Initiative, 2003).

1.6.5.2 Ajuste entre el modelo de indización asistida y las iniciativas de metadatos

El énfasis anteriormente expresado 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 (Netscape Communications Corporation, 1997). 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 (World Wide Web Consortium, 1997), que fue presentado oficialmente en la quinta reunión de la Dublin Core Initiative una semana más tarde. En esta línea, la noticia más reciente es la adopción de este esquema como norma internacional (actualmente en preparación) ISO 15836 (Dekkers & Weibel, 2003).

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 (López & Massa, 1998). 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. 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 (Marchiori, 1998).

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. A pesar de ello, Las etiquetas meta han pasado de incorporarse al 45% de las páginas analizadas en 1998 al 70% en 2002 (del 70 al 82 por ciento en las portadas o "*home pages*"), con un promedio de etiquetas que ha aumentado desde 2,27 a 2,75 (O'Neill *et al.*, 2003).

Los sistemas de recopilación automática más popularizados han aceptado durante años etiquetas META en la indización de páginas Web de una forma irregular: Lycos y Northern Light no los empleaban en la representación de los resultados de búsqueda y sólo HotBot e Infoseek lo hacían en el cálculo de relevancia (Amat, 1999). De entre los sistemas españoles, Olé(Terra) mencionaba la posibilidad de buscar entre las palabras clave pero no hacía referencia explícita, al contrario que Trovator, al empleo de metadatos en la recopilación, la indización o el cálculo de relevancia. En Octubre de 2002, se anunció "la muerte" de la etiqueta "keywords": sólo Inktomi, el sistema mayorista de recopilación automática de documentos, continuaba rastreando los valores de este metadato. La razón esgrimida es su uso abusivo (como reclamo o "*spam*") por quienes desean obtener posiciones de preferencia en las listas de resultados y, así, atraer a sus páginas la mayor cantidad de conexiones (Sullivan, 2002).

Algunos trabajos ponen en cuestión la incorporación de metadatos a las páginas generadas en entornos académicos, comerciales u oficiales en función de su coste. 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 (Thomas & Griffin, 1999). La

experiencia, recientemente comunicada, de una modesta editorial electrónica resulta ilustrativa (Rhind-Tutt, 2003). A pesar de las cifras anteriormente aportadas, se constata el empleo de 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 (O'Neill *et al.*, 2003). Uno de los trabajos más recientes estimaba que sólo el 20% de las páginas Web en español emplean etiquetas meta (Craven, 2004).

1.6.5.3 La estructuración de documentos y la adición de significado

Desde el punto de vista de la recuperación de información, la iniciativa llamada “Web semántica” se puede considerar una continuación de otras iniciativas de metadatos. De hecho, aunque el concepto se popularizó tras la publicación de un trabajo divulgativo en Mayo de 2001 ya existía a finales de los 80 y aparece integrada en las actividades alrededor del Resource Description Framework. Su explicación es simple:

> “El Web se diseñó como un espacio informativo, con el objetivo de que fuera útil no sólo para la comunicación entre personas, sino también para que las máquinas pudieran participar y ayudar. Uno de los principales obstáculos de este objetivo ha sido el hecho de que la mayor parte de la información en la web se ha diseñado para consumo humano, e incluso si se extrae de estructuras de base de datos con significados bien definidos por sus tablas, esa estructura no es evidente para un robot que rastree la web. Dejando aparte el problema de la inteligencia artificial, de entrenar máquinas para que se comporten como humanos, el enfoque de la web semántico desarrolla lenguajes para expresar información de forma comprensible para el procesamiento automático” .
>
> (Berners-Lee, 1998)

Las etiquetas HTML están destinadas a que los procesadores y programas de visualización puedan representar el formato de los documentos. Las etiquetas XML y su cohorte de esquemas de datos (DTD, Document Type Definitions), reglas de descripción de recursos, vocabularios, sintaxis y otras especificaciones, pretenden destinarse a representar el contenido de los documentos:

> "El Web semántico dotará de estructura al contenido significativo de las páginas Web, creando un entorno donde los agentes de software que circulan de página a página puedan realizar tareas sofisticadas para los usuarios. Cuando un agente llegue (…) no sólo sabrá que la página tiene palabras clave…".
>
> (Berners-Lee, Hendler, & Lassila, 2001).

La idea, por tanto, se basa en asociar a cada documento un conjunto de elementos o marcas que conviertan las meras expresiones textuales en valores de unos atributos. El marcaje con XML, que "permite a cualquier usuario crear sus propias marcas-etiquetas ocultas similares a las marcas de las páginas Web o a las anotaciones de las secciones de las páginas de texto (...) añadiendo así una estructura arbitraria a sus documentos que, sin embargo, no dice nada acerca de lo que significan". El significado se añade a través del RDF, que "codifica conjuntos de atribuciones" e informa acerca de que "X es autor de de" por ejemplo (Berners-Lee *et al.*, 2001).

El panorama que esta Web de significados prefigura es el de una base de datos universal de documentos estructurados. Este panorama ideal ha generado cierto escepticismo entre la comunidad de la documentación. Pero ese escepticismo no significa desconfianza hacia el propio concepto de "Web semántica" , cuyos objetivos en relación con la representación y acceso al conocimiento se han calificado de "magníficos" (Codina, 2003). Más bien significa desconfianza en la capacidad de los productores y distribuidores de los documentos para la asignación de metadatos y conjuntos de etiquetas XML. Terrence Brooks emplea una perspectiva más amplia en su crítica. Abarca no sólo la dificultad de incorporar elementos estructurales y etiquetas de significado a los documentos del espacio Web. También se apoya en las limitaciones del empleo de las etiquetas meta (discutidas aquí en los párrafos anteriores). Además, ahonda en las diferencias entre la representación formal y de contenido:

> "Hemos heredado el concepto de documento de los sistemas de archivo vertical y de las bases de datos bibliográficas, dos entornos tecnológicos que separan contenido y representación... Contemplada desde la perspectiva del HTML, sin embargo, la indización de páginas Web confunde representación y contenido" .
>
> (Brooks, 2003).

Introduce, además, la característica volatilidad de los documentos del espacio Web contrastándola con la estabilidad de los documentos impresos y profundiza en su carácter de "instantáneas":

> "Visualizar lo que una página muestra en el navegador entre las etiquetas <HTML> y </HTML> refleja la forma en que el navegador dispone la fuente de bytes que llega del servidor, pero nada dice acerca de cómo se estructuró su contenido en el servidor de partida".

Se basa en diversas estimaciones sobre la proporción de documentos que se generan en respuesta a una consulta a bases de datos para ahondar más en la volatilidad de la información en ellos contenida. Concluye del siguiente modo:

> "...Las páginas Web no son buenos anfitriones para metadatos temáticos. Esto no es un juicio de valor sobre los metadatos en sí mismos, sino la mera observación de que no se aplican correctamente a una tecnología caracterizada por la mezcla de contenido en presentaciones arbitrarias recorridas por algoritmos desconocidos. El coste y esfuerzo de añadir metadatos temáticos a una estructura informativa sólo están justificados si tal estructura persiste en el tiempo con estructura, identidad y contenidos reconocibles".

No obstante, distingue el "closed web" constituído por intranets, bibliotecas digitales y otros espacios, donde admite la validez de procedimientos basados en indización asistida, porque representan entornos en que existe un acuerdo entre la asignación de metadatos y la estabilidad de las estructuras informativas. De hecho, hace equivaler los términos "web semántica" y "web cerrado".

Brookes no afronta la situación paradójica que se desprende de su propia línea argumental: si un alto porcentaje (cita estimaciones que lo aproximan al 75%) del contenido del espacio Web se genera de consultas a bases de datos, ¿ Cuál es la dificultad para generar un etiquetado semántico mediante la "traducción" de, por ejemplo, los nombres de campos de su esquema conceptual?. La experiencia de PubMed (http://www.pubmed.org) , que ofrece un formato de salida etiquetado en XML puede servir de ilustración a esta cuestión.

1.6.6 Sistemas basados en agentes (inteligentes)

Inicialmente, el descubrimiento y recuperación de recursos distribuidos 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 y distribuidores. Este esquema pas 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 distribuida. De forma más concreta: 1) 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; 2) 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; 3) 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; 4) los sistemas no siempre son accesibles; 5) 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; 6) los sistemas de recopilación automática no pueden seguir con el ritmo adecuado la dinámica y falta de estabilidad de los documentos y 7) 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.(Hermans, 1996; Hermans, 1997; Jansen, 1997). Baeza ha observado una coincidencia entre determinados aspectos del soft computing y la naturaleza imprecisa de la búsqueda y recuperación de información (Baeza-Yates, 2003).

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"* (Hermans, 1998).

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 (Wiederhold, 1992). 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.
> (Wiederhold & Genesereth, 1997).

El concepto de agente parece admitir dos acepciones, incluso en el seno de la propia comunidad dedicada a la inteligencia artificial. Ambas nociones, más intuitiva la primera, más formal la segunda, se ofrecen en una extensa revisión, que es una de las aportaciones más ampliamente difundidas en este sector de la literatura:

> “Quizá el uso más generalizado del término agente denota un sistema automático basado en hardware o, más usualmente, en software que goza de las siguientes propiedades: autonomía, capacidad social, capacidad de reacción y otras...”

> Para algunos investigadores (…) el término ‘agente’ tiene un significado más riguroso y específico (...) Estos investigadores generalmente llaman agente a un sistema informático que, además de las propiedades enumeradas antes, se conceptualiza o implanta empleando conceptos que se aplican habitualmente a las personas...”.
>
> (Wooldridge & Jennings, 1995).

La noción “laxa” de los agentes se ha empleado acaso con demasiada laxitud en el entorno de la documentación. 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”.
>
> (Hípola & Vargas Quesada, 1999)

Tanto estos autores como Tramullas y Olvera (Tramullas Saz & Olvera Lobo, 2001) ofrecen en sus trabajos peromenorizadas descripciones de lo que los propios fabricantes denominan “*Desktop MetaSearch Utility”*. Su esquema de funcionamiento es simple: tras la instalación en el ordenador de cada usuario, estos programas conectan con sistemas de recuperación, bases de datos y sedes, en los que realizan búsquedas según el perfil especificado. Posteriormente compactan los resultados procedentes de las diversas fuentes y los presentan al usuario de forma ordenada.

Es posible que este mecanismo general de funcionamiento se aproxime a algunos de los requisitos especificados por Woolbridge y Jennings. Sin embargo, es difícil que cumplan las condiciones adicionales que se atribuyen a los agentes inteligentes: habilidad (competence) y fiabilidad (trust) (Maes, 2003). De hecho, los asistentes de sobremesa, programados mediante un conjunto de preferencias por el usuario final, no cumplen ninguna de las condiciones:

> “...Este enfoque requiere demasiada implicación, demasiado conocimiento y esfuerzo del usuario final, puesto que este usuario debe reconocer la oportunidad para emplear un agente, conferirle conocimiento explícito (especificándolo por tanto en algún lenguage formal) y mantener

sus reglas a medida que pasa el tiempo (a medida que los hábitos de trabajo o los intereses cambian)".

(Maes, 2003).

A pesar de la falta de una definición exacta del término, la esencia de los agentes reside en su capacidad de aprendizaje, base de su adaptabilidad (Hendler, 1999). Maes discute las ventajas del aprendizaje automático (*machine learning*) frente al enfoque basado en el conocimiento (*knowledge based approach*) del dominio de referencia y del usuario. En su trabajo, tras presentar 4 tipos de agentes operativos en el laboratorio de medios del MIT, insiste en la capacidad de aprendizaje como característica esencial de los agentes que "...aprenden de forma gradual la mejor forma de asistir al usuario mediante: la observación e imitación del usuario, la interpretación de las respuestas positivas y negativas del mismo, la recepción de instrucciones explícitas y la obtención de ayuda de otros agentes".

Parecida terminología emplea una de las más amplias panorámicas sobre los sistemas de agentes disponibles (Mladenic, 1999), que distingue los basados en contenido, que emplean el aprendizaje textual (*text learning*) de aquellos otros que, basándose en la colaboración con otros agentes, recurren al aprendizaje social (*social learning*). Esta panorámica identifica 13 sistemas textuales, 5 colaborativos y 3 mixtos, antes de presentar los resultados del proyecto WebWatcher de la Carnegie Mellon University. De la misma época son dos proyectos directamente relacionados con la recuperación de información distribuida en el espacio Web. El proyecto TétraFusion, desarrollado conjuntamente por instituciones irlandesa y francesa, almacena su base de conocimientos mediante la extracción de datos (data mining) de páginas web seleccionadas (Crimmins, Smeaton, Dkaki, & Mothe, 1999) a las que somete a una taxonomía compleja. Por su parte, el sistema Retriever, de la Universidad de Patras, procesa el análisis de los resultados relevantes para obtener un conocimiento del "dominio de interés" (*query domain*) y reacciona incorporando a su base de conocimientos los datos resultantes (Fragoudis & Likothanassis, 1999).

En España y en documentación, sólo el trabajo de Berrocal se orienta hacia el empleo de agentes para la recuperación de información en Internet (Berrocal, Figuerola, Zazo, & Rodríguez, 2003). Desgraciadamente, presenta las principales nociones a nivel meramente introductorio. Otros grupos españoles están desarrollando trabajos originales con cierto éxito (Castillo Sobrino, Serrano Moreno, & Sesmero Llorente, 2003). Por otra parte, la revisión del Laboratorio de Investigación sobre Agentes de la Universidad de Gerona ofrece, de momento, la más amplia clasificación de los agentes asesores (*recommender agents*) empleados en la recuperación en el espacio Web y destinados a la obtención de modelos sobre los usuarios y sus preferencias, la elaboración de modelos de contenido y la construcción de patrones sociales (Montaner, López, & Rosa, 2003).

Casi resulta ocioso añadir que los sistemas de agentes pueden reconocer una "federación de contenidos" distribuidos en el espacio Web e interoperar a través de un protocolo común o de la conversión de protocolos para obtener información de los espacios participantes (Sanchez, Sandra, Fernández, & Chevalier, 2002).

James Hendler, desde la Information Systems Office de DARPA, ha vislumbrado un panorama en que se combinan la generalización del concepto de Web semántica y el empleo de agentes: "En los próximos años, la práctica totalidad de empresas, universidades, organismos gubernamentales y grupos especializados deberán dotar a sus contenidos de ontología para que sean accesibles por las poderosas herramientas preparadas al efecto. La información se intercambiará entre aplicaciones, permitiendo a los programas recopilar y procesar contenidos Web e intercambiar información de forma gratuita. En la cima de esta infraestructura, la automatización basada en agentes será mucho más práctica" (Hendler, 2001). Y prosigue: "Para que esta visión pueda hacerse realidad, debe ocurrir un fenómeno similar al de los primeros días de la web. Los usuarios no marcarán sus páginas a menos que perciban el valor de hacerlo y no se desarrollarán herramientas para hacerles ver ese valor hasta que no se

marquen los recursos Web". Aún antes de discutir los pasos a seguir para resolver este problema "del huevo y la gallina" (sic) se había mostrado lo suficientemente optimista para dirigirse a la comunidad científica de forma rotunda:

> "En pocas palabras, si usted aún no emplea una tecnología basada en agentes, no se preocupe, pronto la utilizará".
>
> (Hendler, 1999)

1.7 Los estudios sobre la búsqueda de información y la evaluación de los sistemas

Todos los observatorios independientes coinciden en resaltar la gran popularidad de los sistemas de recuperación de información distribuida en Internet[2]. En especial, los informes periódicos del Graphic, Visualization and Usability Center muestran en sus más recientes entregas que casi el 80% de los usuarios encuestados localizan páginas Web a través de sistemas de búsqueda y más del 40% a través de directorios temáticos. Estos porcentajes aumentan sustancialmente con la mayor experiencia de los usuarios (GVU's WWW Surveying Team, 1998).

Por otra parte, todos los estudios de evaluación coinciden en el bajo nivel de rendimiento de estos mismos sistemas. Así lo reconocen tres de las más recientes revisiones sobre su evaluación (Oppenheim, Morris, Mcknight, & Lowley, 2000; Jansen & Pooch, 2001; Rasmussen, 2003).

Los resultados de los trabajos sobre satisfacción de los usuarios muestran resultados dispares. Así, Bruce concluye "*users (…) are satisfied with the process regardless of how frequently they use the network or whether or not they have formal training*" (Bruce, 1998). De hecho, estima el grado de satisfacción sobre la búsqueda de información en Internet en un intervalo de 4,0417 a 4,2420 en una escala de 1 a 6. Raya Fidel

[2] Por ejemplo http://websearch.about.com/ o http://www.searchenginewatch.com/

y colaboradores, por su parte, hallaron que " [users] *were satisfied with their searches and the results, but impatient with slow response*" (Fidel *et al.*, 1999). En el primer estudio, los usuarios eran personal académico de una universidad tecnológica australiana; en el segundo, alumnos universitarios estadounidenses. Podría pensarse, a la vista de estas conclusiones, que el nivel educativo o académico es una garantía de éxito en la recuperación de información en Internet. Sin embargo, otro estudio destaca "*Among other findings, the study revealed that searching the World Wide Web (WWW) is not without difficulty*" y su escuesta fue remitida a 1000 miembros de la comunidad universitaria de Amsterdam (Voorbij, 1999). El enfoque holístico del grupo de Carol Tenopir, que emplea estudiantes de documentación y analiza factores cognitivos y afectivos puestos en juego en el proceso de búsqueda, tampoco ofrece resultados alentadores (Wang *et al.*, 2000).

Los estudios sobre usuarios "reales", aquellos que no integran grupos académicos, se han basado en el análisis de transacciones (log analysis) es decir, en el seguimiento de sus acciones tras la conexión con uno u otro sistema de búsqueda: la expresión de perfiles, el tema de las búsquedas, la cantidad de resultados visualizada, el seguimiento de alguno de ellos y otras variables han sido objeto de análisis pormenorizados en 2 amplias series de estudios. En primer lugar, investigadores de Google y Compaq analizaron 285 millones de sesiones de búsqueda planteadas durante 6 meses en el sistema AltaVista. En segundo lugar, un grupo colaborativo ha realizado toda una serie de estudios analizando las búsquedas realizadas en el sistema Excite.

El primer grupo concluye su análisis de forma rotunda:

> "Nuestros datos apoyan la conjetura de que los usuarios de la web se diferencian significativamente del usuario que se considera habitual en la literatura en documentación. Específicamente, mostramos que los usuarios de la web plantean peticiones más cortas, la mayoría sólo examinan los 10 primeros resultados y raramente modifican su perfil. Estos sugiere que las técnicas tradicionales de recuperación podrían funcionar mal para dar respuesta a las peticiones de búsqueda en la web".

(Silverstein, Henzinger, Marais, Moricz, & M, 1999).

El material de análisis de Amanda Spink y colaboradores consiste en tres conjuntos de datos que incluyen más de un millón de peticiones formuladas por unos de 211.000 usuarios del sistema Excite en plazos mensuales de 1997, 1999 y 2001. El estudio longitudinal muestra ciertas variaciones en la temática de las búsquedas, aunque no en la longitud (número de términos) de los perfiles ni en la frecuencia de búsquedas por usuario (Wolfram et al., 2001; Spink, Jansen, Wolfram, & Saracevic, 2002). Las búsquedas de usuarios estadounidenses en Excite y europeos (sobre todo alemanes) en FAST muestra algunas diferencias temáticas y de conducta (Spink, Ozmutlu, Ozmutlu, & Jansen, 2002). En términos generales, se afirma que

> "la mayoría de la gente usa pocos términos de búsqueda, modifica poco sus perfiles, visualiza pocas de las páginas resultantes y raramente usa la modalidad de búsqueda avanzada. Se utiliza un pequeño número de términos con mucha frecuencia y muchos de los términos se emplean sólo una vez" .
>
> (Spink, Wolfram, Jansen, & Saracevic, 2001)

En el estudio más reciente de esta serie, centrado en el análisis de un millón de búsquedas que 220.000 usuarios europeos en Febrero de 2001 y Mayo de 2002 (Jansen & Spink, 2004) se obtienen resultados coherentes con los anteriores estudios:"Se apreció una disminución en la extensión de los perfiles, con expresiones extremadamente simples. Los temas de búsqueda de los europeos se amplían, con un notable declive porcentual en búsquedas de sexo y pornografía. La mayoría de los usuarios visualizan menos de cinco documentos Web, deteniéndose sólo unos segundos en cada. Un 50 % aproximadamente de los documentos eran temáticamente relevantes ".

Que el 22% de los usuarios de los estudios anteriores modificaran su primera formulación de búsqueda, que la mitad de los usuarios que emplearon operadores booleanos lo hicieran de forma errónea, al igual que la tercera parte de

quienes emplearon modificadores, que sólo el 5% de los usuarios utilizaran feedback de relevancia o que menos del 1% formulara frases en sus perfiles... son hallazgos que permiten, entre otras, conclusiones próximas a las del estudio de AltaVista:

> "Mientras los sistemas de recuperación en la Web siguen los principios básicos de la recuperación de información, los usuarios parecen diferir significativamente en los sistemas tradicionales, como los del sistema DIALOG o los que se supone en los experimentos TREC. Es recuperación de información, pero una recuperación muy diferente. Los usuarios de la web no se sienten precisamente cómodos con los operadores booleanos y otros procedimientos de búsqueda avanzados. Ciertamente, no visualizan los resultados más allá de la primera página. Estos hechos en sí mismos ponen énfasis en la necesidad de enfocar el diseño de los sistemas de recuperación en la web, e incluso el diseño de las sedes Web, es una forma significativamente diferente a la que se ha adoptado hasta el momento".
>
> (Jansen, Spink, & Saracevic, 2002).

Y no deben resultar extrañas si se tienen en cuenta las dificultades, de índole más general, que el público experimenta en su obtención de información a partir de fuentes distribuidas: tecnológicas, económicas, geográficas, de limitaciones personales, cognitivas y psicológicas (Pettigrew, Durrance, & Unruh, 2002). La principal dificultad, sin embargo, se relaciona con la búsqueda y obtención de información. El trabajo de Karen Pettigrew y colaboradores, que toma como base la experiencia de los usuarios y profesionales de tres bibliotecas públicas y redes ciudadanas durante dos años identifica los siguientes problemas:

> "**Recuperación deficiente, sobrecarga informativa y baja precisión**: a causa de la deficiente indización de los sistemas, los usuarios recuperaron demasiada información comunitaria y tuvieron que discernir qué información resultaba relevante a sus búsquedas;
>
> **Diseño de la interface mejorable**: Los usuarios se desalentaron con frecuencia por el aspecto de determinadas sedes: demasiado ocupada, con demasiados resortes, mal juego

de caracteres (sobre todo para quienes no apreciaban bien los colores) y demasiado texto presentado en una única página;

Información mal organizada (clasificada): los usuarios no encontraban la información comunitaria donde esperaban y había muy pocos reenvíos;

Información anticuada e inapropiada: o la información comunitaria estaba pasada o no había modo de distinguir la fecha de creación o de más reciente actualización de una página. También se observaron contenidos inapropiados*;*

Falta de autoridad: sin identificación apropiada de las credenciales de un autor o su filiación, los usuarios experimentaron dificultades en valorar la calidad de un recurso, es decir, su credibilidad;

Pérdidas: en ocasiones, los usuarios comentaron que la información había desaparecido, aunque se anunciara al inicio de la correspondiente página*;*

Enlaces inactivos: los usuarios experimentaron gran frustración cuando se encontraron con que los enlaces a una página aparentemente de gran relevancia, eran inactivos o devolvían errores;

Dificultades lingüísticas: Dejando aparte que la mayoría de la información sólo aparecía en inglés, algunas sedes contenían información escrita en argot o a un nivel de difícil comprensión;

Confidencialidad: los usuarios deseaban garantías de que la información que enviaban y recuperaban fuera confidencial;

Proximidad: los usuarios deseaban información sobre su entorno más próximo y las personalidades más cercanas;

Pasividad de los sistemas: los usuarios indicaron que su búsqueda de información se vería muy facilitada si los sistemas de información comunitaria fueran capaces de anticipar sus siguientes necesidades informativas (basándose en las realizadas anteriormente). Con demasiada frecuencia, los usuarios describieron que las sedes halladas no eran las deseadas, pero no habían sabido qué pasos tomar a continuación."

(Pettigrew *et al.*, 2002).

Un estudio de campo realizado desde el punto de vista antropológico con "usuarios experimentados" de 4 áreas urbanas resulta en hallazgos altamente coincidentes, aunque tuviera por objetivo principal investigar el efecto sobre los usuarios del pago por posición:

"La mayoría de los participantes tenían poco conocimientos sobre la forma en que los sistemas recuperan información de la web o cómo ordenan los enlaces en las páginas de resultados…

> La mayoría no examinó los resultados más allá de la primera página porque tenían una confianza ciega en que los sistemas presentaban sólo los mejores o más apropiados resultados en esa página".
>
> (Marable, 2003).

En los mismos términos se puede describir la experiencia en recuperación de usuarios "domésticos" (Rieh, 2004).

Los sistemas de recuperación de la información distribuida en Internet son populares, imprescindibles o ambas cosas. Su manejo, sin embargo, es dificultoso y sus resultados poco satisfactorios. Esta situación contradictoria parece requerir la realización de trabajos de evaluación de los sistemas. Sin embargo, es común reconocer las limitaciones de los estudios hasta ahora desarrollados:

> "... no se han desarrollado investigaciones coherentes que evalúen los sistemas de búsqueda y, en consecuencia, no es posible comparar el rendimiento de los sistemas que cada investigador comunica".
>
> (Oppenheim *et al.*, 2000).

Bernard Jansen ha expuesto la situación de forma más sucinta: "El análisis de los limitados estudios disponibles indica que tanto los métodos experimentales como la terminología aún divergen" (Jansen *et al.*, 2001).

Ambos trabajos realizan una clasificación de los trabajos de evaluación realizados. Es curioso reconocer que también sus criterios clasificatorios son "divergentes". De hecho, Oppenheim parte de reconocer como "paradigma para la evaluación de la eficacia y el rendimiento de los sistemas de recuperación de la información el modelo Cranfield". En consecuencia, su tipología se ciñe a las variantes que, sobre todo en el cálculo de la exhaustividad de la recuperación, presentan los trabajos revisados. Jansen se muestra más limitado e incorpora a su clasificación tres grandes grupos de estudios: los basados en el comportamiento de los usuarios que buscan información en Internet, llamados primarios;

aquellos otros que analizan las búsquedas realizadas en sedes únicas o múltiples y los estudios sobre búsqueda de información a través de OPAC y de sistemas de recuperación tradicionales. A grandes rasgos, se puede afirmar que la revisión de Oppenheim se centra en los sistemas, mientras que la de Jansen se centra en los usuarios. Aunque ambos puntos de vista no son equivalentes, su objetivo general es común: el desarrollo de un marco para normalizar la evaluación de los sistemas de recuperación en Internet.

La propuesta de Jansen se centra en la exigencia de tres secciones en la comunicación de resultados: una sección descriptiva, un análisis de la presentación y un análisis estadístico. El hecho de que los estudios de búsqueda en Internet cuenten con esas secciones puede contribuir a comparar unos y otros. Hasta cierto punto es una propuesta formal. En contraste, Oppenheim se centra en la evaluación del rendimiento de los sistemas, revisa los criterios empleados y concluye con un conjunto de recomendaciones para su empleo en pruebas de evaluación que incluye las siguientes:

> “precisión;
> exhaustividad relativa;
> velocidad de respuesta;
> coherencia de los resultados a lo largo de un periodo extenso;
> proporción de enlaces fallidos;
> proporción de resultados duplicados;
> calidad general de los resultados juzgada por los usuarios;
> evaluación de la familiaridad de la interface para los usuarios;
> utlidad del sistema de ayuda y variaciones del programa para usuarios de diversos niveles;
> opciones de presentación de los resultados;
> presencia de anuncios;
> cobertura;
> extensión de los perfiles;
> extensión y legibilidad de los resúmenes;
> eficacia de los sistemas.”

Más modesta resulta la lista de objetivos de Martínez Méndez:

> "1. Cálculo de una medida de valor simple basada en la exhaustividad y en la precisión adaptadas al contexto de la World Wide Web
> 2. Tipificación de esta medida de valor simple con respecto al volumen de la respuesta que se le entrega al usuario
> 3. Determinación de los porcentajes de enlaces duplicados y de la frecuencia de actualización de los índices de los motores a partir del porcentaje de errores en las referencias ofrecidas
> 4. Establecimiento de la similitud existente entre los índices de los motores en función del contenido y del alineamiento de su respuesta
> 5. Identificación de posibles agrupamientos entre los distintos motores con base en su similitud".
>
> (Martínez Méndez, 2001).

Desde las revisiones de Oppenheim y Jansen se han producido aportaciones de interés al campo de la evaluación de estos sistemas. Su orientación es similar a la de los trabajos mencionados: la búsqueda de un modelo de evaluación objetivo y, en consecuencia, generalizable. Pia Borlund ahonda en la búsqueda de un procedimiento experimental que se aparte del modelo Cranfield. Este modelo, en su opinión, no se ajusta a las condiciones en que los usuarios interaccionan con los sistemas de recuperación en Internet (Borlund, 2003). Se debe tener en cuenta que el trabajo de Borlund se situa en una corriente de investigación, la de la recuperación de información en contexto (*information retrieval in context*) cuyos máximos exponentes comparten su filiación: la Royal School of Library and Information Science danesa. Por otra parte y en la línea de Cranfield y la serie de experimentos TREC, se ha elaborado colecciones de referencia aplicables a la evaluación de la recuperación en la web (Bailey, Craswell, & Hawking, 2003) después de que el mismo grupo realizara un amplio trabajo de evaluación en este contexto y en esta línea (Hawking, Craswell, Bailey, & Griffiths, 2001). También se ha tratado de buscar un equilibrio entre la aplicación de métodos tradicionales de evaluación y el análisis de los sistemas interactivos:

> "... Así, mientras la comunidad de RI está ocupada tratando de redefinir y reevaluar la necesidad de criterios e indicadores, la comunidad Web comienza una búsqueda algo caótica de parámetros para evaluar el rendimiento de varios sistemas de recuperación. Se diría que ambas

comunidades tienen mucho en común y que podrían obtener ventaja de algún tipo de colaboración".

(Landoni & Bell, 2000).

Este intento de conciliación contiene la propuesta de una compleja batería de indicadores y directrices que, partiendo de un grupo de cinco grandes objetivos, se desglosa hasta alcanzar cuestiones de gran detalle. Por ejemplo, la operación "revisión de criterios y de los indicadores aplicables" desciende hasta obligar al analista a decidir sobre la puntuación a asignar a los resultados repetidos o a los resultados indirectos (aquellas páginas juzgadas no relevantes que, sin embargo, contienen enlaces a documentos que sí lo son).

La más reciente revisión divide los trabajos de evaluación en dos épocas. La primera, que abarca los años 1995 y 1996 se caracteriza por la diversidad (divergencia) metodológica ya mencionada y por la falta sistemática de juicios de relevancia a cargo de los usuarios finales. En el periodo más reciente (entre 1997 y 2000) se han realizado experimentos controlados, otros con usuarios "reales", estudios "naturalistas"y encuestas (Su, 2003). La propia Louise Su formula un modelo de evaluación orientado hacia los usuarios que, a continuación, aplica a un grupo de estudiantes universitarios (Su, 2003). Su propuesta incluye dos tipos de variables, directa e indirectamente relacionadas con el rendimiento y un conjunto de procedimientos. Entre las primeras se incluyen medidas de relevancia, de eficacia, de utilidad, de satisfacción de los usuarios con los resultados de búsqueda y de conectividad. La segunda serie de variables trata de determinar el estado de conocimiento y las habilidades previos de los participantes, sus necesidades informativas y sus requisitos. Su protocolo o procedimiento se articula en dos sesiones:

"Sesión 1:

Complete una autorización de participación y un cuestionario sobre la experiencia previa de cada usuario;
Consiga información sobre la necesidad informativa del usuario en un formulario;
Reciba una sesión instructiva;

Realice las búsquedas sobre el propio problema informativo del usuario empleando sistemas seleccionados;

Use instrucciones online pulsando en INFO, TIPS, FAQs o HELP;

Obtenga una versión impresa de los resultados de búsqueda de cada sistema de una extensión dada;

Cumplimente un cuestionario sobre la satisfacción de los usuarios con las características e interacción de cada sistema en cuanto el usuario completa la búsqueda en cada sistema;

Elija un periodo propicio para las tareas de la segunda sesión.

Sesión 2:

Juzgue la relevancia de los resultados según unas directrices preestablecidas;

Seleccione y ordene los items resultantes;

Intervenga en una entrevista posterior a las búsquedas para recabar las reacciones al proceso de búsqueda, el producto y el rendimiento global de un sistema de recuperación" .

(Su, 2003).

Junto con el trabajo de Gordon (Gordon & Pathak, 1999) y a pesar de las diferencias entre ambos (véase la sección 5.1) se trata del intento más firme de fijar una metodología de evaluación. Los trabajos más recientes (Can F, Nuray, & Sevdik, 2004; Vaughan, 2004) están abundando en esta línea.

2. Objetivos y plan de trabajo

Las razones que justifican el análisis de los sistemas de recuperación de información distribuída en Internet son variadas y de gran peso. En el editorial de presentación del número del ASIS Bulletin dedicado a la evolución de la recuperación de textos, Irene Travis hace énfasis en los motivos que apoyan el análisis de esos sistemas (Travis, 1998). De su enfoque, y de otros (Bellardo Hahn, 1998) incluidos en el mismo monográfico, se pueden enumerar las siguientes:

1. Los Sistemas de Recuperación de Información distribuída en Internet (SIRI) ofrecen la oportunidad de estudiar las posibilidades de consulta en texto libre y recuperación de ingentes colecciones de documentos (electrónicos), una situación que parece avecinarse en el entorno de la información ligada al conocimiento y al que no han de ser ajenas las iniciativas de “colecciones digitales” ahora incipientes.
2. El acceso de usuarios no cualificados sin intermediación alguna, los “casual users” frente a “end users”, brinda la oportunidad de estudio de sus hábitos de consulta y recuperación y la extracción de conclusiones de extrema importancia práctica.
3. Las características de los documentos y textos a recuperar (inestabilidad, desaforado crecimiento, heterogeneidad máxima) plantean una serie de desafíos especialmente interesantes al diseño de los sistemas, a sus mecanismos de actualización y a su arquitectura de almacenamiento.
4. Todas estas motivaciones multiplican su valor cuando se plantea el establecimiento de mecanismos de recuperación de información distribuída en entornos académicos, de investigación o, en general, ligados al conocimiento. Esto incluye la selección o el diseño de sistemas de recuperación de los contenidos en intranets y redes corporativas

especializadas, en el marco de lo que se ha dado en llamar *Enterprise Content Management*.

Varios autores han tratado el tema en la literatura española. López Alonso y Mares Marín presentaron un trabajo original donde evaluaban el rendimiento de los sistemas Excite, Lycos, OpenText y Savvysearch (López Alonso y Mares Marín, 1996). Sánchez Montero ha tratado el problema en el marco de su trabajo sobre diseño de contenidos para intranets corporativas (Sánchez Montero, 1997). Más recientemente, Baró aportó una relación, con descripciones muy genéricas de algunos sistemas de búsqueda en el entorno World Wide Web (Baró i Queralt, 1997). La revisión de Senso (Senso, 1998), y la aportación al tema de Marcos Mora resultan más descriptivas de los servicios asociados que de los propios sistemas (Marcos Mora, 1998). Otro trabajo destaca por el número de sistemas analizados y por incluir entre ellos, por vez primera, a 10 sistemas nacionales (Maldonado Martínez y Fernández Sánchez, 1998). En él, las autoras presentaron un estudio pormenorizado de las “características documentales” de 10 SIRI internacionales y 10 españoles. Por características documentales las autoras entendían el estudio del esquema de datos, las posibilidades de recuperación y la presentación de resultados de cada sistema. Como resultado, emitían un juicio sobre la operatividad de todos y un análisis global de ambos grupos.

Los trabajos españoles de tesis no han estudiado los sistemas españoles de recuperación distribuída en Internet. Olvera se centra en 10 servicios de búsqueda internacionales: Altavista, Excite, Hotbot, Infoseek, Lycos, Magellan, OpenText, WebCrawler, WWWWorm y Yahoo (Olvera Lobo, 1999). Por su parte, Martínez emplea AltaVista, AlltheWeb (FAST), Google, MSN, Terra y WISEnut (Martínez Méndez, 2001). En ninguno de los dos casos existe formulación de peticiones y juicio de relevancia por parte de los usuarios. Alonso Berrocal, por su parte, no realiza experimento de recuperación alguno (Alonso

Berrocal, 2000). Y, sin embargo, “De entre los recursos de Internet, el tipo más usado por las pequeñas empresas españolas (94%) corresponde a los buscadores” (PricewaterhouseCoopers, 2004).

Si a las motivaciones antes enumeradas se añade otra, basada en el valor de la información relevante y la consideración “patrimonialista” de los recursos distribuidos en un medio social y cultural específico (un argumento ampliamente esgrimido en lo tocante a la producción y el consumo de información científica “nacional”) se reconocerá el interés del estudio de los SIRI desarrollados en España. Para llevarlo a cabo, es rigurosamente necesario analizar los tres componentes principales de cualquier sistema de recuperación de información: 1) los documentos; 2) las características operativas y el esquema conceptual de los sistemas y, naturalmente, 3) la interacción usuario-sistema y la satisfacción de las demandas planteadas.

Por estudio de los documentos ha de entenderse una profundización en el manido comentario (camino de convertirse en frase hecha) que une heterogeneidad, ingente volumen y falta de estabilidad como características del espacio Web. El análisis del esquema conceptual y la operatividad de los sistemas exige la comparación entre estos sistemas y otros, tradicionales o no. La línea de los usuarios y la satisfacción de sus demandas se ha de plantear teniendo en cuenta no sólo la mecánica de ajuste entre perfiles y documentos. La cultura, el estado de conocimiento previo y otros factores se han de tener en cuenta en este tercer componente de la evaluación.

El presente trabajo se propone el análisis del rendimiento de los sistemas españoles de recuperación de información en Internet. Previamente, incorpora al estudio global la determinación de las características del espacio Web español y, en la línea iniciada por Maldonado y Martínez, un análisis descriptivo de los sistemas españoles y los recursos que emplean para la cobertura, la indización, la representación, la recuperación y la ordenación de resultados en respuesta a las

búsquedas. El esquema propuesto se ajusta al menos a dos de los grandes grupos de trabajos identificados por Francisco Javier Martínez y José Vicente Rodríguez: existe un análisis de las características formales o externas de los sistemas y también se realiza un ensayo de su rendimiento (Martínez Méndez y Rodríguez Múñoz, 2003). Los tres objetivos generales se estructuran en operaciones concretas.

En una primera fase, la caracterización de espacio Web español parte de la extracción de una muestra de sedes y la investigación de su composición cualitativa y cuantitativa. La segunda fase abarca el estudio de los sistemas de recuperación; parte de una selección e investiga sus métodos de recopilación, su esquema de datos y sus mecanismos de indización. Por último, el estudio del rendimiento emplea a usuarios reales que expresan sus solicitudes de información y, posteriormente, juzgan la relevancia de los resultados obtenidos en los procesos de búsqueda, permitiendo el cálculo de exhaustividad y precisión de cada sistema. Los datos pueden ponerse en relación con la segunda fase del estudio.

3. Caracterización del espacio web en España

El diseño, la implantación y el mantenimiento de cualquier sistema de información ha de tener en cuenta las características de los documentos que incorpora. Mucho más si, como en el caso de los sistemas de recuperación del espacio Web, se ofrece la recuperación de documentos íntegros a través de la indización exhaustiva de su contenido. El volumen, el número de elementos o la jerarquía de páginas de una sede son factores con alguna incidencia en su accesibilidad a través de los sistemas de recuperación de información distribuida en Internet. Otros, como el número de enlaces a otras sedes, la abundancia de textos en soportes indizables o la homogeneidad estructural de las páginas tienen una importancia crucial. Aunque el concepto de "caracterización" admite un triple significado, se aplica aquí al estudio del contenido de los documentos distribuidos en el espacio web. En este sentido, se presenta un análisis cuantitativo de las características de una muestra aleatoria de sedes Web encuadradas en el dominio .es. Además del estudio de su accesibilidad directa, las variables analizadas incluyen tamaño, estructura jerárquica, número y tipo de los elementos incluidos en las páginas, grado de conexión a través de enlaces y grado de interactividad mediante funciones. El modelo general resultante corresponde a una distribución exponencial de la práctica totalidad de las variables, lo que permite definir 3 grupos de sedes. Se analizan las implicaciones

de estos hallazgos para los sistemas de recuperación de información distribuida en Internet y se propone que el estudio dinámico de esta u otras muestras permitirá seguir la evolución de las sedes Web españolas y mejorar su accesibilidad a través de los buscadores.

3.1 El concepto de caracterización y sus modalidades

Los términos *Web demographics, Web statistics, Web metrics, Web characterization, Cybermetrics* e incluso *Web archaeology* se han empleado desde hace tiempo para referirse a la variada gama de proyectos y trabajos de análisis cuantitativo de ese componente de Internet. Junto a los recuentos estadísticos realizados periódicamente por organismos reguladores, cabe encuadrar en esta área las investigaciones cuantitativas dirigidas a la obtención de hallazgos sobre el volumen de este espacio informativo, su dinámica, su estructura y sus características. A grandes rasgos, se han identificado inicialmente tres líneas de investigación (Woodruff, Aoki, Brewer, Gauthier, & Rowe, 1996):

1) el análisis del rendimiento y de los flujos de información y datos a nivel global;
2) el análisis del contenido de los documentos distribuidos en el espacio Web; y
3) el estudio de la interrelación entre esos mismos documentos y sus usuarios.

Las investigaciones encuadradas en el primer grupo, de los que son buen ejemplo los trabajos de KC Claffy para el National Laboratory for Applied Network Research (Monk & Claffy, 2002) y la Cooperative Association for Internet Data Analysis (Claffy, 2000), se dirigen al estudio de los modelos de flujo informativo, siendo sus resultados aprovechables en la mejora de infraestructuras y de los sistemas de medición de audiencias. Los análisis estructurales de la web considerada globalmente, que también podrían encuadrarse en esta línea, ofrecen además resultados directamente relacionados con la accesibilidad de la información en la Web. Es el caso del trabajo de Andrei Broder y colaboradores que, entre su lista de motivaciones, incluye “el diseño de estrategias de recopilación en la

Web" y "la comprensión sociológica de la creación de contenidos en la Web" (Broder, 2000). Otro tanto ocurre con trabajos, también eminentemente numéricos, que comienzan a ofrecer interesantes modelos predictivos (Adamic & Huberman, 2001).

Buenos exponentes de los análisis sobre la interacción entre documentos y usuarios en el espacio Web son, por una parte, los trabajos de James Pitkow y su grupo, iniciados en el Instituto Tecnológico de Georgia y continuados en Xerox, cuyo objetivo último pasa por dilucidar las propiedades de la Web como "sistema ecológico del conocimiento" (Huberman, Pirolli, Pitkow, & Lukose, 1998). Por otra parte, los ingentes análisis del grupo de investigación de Digital sobre las transacciones de Altavista (Silverstein, Henzinger, Marais, Moricz, & M, 1999), de IBM sobre las de WebCrawler (Zien, Meyer, Tomlin, & Liu, 2003) y de Jansen, Spink y Saracevic sobre Excite (Jansen, Spink, & Saracevic, 2002) concretan su análisis en las operaciones de recuperación de información de los usuarios. En fechas muy recientes, se han empezado a realizar estudios sobre la accesibilidad de las sedes para colectivos desfavorecidos (usuarios con algún tipo de minusvalía). Estos estudios emplean también análisis cuantitativos de determinados indicadores y elementos de la sintaxis de las páginas. Uno de los más recientes examina el ajuste a las recomendaciones técnicas del Consorcio Web de las sedes de las universidades españolas (Térmens Graells, Ribera Turró, & Sulé Duesa, 2003).

Uno de los más antiguos ejemplos de estudio del contenido de los documentos Web está contenido en la comunicación de Tim Bray en la quinta Conferencia Internacional sobre la Web (Bray, 1996). Su trabajo no sólo presenta una descripción estadística de variables, como el tamaño de las páginas o su formato, sino que, además el recuento de los enlaces permite identificar las "mayores y más visibles sedes", dotando de un componente estructural a su análisis. En la misma conferencia Woodruff y colaboradores emplearon una muestra mucho mayor que la de Bray para analizar un completo conjunto de variables, que abarcaban desde el tamaño de los documentos al número de errores de código de las páginas (Woodruff *et al.*, 1996). Desde estos trabajos, dos elementos han animado la línea de caracterización cuantitativa del contenido de los documentos Web: el refinamiento metodológico y, especialmente, la

puesta en relación de los hallazgos con la accesibilidad de los documentos en ese espacio.

Los estudios iniciales tomaban como muestra los documentos recopilados por sistemas de recuperación (OpenText, Inktomi) con una innegable intención promocional. Pero, a finales de los años 90, el grupo de investigación de los laboratorios NEC publicó unos resultados demoledores sobre la cobertura y el nivel de actualización de los mayores sistemas de búsqueda en Internet (Lawrence & Giles, 1998; Lawrence & Giles, 1999). Por su parte, el grupo de investigación de Digital publicaba sus análisis sobre la cobertura de los sistemas (Bharat & Broder, 1998) y el de Compaq aportaba importantes refinamientos metodológicos a la técnica de muestreo empleada por el equipo de NEC (Henzinger, Heydon, MIzenmacher, & Najork, 1999) que, previamente, el Web Characterization Project de OCLC había divulgado (O'Neill, McClain, & Lavoie, 1998).

Los trabajos más recientes comienzan a atribuir una semántica a los enlaces entre sedes y páginas, de forma que la interconexión entre documentos puede contribuir a definir “temas comunes” (Eckmann & Moses, 2002). Estos estudios “topológicos”, claramente basados en conceptos y técnicas bibliométricas (García Santiago, 2000) apoyan la conveniencia del análisis estructural, que paulatinamente se encamina a posibilitar el análisis automático de contenido (clasificación e indización) basado en la interconexión entre documentos Web (Glover, Tsioutsiouliklis, Lawrence, Pennock, & Flake, 2002).

La mera obtención de muestras del espacio Web es un factor ineludible en el cálculo de muchos indicadores empleados en la evaluación de los sistemas de recuperación de información distribuida, véase, si no, la tabla que aporta Ford (Ford, 2001). El presente trabajo limita sus objetivos, precisamente, a la obtención y caracterización de una muestra de sedes Web españolas como base para la evaluación de los sistemas de recuperación de información distribuida en Internet en el entorno nacional.

3.2 Selección de las sedes y criterios de análisis

A lo largo del presente trabajo, se consideran equivalentes los términos *dominio de segundo nivel* y *sede*, que constituyen la unidad de análisis. A este respecto, se sigue la definición propuesta por la Web Characterization Activity del W3C: “Una sede Web es una colección de páginas Web interrelacionadas, que incluye una como principal y residen en la misma localización de red” (Lavoie & Frystyk Nielsen, 2003). Aunque esta definición parece demasiado inclinada hacia la concepción física de las sedes web, será matizada más adelante, siguiendo la propuesta terminológica inicial (Web Characterization Project, 2003) que, curiosamente, se modificó en fecha posterior a la liquidación del grupo de W3C.

Las páginas web integradas en cada sede constituyen sus unidades de contenido. Sin embargo, cabe considerarlas como documentos compuestos de una serie de elementos: 1) indicadores de relación con otros documentos; 2) llamadas de función; 3) textos y 4) elementos gráficos en diversos formatos que pueden presentarse aislados o en combinación con los anteriores.

Sólo de forma aproximada se puede abordar el ingente contenido del espacio Web. La extracción de muestras es la metodología obligada cuando se trata de obtener indicadores globales, aunque se apliquen a un espacio limitado geográfica o culturalmente. Existen tres procedimientos de obtención de muestras:

1. un procedimiento general, destinado a la Web en su conjunto, considerado como universo informativo;
2. un procedimiento basado en recopilaciones existentes (normalmente las propias bases de datos de los sistemas de recuperación) que se destina al estudio de espacios nacionales o culturales determinados y
3. un procedimiento basado en la existencia de registros administrativos de sedes.

Diversos estudios han aplicado al universo Web un método de extracción de muestras aleatorias (Lawrence et al., 1998; Lawrence et al., 1999; O'Neill, Lavoie, & Bennett, 2003) que, básicamente,

sigue el procedimiento propuesto por el Web Characterization Project de OCLC (O'Neill *et al.*, 1998).

Este procedimiento se basa, básicamente, en la extracción aleatoria de cuatro grupos de números, cada uno entre 0 y 255 y su posterior combinación para formar los cuatro octetos que definen una dirección IP. Se trata de obtener, del universo de 4.294.967.296 direcciones posibles, un número manejable de URLs. Para asegurarse de que corresponden a documentos Web, se exige normalmente que se acompañen de la designación de puerto 080, el correspondiente al protocolo http. Monika Henzinger y su equipo han modificado la metodología para obtener mayor equilibrio de las muestras (Henzinger *et al.*, 1999).

Los estudios basados en el contenido recopilado por los sistemas, que se iniciaron con las comunicaciones de Tim Bray (Bray, 1996) y Allison Woodruff (Woodruff *et al.*, 1996) han continuado. Especialmente relevante resulta el análisis de Baeza Yates, quien basa su caracterización en el contenido recopilado por el sistema Buscopio[1] en el mes de marzo de 2001 (Baeza-Yates, 2002). Un defecto del empleo de estas fuentes es el sesgo introducido por las propias condiciones de recopilación del sistema en cuestión. El "muestreo" en este caso se fía a la cobertura del sistema utilizado. Así, Baeza se refiere a la "infrautilización del dominio .es" que, según sus cálculos, sólo representa el 30% de las sedes Web españolas. Sin embargo, un rápido cálculo empleando el sistema AltaVista arroja proporciones diferentes: 6.551.316 páginas en el dominio .es frente a 4.208.793 en el dominio .com, cifras siempre referidas a España. Aunque no aportan cifras, otras estimaciones se inclinan por considerar los nombres genéricos de dominio como mayoritarios frente a la denominación geográfica de las sedes españolas (Zook, 2000).

El tercer procedimiento de toma de datos también presenta limitaciones. El registro centralizado de sedes ha dejado de ser un procedimiento obligatorio y sólo recientemente se ha posibilitado la existencia de subdominios (Ministerio de Ciencia y Tecnología, 2003) algo que no sucede en otros entornos. Por otra parte, aunque se trata de un conjunto limitado, se ha puesto en evidencia

[1] http://www.buscopio.com ya no es operativo y en la actualidad (15 de Octubre de 2003) remite al sistema de búsqueda de PRISA y sus ediciones electrónicas

repetidamente que su composición, desde el punto de vista económico, social y cultural de España, corresponde a un conjunto equilibrado, tanto si se consideran la totalidad de las sedes registradas (Martínez de Lejarza Esparducer, 1999) como si se muestrean fragmentos del registro (Comisión del Mercado de las Telecomunicaciones, 2001).

No es posible extraer muestras útiles del espacio Web español mediante el muestreo de URLs. Tampoco se puede correr el riesgo de sesgar la muestra apoyándose en la cobertura de algún sistema. Por tanto, se recurrió al tercer procedimiento: a partir de la base de datos de ES-NIC, el Registro Delegado de Internet en España, se obtuvieron aleatoriamente 500 dominios de segundo nivel el día 4 de enero de 2001. En esa fecha, la base de datos contaba con 29.858 registros. Se debe hacer constar que, en septiembre de 2003, la cifra ha alcanzado las 58.580 sedes.

Para el análisis de las sedes, se empleó Xenu's Link Sleuth 1.1c, un programa de verificación de enlaces que, además, permite identificar algunas características de los dominios que analiza (Hausherr, 2001). Entre las variables que identifica destacan:

1. URL de cada página de la sede y de las páginas enlazadas
2. Tipo de fichero o aplicación correspondiente
3. Tamaño en bytes
4. Título de la página
5. Fecha de actualización de la página
6. Nivel jerárquico en la estructura de la sede
7. Número de enlaces de partida
8. Número de enlaces de llegada
9. Servidor que soporta la sede

De esta forma se examinó la estructura en niveles de cada sede y se individualizaron los elementos integrantes, su tipo y volumen. También se cuantificaron los enlaces de cada sede a otras, así como los enlaces de partida de cada página. La figura 3.1a muestra un ejemplo de análisis resultante.

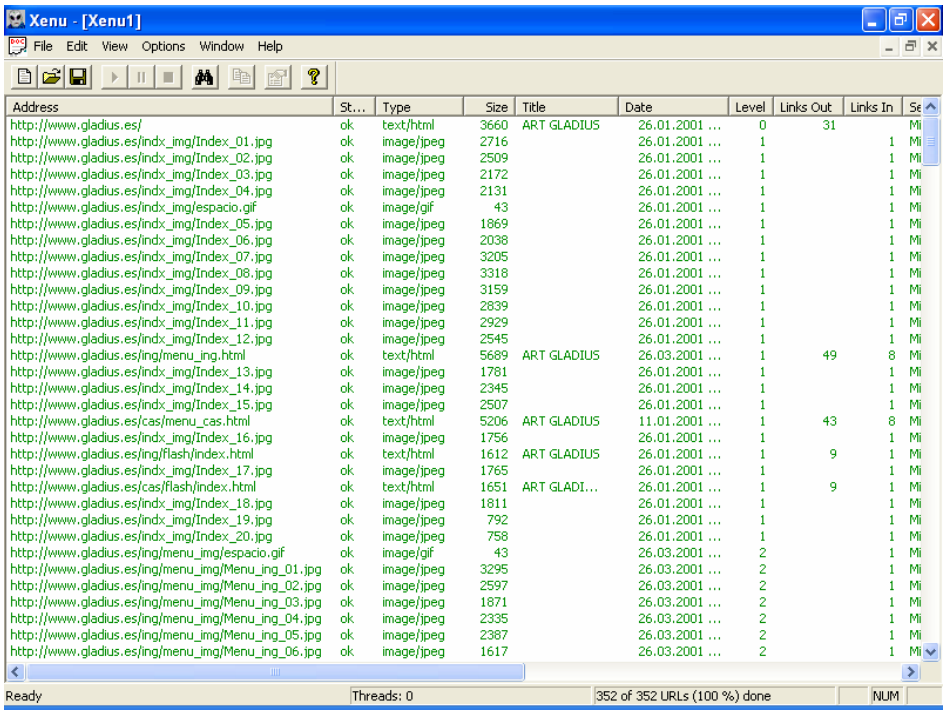
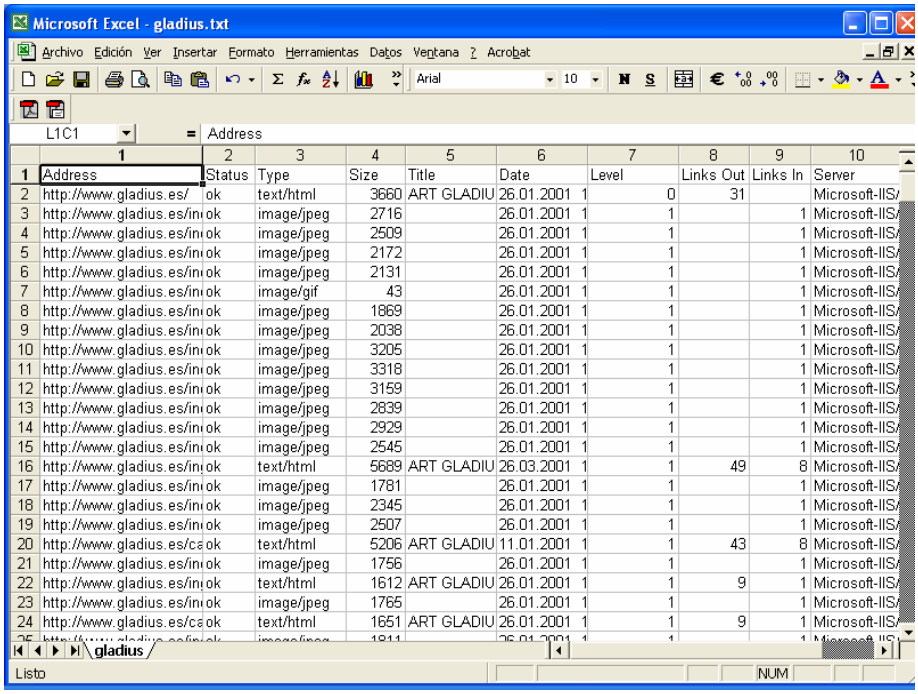


Figura 3.1a: Ejemplo del análisis de una de las sedes mediante Xenu's.

Los mismos datos se muestran en la figura 3.1b, incorporados a una hoja de cálculo para proceder a su análisis cuantitativo.



Microsoft Excel - gladius.txt

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ? Acrobat

Arial 10

L1C1 = Address

	1	2	3	4	5	6	7	8	9	10
1	Address	Status	Type	Size	Title	Date	Level	Links Out	Links In	Server
2	http://www.gladius.es/	ok	text/html	3660	ART GLADIU	26.01.2001 1	0	31		Microsoft-IIS/
3	http://www.gladius.es/in	ok	image/jpeg	2716		26.01.2001 1	1		1	Microsoft-IIS/
4	http://www.gladius.es/in	ok	image/jpeg	2509		26.01.2001 1	1		1	Microsoft-IIS/
5	http://www.gladius.es/in	ok	image/jpeg	2172		26.01.2001 1	1		1	Microsoft-IIS/
6	http://www.gladius.es/in	ok	image/jpeg	2131		26.01.2001 1	1		1	Microsoft-IIS/
7	http://www.gladius.es/in	ok	image/gif	43		26.01.2001 1	1		1	Microsoft-IIS/
8	http://www.gladius.es/in	ok	image/jpeg	1869		26.01.2001 1	1		1	Microsoft-IIS/
9	http://www.gladius.es/in	ok	image/jpeg	2038		26.01.2001 1	1		1	Microsoft-IIS/
10	http://www.gladius.es/in	ok	image/jpeg	3205		26.01.2001 1	1		1	Microsoft-IIS/
11	http://www.gladius.es/in	ok	image/jpeg	3318		26.01.2001 1	1		1	Microsoft-IIS/
12	http://www.gladius.es/in	ok	image/jpeg	3159		26.01.2001 1	1		1	Microsoft-IIS/
13	http://www.gladius.es/in	ok	image/jpeg	2839		26.01.2001 1	1		1	Microsoft-IIS/
14	http://www.gladius.es/in	ok	image/jpeg	2929		26.01.2001 1	1		1	Microsoft-IIS/
15	http://www.gladius.es/in	ok	image/jpeg	2545		26.01.2001 1	1		1	Microsoft-IIS/
16	http://www.gladius.es/in	ok	text/html	5689	ART GLADIU	26.03.2001 1	1	49	8	Microsoft-IIS/
17	http://www.gladius.es/in	ok	image/jpeg	1781		26.01.2001 1	1		1	Microsoft-IIS/
18	http://www.gladius.es/in	ok	image/jpeg	2345		26.01.2001 1	1		1	Microsoft-IIS/
19	http://www.gladius.es/in	ok	image/jpeg	2507		26.01.2001 1	1		1	Microsoft-IIS/
20	http://www.gladius.es/ca	ok	text/html	5206	ART GLADIU	11.01.2001 1	1	43	8	Microsoft-IIS/
21	http://www.gladius.es/in	ok	image/jpeg	1756		26.01.2001 1	1		1	Microsoft-IIS/
22	http://www.gladius.es/in	ok	text/html	1612	ART GLADIU	26.01.2001 1	1	9	1	Microsoft-IIS/
23	http://www.gladius.es/in	ok	image/jpeg	1765		26.01.2001 1	1		1	Microsoft-IIS/
24	http://www.gladius.es/ca	ok	text/html	1651	ART GLADIU	26.01.2001 1	1	9	1	Microsoft-IIS/

gladius

Listo NUM

Figura 3.1b: Resultado de la importación de los datos del mismo ejemplo.

Pueden evidenciarse en columnas sucesivas la URL de cada página, el estado de conexión, el tipo de fichero, su volumen en bytes, el título, la fecha de carga o modificación, su nivel jerárquico y los enlaces de partida y llegada a cada página.

3.3 Resultados sobre la concentración, la accesibilidad y otras características

3.3.1 Accesibilidad de la información

Las 500 sedes estaban albergadas en 171 proveedores de dominio distintos (mediana = 25, máximo = 49). La accesibilidad de cada sede se verificó mediante conexión directa. De las 500, sólo 168 presentaban información accesible. Las sedes en construcción (27’8 %), las direcciones erróneas (19’6 %) y las reservas de dominio, asociadas en la mayoría de los casos con marcas comerciales (8’2 %) constituían los 3 grandes grupos cuya información resultó inaccesible, dos tercios del total (Tabla 3.1).

Tabla 3.1: Accesibilidad por conexión directa de las sedes

	Casos	Porcentaje	Acumulado
Accesibles	170	34,0	34,0
Acceso prohibido	28	5,6	39,6
Conexión rechazada	1	,2	39,8
En construcción	139	27,8	67,6
Error	51	10,2	77,8
No encontrado	47	9,4	87,2
Otros	5	1,0	88,2
Redirección	18	3,6	91,8
Reserva de dominio	41	8,2	100,0
Total	500	100,0	

En algunos casos se detectaron elementos inaccesibles. Correspondían en su mayor parte a ficheros gráficos cuyo análisis rebasaba el tiempo fijado por el programa de verificación (ver párrafos siguientes) a pesar de que cada sede problemática se conectó en 3 ocasiones diferentes. El resto de estos elementos no

eran accesibles por la existencia de errores de ubicación, directorios inexistentes o encaminamientos fallidos.

3.3.2 Asignaciones múltiples y naturaleza de las sedes

También se verificó mediante conexiones directas la existencia de múltiples denominaciones o asignación a diferentes dominios de primer nivel de idénticos contenidos en los dominios de segundo nivel. 40 sedes ofrecían una doble denominación (.es y .com); 2 más ofrecían idénticos contenidos bajo .es y .net y una sede ostentaba triple asignación: .es, .com y .net. También se detectó denominación múltiple bajo dominio .es. Así, tpi.es ofrecía la misma información que paginas-amarillas.es en el momento de las conexiones.

Para investigar la naturaleza de las 500 sedes, se procedió a su examen y conversión al esquema inicial de dominios genéricos. La mayoría (N=450) se encuadraban en el dominio .com. Seguían grupos de los dominios .org y .net y había una proporción muy minoritaria de sedes .edu y .gov. Las correspondientes proporciones se muestran en la figura 3.2

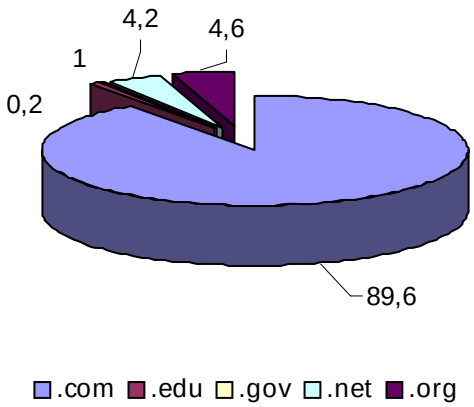


Figura 3.2: Distribución porcentual de las sedes por dominios genéricos

3.3.3 Tamaño y estructura

Las sedes analizadas contienen un total de 15.812 páginas que incluyen 47.735 elementos y ocupan 5’68 terabytes. Los estadísticos relativos a la variable tamaño (en bytes) se ofrecen en la Tabla 3.2:

Tabla 3.2: Tamaño de las sedes (en bytes)

Media		3.653.745,99
Mediana		659.477
Moda		390.063
Desv. típ.		11.729.034,68554
Varianza		137.570.254.654.683
Mínimo		2.068,00
Máximo		93.133.980
Suma		610.175.581
Percentiles	25	224.423
	50	659.477
	75	2.114.743

Por término medio, cada sede distribuye la información en 90,4 páginas, estructuradas en algo menos de 6 niveles jerárquicos (Tablas 3.3 y 3.4)

Tabla 3.3: Niveles jerárquicos en que se estructuran las sedes

Media		5,66
Mediana		4,00
Moda		4
Desviación típica		16,698
Mínimo		0
Máximo		218
Percentiles	25	3,00
	50	4,00
	75	5,00

Sin embargo, ni éstas ni las restantes variables estudiadas se ajustan a una distribución normal. Así, el número de niveles fluctúa entre 1 y 218, con desviación típica de 16,7, muy superior a la media de 5,66. La mediana de la distribución de esta variable se sitúa en 4. De igual modo, las medianas de las restantes distribuciones se alejan considerablemente de las medias: la del número de páginas se sitúa en 22’5 por sede, el volumen en 644,02 Kilobytes y el número de elementos en 81,5 por sede.

Tabla 3.4: Distribución estadística del número de páginas de la sedes

Media		90,37
Mediana		22,50
Moda		1
Desviación típica		209,199
Varianza		43764,174
Mínimo		1
Máximo		1831
Percentiles	25	11,25
	50	22,50

75 62,00

En la figura 3.2 se distribuye sobre escala logarítmica la variable número de elementos por sede (tomada como ejemplo) para evidenciar el carácter exponencial de la distribución. Las restantes variables muestran similar comportamiento. Por otra parte, la existencia de contenidos duplicados (traducciones a varias lenguas de las mismas páginas) y de versiones de acceso diferentes (con “flash”, sin marcos...) limita el valor de las variables referidas al volumen.

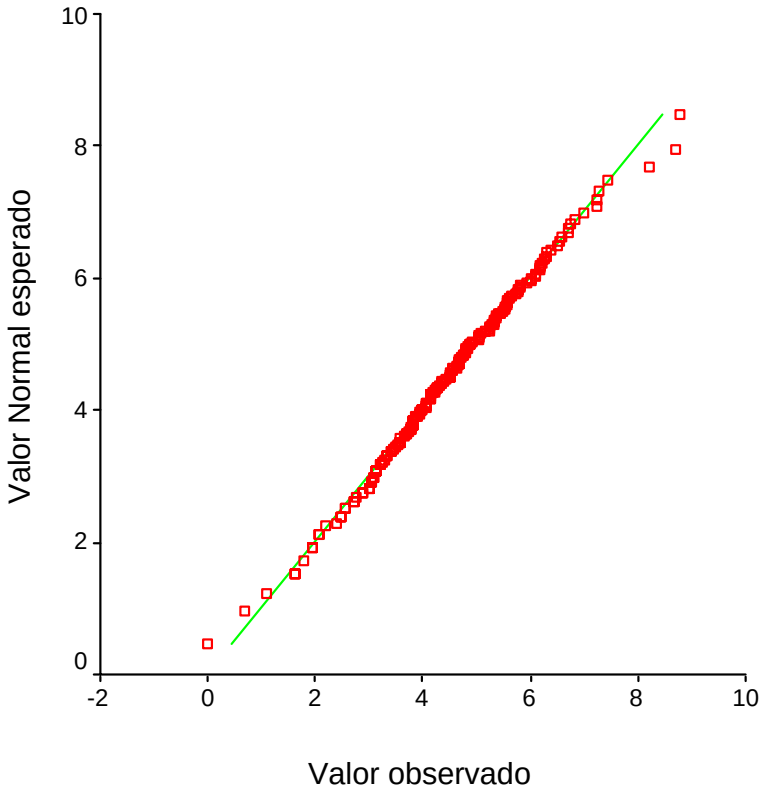


Figura 3.2: Ajuste exponencial del número de elementos por sede

3.3.4 Tipo de archivos y aplicaciones

Se han individualizado los elementos textuales bajo cinco formatos diferentes: los integrados en páginas HTML, los documentos resultantes de aplicaciones diversas (Acrobat y MS Word) y los resultantes de consultas a bases de datos: ASP y CGI. La figura 3.4 muestra la proporción de cada uno de estos segmentos y la preponderancia de las páginas HTML estáticas sobre las páginas dinámicas y los archivos preparados con otras aplicaciones. Sólo tres sedes empleaban documentos xml, si bien este formato representaba el 19 por ciento de las páginas de una de ellas.

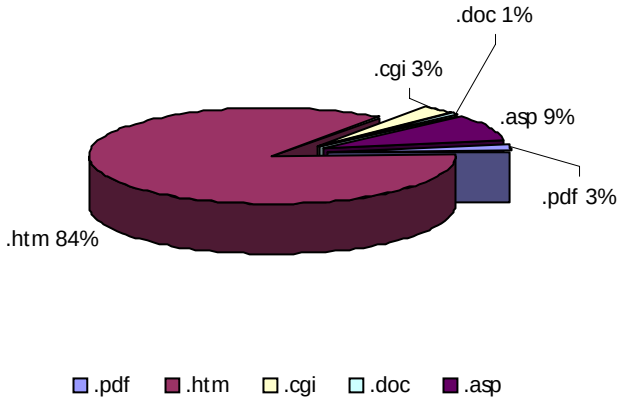


Figura 3.4 : Distribución por tipo de los elementos textuales de las sedes

Las páginas disponen de títulos significativos en muy alta proporción: en más de la mitad de las sedes, el número de páginas tituladas supera el 90%.

Tabla 3.5: Número de páginas con título

Media		76,1711
Mediana		90,5882
Desviación típica		72,75504
Mínimo		0
Máximo		900
Percentiles	25	46,5719
	50	90,5882
	75	100,0000

Los elementos gráficos suponen el 56,72 % del contenido de las sedes en términos de volumen. La mitad de las páginas contienen 1’7 archivos gráficos o menos. Existen, de nuevo, grandes desequilibrios en la distribución: sedes con gran riqueza gráfica y una con una media de 36 gráficos por página (Tabla 3.6).

Tabla 3.6: Distribución del número de elementos gráficos por página

Media		2,5490
Mediana		1,7251
Desviación típica		3,46596
Varianza		12,01287
Mínimo		0
Máximo		36
Percentiles	25	1,0000
	50	1,7251
	75	2,9891

Los formatos GIF y especialmente JPG son los dominantes, con proporciones que superan el 41 y el 58 por ciento respectivamente (Figura 3.5).

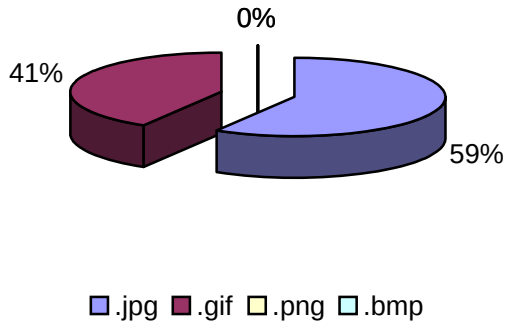


Figura 3.5: Proporción de formatos en los ficheros gráficos de las sedes

Presencia testimonial cabe atribuir a los formatos PNG y BMP. De igual modo, el contenido de elementos audiovisuales es muy escaso: menos del 5 % de las sedes incluyen archivos videográficos, una única sede contiene archivos sonoros y también es única la sede que ofrece representaciones de realidad virtual (archivos .vrml).

Las llamadas de función invocan operaciones a expensas de las acciones de los usuarios. Se han identificado operaciones de envío de correo y funciones javascript. Resulta chocante que 35 sedes no ofrezcan posibilidad alguna de contacto por correo. La mediana de esta distribución, de nuevo muy alejada de la normalidad, es de 1 y el valor máximo de 286. Los elementos *JavaScript* se utilizan sólo en el 30’5 % de las sedes, con un máximo de 100 y, nuevamente, una distribución muy desequilibrada.

3.3.5 Generación dinámica de páginas

Mientras cada sede contiene un promedio de 82 páginas estáticas (HTML), la proporción de contenidos generados dinámicamente es muy baja, con medias de 2,87 y 8,34 para las funciones CGI y ASP respectivamente. La distribución (Tabla 3.7) presenta, como en el resto de los estadísticos, grandes desequilibrios. Existen sedes con hasta 290 elementos ASP y otras

con más de 300 elementos CGI. Las funciones de generación dinámica de páginas están ausentes en el 91, 7% de las sedes para CGI y en el 94,6 para ASP. Existe una correlación directa y significativa ($p < 0,05$) entre el número de elementos HTML y las funciones de generación dinámica de página.

Tabla 3.7: Comparación entre el número de páginas estáticas y de páginas generadas dinámicamente

	CGI	HTML	ASP
Media	2,87	82,03	8,34
Mediana	0	19,00	0
Moda	0	1	0
Desviación típica	24,94	206,52	43,06
Mínimo	0	0	0
Máximo	303	1831	290
Total	482	13781	1401

3.3.6 Conectividad

El conjunto de sedes analizadas muestran en total 207.208 enlaces de partida (*links out*), de los cuales sólo 10.330 (un exiguo 5%) se dirigen a otras sedes. Las medianas de estas distribuciones son 218 y 2, respectivamente. 63 sedes, más de la tercera parte, no han establecido enlace alguno con otras.

Tabla 3.8: Conectividad de las sedes expresada por el número de enlaces

	Enlaces totales	Enlaces externos

Media	1233,38	61,49
Mediana	218,00	2,00
Desviación típica	3466,704	314,709
Total	207.208	10.330

La correlación entre el número de enlaces internos y el número de niveles jerárquicos de cada sede es significativa ($p < 0,01$) lo que ilustra la función de los enlaces como elementos estructuradores de las sedes. Por otra parte, el bajo nivel de conectividad que revela la escasa proporción de enlaces externos (a otras sedes) habla de cierta situación de “aislamiento”.

Existe correlación igualmente significativa ($p < 0,01$) entre el número de páginas de las sedes y el número de niveles jerárquicos.

3.3.7 Evolución de la accesibilidad

En diciembre de 2003 se examinaron de nuevo las sedes que habían resultado inaccesibles. La figura 3.6 ofrece la comparación en valores absolutos de las sedes inaccesibles en 2001 y 2003. De forma global, los trazados muestran que la proporción de sedes accesibles ha pasado del 34 % (véase la tabla 3.1) al 69,6 %. Esta práctica duplicación se ha producido a expensas, sobre todo, de la transformación en sedes operativas de aquellas que mostraban errores de acceso o se hallaban en construcción. Se debe tener en cuenta, en cualquier caso, que los datos porcentuales muestras proporciones similares en algunos grupos de sedes inaccesibles: los accesos prohibidos (8,48 frente a 8,55 %) y las sedes en construcción (42,12 frente a 45,39). En otros apartados, las proporciones correspondientes a 2003 superan a las originales de 2001: reserva de dominio (12,42 frente a 20,39) y conexión rechazada (0,3 frente a 1,32). Sólo la proporción de errores de conexión se ha reducido notablemente desde el 15,45 al 7,89 %.

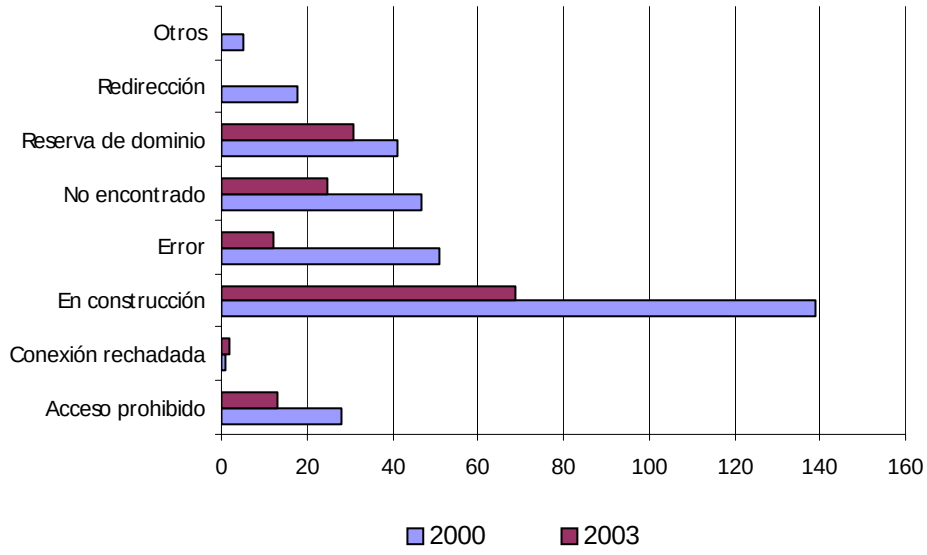


Figura 3.6: Diferencias en los grupos de sedes inaccesibles en 2001 y 2003

3.4 Repercusiones para los sistemas de recuperación

En general, el reducido número de sedes accesibles no es un dato sorprendente: un análisis del espacio Web británico reveló que de un total de 39162 dominios registrados, 18754 (47,89%) no conducían a página o sede alguna (Beckett, 1997).

Se ha mencionado ya que el volumen de las sedes depende a partes iguales de los elementos textuales y gráficos. Cabe añadir que la existencia de documentos .pdf tiene una gran incidencia en el volumen de archivos textuales, si bien es cierto que el formato de intercambio de documentos puede combinar tanto elementos textuales como gráficos en su contenido. Desde el punto de vista de la recuperación de la información contenida en sedes Web, las variables que se refieren al volumen de las sedes no son las más importantes. Mayor relevancia tiene la estructuración de las sedes y la distribución de los diversos tipos de documentos. Aunque las sedes que contienen documentos en formato de intercambio (.pdf)

apenas rebasan el 10 %, la información que contienen no podrá recuperarse a menos que un SIDI disponga de mecanismos de indización de este tipo de archivos. Por otra parte, existe una fuerte correlación positiva entre el número de páginas .html de una sede y su estructuración jerárquica. Sin embargo, esta correlación disminuye mucho cuando las páginas son .asp o resultado de una consulta .cgi. Y, si bien es cierto que las páginas dinámicas así generadas representan una pequeña parte en la muestra analizada, también lo es que su naturaleza dinámica y "volátil" impide que los sistemas capturen, indicen y posibiliten la recuperación adecuada de su contenido. En todo caso, el hecho de que determinadas sedes aporten la mayor parte del contenido (medido en número de páginas o en volumen de textos) podría suponer un desequilibrio en la cobertura de los sistemas de recopilación automática.

Así mismo, la distribución de enlaces tiene una incidencia muy directa sobre la accesibilidad de la información contenida en las sedes. El recurso a programas (robots) de recopilación automática que emplean los enlaces para desplazarse de sede en sede es el más habitual en los sistemas. Se ha evidenciado que la existencia de enlaces se justifica sobre todo por las necesidades de estructuración de las sedes. De hecho, es significativa ($p < 0,01$) la correlación entre el número de enlaces internos de las sedes y el correspondiente número de niveles jerárquicos. En general, la media de enlaces por página baja desde 12,77 enlaces internos (a otras páginas de la misma sede) hasta 0,5 enlaces externos (a otras sedes).

Sería aventurado esbozar un retrato robot de las sedes Web bajo dominio .es. Se ha demostrado que la naturaleza de las distribuciones lo desaconseja o lo imposibilita. En todo caso, sería posible definir tramos y adscribir una sede Web española al grupo dominante, al intermedio o a un tercer grupo residual. A la vista de los resultados globales, se puede decir que un alto número de páginas y elementos, con gran nivel de estructuración jerárquica o, alternativamente, el empleo significativo de procedimientos de generación dinámica de páginas (.asp, .php, .dtl u otros), formatos avanzados de representación (xml, css) y un gran número de enlaces a otras sedes caracterizan al grupo "avanzado" de sedes. En la medida en que las cifras descienden, se halla un grupo intermedio cuya característica distintiva es la baja interactividad:

pocos indicadores de relación (enlaces externos) y pocas llamadas de función (correo). Desde el punto de vista del contenido informativo, se dirían sedes “autosuficientes”, poco colaboradoras o aisladas. El grupo restante de sedes se muestran en estado incipiente de desarrollo, apenas sobrepasado el nivel cero de las meras reservas de dominio.

El procedimiento de análisis utilizado ofrece una visión estática, casi una instantánea de las sedes. Y, más que esta imagen sincrónica, que se limita al periodo de toma de datos, se requiere un seguimiento de la evolución de las sedes. El ritmo de crecimiento del dominio .es parece acelerarse: se observa un aumento del 34,74% en el número de dominios registrados en los meses transcurridos en 2003 frente a variaciones de 22,23 y el 20,21% en los dos años anteriores (2003) y el establecimiento de un observatorio permitiría un estudio evolutivo de las sedes en sus aspectos cuantitativos y también tecnológicos, relativos al tipo de aplicaciones empleadas para su mantenimiento y para el soporte de los contenidos que distribuyen.

No obstante las limitaciones apuntadas, el diseño, la implantación y la evaluación de sistemas de recuperación de la información distribuida en el espacio Web español ha de tener en cuenta los resultados de éste y otros trabajos de caracterización de ese mismo espacio, al que deben ajustarse al máximo sus características operativas, además de su cobertura.

4. Nivel de representación, opciones de recuperación y cobertura relativa de los sistemas españoles de recuperación de información distribuida en Internet

El apartado anterior, que trataba de caracterizar el espacio Web español, corresponde a grandes rasgos al estudio de la “línea de los documentos” en el esquema tradicional de los sistemas de recuperación de información. Sin abandonar este esquema, es necesario atender a otro de los componentes fundamentales: el esquema conceptual. Por esquema conceptual se entiende el conjunto de dispositivos, reglas y métodos que permiten al sistema la representación y organización de la información contenida en los documentos de la colección y también la representación de las demandas planteadas por los usuarios al sistema. Igualmente se integran en el esquema conceptual, auténtico núcleo del sistema, los procedimientos de presentación de los resultados en respuesta a las demandas; pero, por encima de todo, forma parte de este esquema la “función de similitud”, el cálculo del nivel de ajuste entre la expresión de cada demanda y la información de los documentos representados en el sistema.

El presente trabajo avanza en la línea trazada. Su objetivo general es la investigación de dos elementos clave del esquema conceptual de los sistemas españoles de recuperación de información distribuida y un tercer elemento crucial en la valoración de esos mismos sistemas. Los dos primeros se refieren al esquema de datos y a la dinámica de recuperación. El tercero a la cobertura.

El esquema de datos no sólo refleja el nivel de representación de los documentos en cada sistema sino que, junto con los datos relativos a su cobertura, ilustra el funcionamiento del principal módulo de cualquier buscador, su “crawler” o módulo de recopilación. Por otra parte, las funcionalidades del “indexer”

(módulo de indización) y el “searcher” (de recuperación) quedan reveladas (aunque sea parcialmente) por el análisis de la mecánica de recuperación. Puede verse la excelente descripción que Risvik y Michelsen realizan del sistema FAST (Risvik y Michelsen, 2002).

4.1 Fuentes y método

Los principales apartados metodológicos del presente estudio son 1) la selección de los sistemas objeto de estudio; 2) el examen de cada esquema de datos y su comparación con algún conjunto autorizado y normalizado; 3) la caracterización de la mecánica o de las opciones de recuperación y 4) la estimación de su cobertura.

4.1.1 Selección de los sistemas

Aunque se dispone de un sinfín de "listas de buscadores recomendados", la selección tuvo como punto de partida la lista distribuida en Red IRIS hasta mediados de 2001 y la recopilación de Buscopio[1]. En ambos casos se incluyeron inicialmente en el estudio los sistemas españoles de ámbito y temática general operativos a finales del año 2000. Siguiendo el procedimiento de Maldonado y Martínez, se realizó una "prueba de popularidad" en dos sistemas que permitían cuantificar el número de enlaces dirigidos a la dirección de cada servicio. Para minimizar el efecto de la integración de los sistemas de búsqueda en portales, la URL empleada correspondió a la dirección específica de las páginas de búsqueda (cuando fue posible). Así, por ejemplo, se sustituyó la expresión *link:www.telepolis.com* por la más exacta *link:buscador.telepolis.com*. La Tabla 4.1 muestra las cifras correspondientes a esta estimación (30 de Agosto de 2002) en los sistemas Google y AltaVista , junto con los mismos datos obtenidos por Maldonado y Martínez en 1998.

[1] Buscador de buscadores. Accesible en http://buscadores.buscopio.com [10 de Septiembre de 2002]

Tabla 4.1: “Popularidad” de los sistemas por el número de páginas que enlazan

Sistema	Google	Altavista	Maldonado
Altavista	580	58	
Apali	440	229	
BIWE	2110	635	1415
Buscopio	1630	664	
Elcano	2320	24	1742
Elindice	1830	6175	
Enlaweb	444	118	
Eureka	406	382	
Hispavista	4230	18025	869
Lycos	8980	5	
Ole (Terra)	268	3	5770
Ozú	3120	1865	1229
Salman	260	28	
Sol	1070	547	706
Telepolis	362	70	
Trovator	1390	18	1155
Ya	1330	164	
Yahoo	42500	1366	

La lista inicial incluía la porción española del sistema Excite y los sistemas llamados El buscador y Ozu.com. Desafortunadamente, la inestabilidad del espacio Web ha pasado factura a los propios sistemas: Excite.es cesó en su actividad el 11 de Junio de 2001 (Enos, 2001), Ozu.com se fusionó con Ozú.es y la URL de El buscador (http://www.elbuscador.com/) está ocupada actualmente por un proveedor de contenidos eróticos para Internet y servicios 906. Además, los intentos de conexión con elíndice.com ofrecen un error desde finales de agosto de 2002. Naturalmente, sólo se seleccionaron las porciones españolas de los sistemas Altavista, Lycos y Yahoo.

En Marzo y Abril de 2001 se realizaron envíos de una encuesta electrónica dirigida a todos los servicios. La encuesta se basaba en el modelo de formulario de DESIRE (Fase 1) (Koch, Ardo, Brümer y Lundberg, 1996) y tenía por objeto la determinación de los mecanismos de recopilación e indización empleados por cada servicio, además de otras características. El Anexo 1, al final de esta sección, reproduce el formulario enviado. Sólo uno de los servicios cumplimentó la encuesta, aunque sin ajustarse a la estructura requerida. Se impuso el examen directo de los servicios.

4.1.2 Determinación de los esquemas de datos

Las fuentes para la determinación de los datos recopilados por cada sistema fueron las páginas de búsqueda avanzada y/o las páginas de ayuda tanto a la inscripción como a la recuperación, los formularios de solicitud de incorporación al sistema, las listas de preguntas frecuentes (FAQ) disponibles y determinadas comprobaciones en la interrogación a los sistemas.

Como base de comparación se eligió la segunda versión del conjunto de elementos del Dublín Core Metadata (Dublin Core Metadata Initiative, 2003), descrito en 1.6.5.1. La correspondencia entre unos y otros esquemas se estableció de modo que, además de la relación obvia entre los elementos *title, date, type-format, identifier* y *language* (respectivamente título, fecha, formato, URL e idioma) se emparejaron *coverage* y ámbito geográfico, *subject* con palabras-clave y clasificación, *creator* con responsable, *publisher* con organismo editor y *description* con descripción. Se asignó a cada sistema una puntuación para cuantificar el nivel de representación de los documentos Web recopilados. Así, la correspondencia entre un elemento del esquema y su correspondiente elemento del Dublín Core equivalía a un punto, a excepción de *type-format*, valorado con puntos cuando se incluía, además del formato, su tamaño y de *subject*, puntuado doble también cuando, además de los epígrafes clasificatorios, se incluían palabras clave. Como los esquemas clasificatorios identificados dependían de cada sistema, no hubo oportunidad de asignar 3 puntos a esta variable, lo que hubiera significado que el esquema clasificatorio elegido estaba apoyado por alguna autoridad (LC, CDU, o cualquier otra). Igualmente, la

combinación *type-format* hubiera totalizado 3 puntos si se hubiera incorporado en algún caso el subelemento *version*.

4.1.3 La mecánica (las opciones) de recuperación

Las páginas de búsqueda simple y avanzada, la ayuda a la búsqueda y las fuentes empleadas en la determinación del esquema de datos se utilizaron para caracterizar los mecanismos de recuperación de cada sistema. Además de determinar las posibilidades y sintaxis de la expresión de perfiles, las posibilidades de combinación de criterios y la limitación por elementos (campos) se investigaron la existencia de mecanismos de refinamiento de búsqueda (*relevance feedback*) y búsqueda por similitud (*query by example*). Se procedió a la construcción de una escala de mecanismos de recuperación, de forma que el análisis pudiera traducirse en puntuaciones. Desde el nivel inferior, los tramos de esta escala son los siguientes: 1) empleo de operadores lógicos (búsqueda booleana); 2) empleo de expresiones compuestas (búsqueda por frase); 3) búsqueda por proximidad; 4) anidamiento de expresiones; 5) búsqueda en elementos definidos (limitación por campo); 6) truncación; 7) enmascaramiento; 8) refinamiento de resultados; 9) búsqueda por similitud y 10) posibilidad de seguimiento y alerta.

A grandes rasgos se puede decir que las diferentes fases de esta secuencia corresponden a niveles progresivos de procesamiento. En su valoración, sin embargo, se ha de tener en cuenta el hecho de que varios servicios españoles han licenciado la arquitectura de construcción de índices y de recuperación de otros sistemas. Por otra parte y al igual que en el sistema de puntuación propuesto para los esquemas de datos, existe una variable que puede alcanzar valores superiores a la unidad: la limitación por campos. En función del número de elementos que permitan precisar así la expresión de búsqueda, los sistemas obtendrán mayor o menor puntuación.

4.1.5 Estudio de la cobertura

Se realizaron dos series de operaciones para determinar la cobertura de los servicios. En primer lugar, se procedió a la extracción de una muestra aleatoria de dominios españoles de segundo nivel (sedes). La muestra se extrajo el 4 de enero de 2001 y consistió en un conjunto de 500 sedes bajo dominio .es del total de 29.858 registrados en esa fecha por ES-NIC, tal y como se describe en 3.2. En segundo lugar, se emplearon los indicadores de cobertura sumarizados por Abad (Abad García, 1997), utilizando el mismo método que Castillo, Martínez y Server (Castillo Blasco, Martínez de Pablos y Server, 1999). En concreto, se determinaron la tasa de solapamiento, la tasa de cobertura y el índice de aporte específico de cada servicio en relación con la muestra de sedes. Tal y como se pone de manifiesto en el trabajo de caracterización de la muestra, en la sección anterior, no todos los dominios contenían información o eran accesibles.

En segundo lugar, se determinó si un dominio era incluido en un sistema, mediante un conjunto de búsquedas (Marzo de 2001). Los perfiles empleados correspondieron a elementos de la URL de los dominios en aquellos sistemas que permitían tal modalidad de búsqueda. En los restantes, se utilizaron palabras del título de la portada (*home page*) de cada dominio en combinación booleana o como frase. Los resultados se examinaron para garantizar que cada sistema devolvía exactamente el dominio en cuestión.

4.2 Resultados y discusión

Atendiendo a la escala propuesta, los sistemas españoles de recuperación de información en Internet presentan esquemas de datos insuficientes. Sobre un total de 19 puntos, que corresponderían a un ajuste total a la lista de elementos del Dublin Core y a la interpretación cuantitativa que se ha propuesto, un sistema alcanzan una puntuación de 9 (ver Tabla 4.2) le siguen otro con 8 puntos y tres puntuados con 7. La puntuación mínima adjudicada es de 2. La tabla permite apreciar que, salvo en el caso de Altavista, los elementos *Subject* y *Type-Format* en combinación, son los criterios que en mayor medida contribuyen al aumento de las puntuaciones. El servicio con el esquema de datos más valorado (www.altavista.es o es-es.altavista.com) distribuye sus méritos relativos entre todos los elementos. La existencia de la combinación

palabras-clave y esquema clasificatorio rinde las mejores puntuaciones en lo que al elemento *Subject* respecta. También contribuye a ellas la inclusión de valores relativos al tipo de fichero y a su tamaño en la variable *type-format*. Ninguno de los sistemas analizados alcanza puntuaciones máximas en estos dos elementos y sólo uno obtiene una puntuación doble en la variable *date*: incluye no sólo la fecha de creación de los documentos recopilados, también la de modificación.

Tabla 4.2: Comparación entre el Dublín Core Element Set y los esquemas de datos de los servicios analizados

	T	Su	De	P	Da	T/F	I	So	L	R	C	Puntos
Altavista	1	1	1	0	1	1	1	1	1	1	1	**9**
Apali	1	1	1	0	1	0	1	0	1	0	0	**5**
BIWE	1	2	1	0	0	1	1	0	0	0	1	**6**
Buscopio	1	0	0	0	0	0	1	1	0	0	0	**2**
Elcano	1	1	1	1	2	2	1	0	0	0	0	**8**
Elindice	1	1	1	0	1	0	1	0	0	0	0	**4**
Enlaweb	1	0	1	0	0	0	1	0	0	0	0	**2**
Eureka	1	2	1	0	1	0	1	0	1	0	1	**7**
Hispavista	1	1	1	0	0	0	1	0	1	0	1	**5**
Lycos	1	1	0	0	1	2	1	0	1	0	1	**7**
Ole (Terra)	1	1	0	0	1	2	1		1	0	0	**6**
Ozú	1	0	0	0	0	2	0	1	0	1	0	**5**
Salman	1	1	1	0	1	0	1		1	0	1	**6**
Sol	1	1	1	0	0	2	1	1	0	1	0	**7**
Telepolis	1	2	1	0	0	0	1	0	1	0	1	**6**
Trovator	1	1	1	0	1	0	1	0	0	0	1	**5**
Ya	1	1	1	0	0	0	1		1	1	1	**6**
Yahoo	1	1	1	0	1	0	1	0	1	0	1	**6**

Equivalencia de las siglas de las columnas:

T	Title	T/F	Type-Format	C	Coverage
Su	Subject	I	Identifier		
De	Description	So	Source		
P	Publisher	L	Language		
Da	Date	R	Relation		

Los diez “tramos” en la escala de valoración de las opciones o mecanismos de recuperación son cubiertos de forma irregular por los sistemas. Las puntuaciones máximas (ver Tabla 4.3) corresponden a los fragmentos españoles de sistemas internacionales o a aquellos que han licenciado sus programas: Hispavista (8 puntos) y Ya (también 8) emplean tecnología FAST, Ozú (9) y Sol (6) emplean los programas de Google y Ole-Terra (8) los de Lycos. Las mayores aportaciones a la puntuación final proceden, naturalmente, de la variable “limitación por campos”, donde las puntuaciones más elevadas son, precisamente, las que corresponden a estos sistemas. Sólo el sistema Apali, a través de su mecanismo “*personal search*”, que requiere registro previo, ofrece la posibilidad de almacenamiento y reejecución diferida de perfiles de búsqueda. Las puntuaciones se han otorgado a los servicios atendiendo a las instrucciones de búsqueda, tal y como se ha especificado en el apartado metodológico. Se debe precisar, no obstante, que el sistema Altavista (puntuación total de 8 y de 2 en el componente de limitación) posee muchas más opciones de limitación por elemento (por campo) de las especificadas en sus instrucciones de búsqueda. Finalmente y en relación con la truncación, merece la pena que se anote que dos sistemas (Hispavista y Salman) realizan tanto truncación como enmascaramiento automáticos.

Tabla 4.3: Valoración de las opciones y mecanismos de recuperación de los sistemas analizados*

	Bool	Fras	Prox	Anid		Limi	Trun		Mas	Refi	
	Simi	Aler	Puntuación								
Altavista	1	1	1	1	2	1	1	0	0	0	8
Apali		1	1	0	0	0	0	0	0	0	1 3
BIWE		1	1	0	0	1	1	0	1	0	0 5
Buscopio	1	1	0	0	0	0	0	1	0	0	3
Elcano	1	1	0	0	1	1	1	0	0	0	5
Enlaweb	1	1	0	0	0	1	1	0	0	0	4
Eureka	1	1	0	0	1	1	0	0	0	0	4
Hispavista		1	1	0	0	4	1	1	0	0	0 8
Lycos		1	1	0	1	5	0	0	0	0	0 8
Ole/Terra	1	1	0	1	5	0	0	0	0	0	8
Ozú	1	1	0	0	4	1	0	1	1	0	9
Salman	0	1	0	0	2	1	1	0	0	0	5
Sol	1	1	0	1	1	1	1	0	0	0	6
Telepolis	1	1	0	1	0	1	1	0	0	0	5
Trovator	1	0	0	1	1	0	0	0	0	0	3
Ya	1	1	0	0	6	0	0	0	0	0	8
Yahoo	1	1	0	1	3	1	0	0	1	0	8

* Puntuaciones parciales y total obtenidas por los sistemas analizados en razón a las opciones de recuperación que ofrecen: 1) empleo de operadores lógicos (búsqueda booleana); 2) empleo de expresiones compuestas (búsqueda por frase); 3) búsqueda por proximidad; 4) anidamiento de expresiones; 5) búsqueda en elementos definidos (limitación por campo); 6) truncación; 7) enmascaramiento; 8) refinamiento de resultados; 9) búsqueda por similitud y 10) posibilidad de seguimiento y alerta. Se ha eliminado elindice.com, que ofrece un mensaje de error (error http 404: servidor no encontrado) desde finales de Agosto de 2002.

De las 500 sedes seleccionadas aleatoriamente, sólo 166 eran accesibles en el momento del estudio. Las razones de esta limitación se detallan y comentan en la sección anterior. Partiendo, en consecuencia, de un global de 166 sedes, los datos relativos a la cobertura de los sistemas estudiados se ofrecen en la tabla 4.4.

Tabla 4.4: Cobertura relativa de los sistemas analizados*

	S(a)	Global	S(x)	Específico
Altavista	58	34,94	6	3,61
Apali	0	0	0	0
BIWE	24	14,46	0	0
Buscopio	61	36,75	1	0,6
Elcano	5	3,012	0	0
Elindice	30	18,07	2	1,2
Enlaweb	0	0	0	0
Eureka	4	2,41	1	0,6
Hispavista	44	26,5	5	3,01
Lycos	54	32,53	1	0,6
Ole (Terra)	23	13,86	0	0
Ozú	21	12,65	1	0,6
Salman	7	4,22	0	0
Sol	25	15,06	0	0
Telepolis	9	5,42	0	0
Trovator	19	11,45	0	0
Ya	34	20,48	3	1,81
Yahoo	24	14,46	1	0,6

teórico global = 166

*Sobre la muestra de 166 sedes, S(a) es el número absoluto de sedes cubiertas por cada sistema, S(x) el número de sedes incluidas exclusivamente en cada sistema y las columnas restantes, los aportes global y específico de cada SIRI.

Puede apreciarse en ella que 2 sistemas presentaban nula cobertura y que los valores máximos en relación con el aporte global de cada sistema (el porcentaje de dominios cubiertos) rebasa un tercio del conjunto de dominios sólo en dos casos (Buscopio, 36,75% y Altavista 34,94%). En cuanto al aporte específico, es decir, al número de dominios que cada buscador ha incluido en su cobertura de forma específica, Altavista e Hispavista rebasan el 3%. De los restantes sistemas, 9 ofrecen una aportación exclusiva nula: la búsqueda de contenidos en sedes Web españolas podría prescindir de los mismos.

Los resultados expresados están sujetos a las mismas limitaciones que los obtenidos por trabajos anteriores: las determinaciones se han realizado en fechas concretas y no es posible que la repetición de los cálculos ofrezca resultados idénticos. Por otra parte, tal y como se ha expresado en el plan, el presente trabajo forma parte de una investigación que, sólo a su conclusión, permitirá la formulación de conclusiones sólidas.

No obstante, algunos hallazgos merecen ser destacados: el esquema de datos de la mayoría de los sistemas es rudimentario. Teniendo en cuenta que la mayor parte de los elementos del esquema de comparación son descriptivos (en oposición a analíticos, como sería el caso de las palabras-clave, la clasificación y la descripción) y dependen sólo de la configuración de los programas de recopilación, no se explica la falta de eficiencia que se ha hallado. Otro tanto sucede con las propiedades de los respectivos módulos de indización. A excepción de las variables situadas en los tramos finales de la escala empleada (*query by example*, *relevance feed-back* y alerta) y de aquellos que requieren una especial arquitectura de índices (truncación y enmascaramiento sobre todos), es difícil justificar las bajas puntuaciones de gran parte de los sistemas analizados en la limitación por campos o la búsqueda de términos próximos, precisamente las opciones de interrogación que permiten un mejor ajuste entre necesidad de información y resultados obtenidos.

No existe correlación (Figura 4.1) entre las puntuaciones asignadas a los esquemas de datos y las obtenidas por las

respectivas opciones de recuperación, hallazgo que habla a las claras de la mencionada falta de eficiencia de los sistemas:

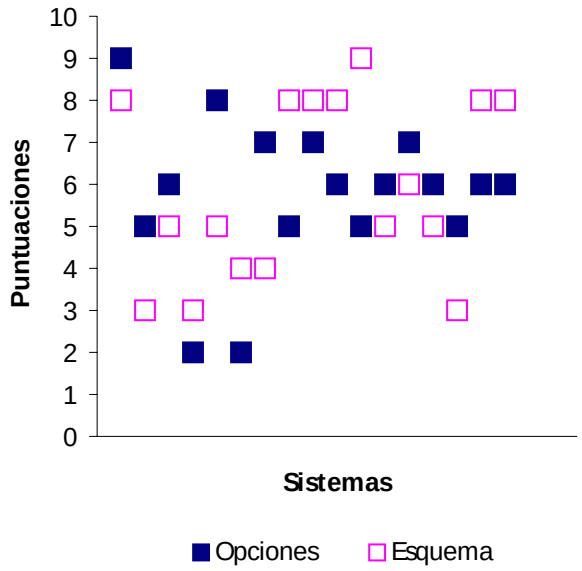


Figura 4.1: Correlación entre las puntuaciones asignadas a los esquemas de datos y las asignadas a las opciones de recuperación de los sistemas analizados.

Si hubiera que atenerse a los datos de cobertura resultantes del presente trabajo, se concluiría que las dos terceras partes de los contenidos de las sedes Web españolas resultan inaccesibles a los usuarios de los sistemas analizados. Por fortuna y además de la existencia de SIRI internacionales con otros presupuestos, la muestra empleada se limitaba a las sedes bajo dominio de primer nivel .es y no tenía en cuenta la plétora de sedes registradas en otros dominios más liberalizados (sobre todos .com). En todo caso, cabe anotar que, aunque existe una correlación positiva entre la aportación global y la específica de los sistemas estudiados ($p<0,01$) los valores son relativamente débiles (entre 0,53 y 0,66, dependiendo del estadístico empleado).

En términos generales, los SIRI españoles no se pueden considerar válidas “islas de información filtrada”, por emplear otra de tantas metáforas náuticas al uso, acuñada en el contexto de la edición electrónica (Butler, 1999). Los sistemas estudiados, en las condiciones observadas y con los métodos aquí empleados, ofrecen limitado acceso a poca información insuficientemente representada.

5. Rendimiento de los sistemas españoles de recuperación de información en Internet

Tras el análisis de algunas características de los sistemas analizados, es necesario proceder a la determinación de su rendimiento. En concreto, son dos cuestiones las que interesan:

1. ¿ Cuán eficaz es un sistema recuperando sólo páginas relevantes?
2. ¿ Es ése un buen sistema para encontrar la mayoría, todas o una proporción suficiente de las páginas relevantes existentes?.

Ambas cuestiones se relacionan con los indicadores más usuales de medida del rendimiento de los sistemas de recuperación de información: precisión y exhaustividad.

El empleo de la metodología tradicional para la evaluación de los sistemas de recuperación en Internet ha suscitado algunas críticas (Harter y Hert, 1997). En su mayoría se centran en la dificultad de determinar el número total de documentos relevantes en las colecciones. Sin embargo, la aplicación del concepto de relevancia y sus indicadores asociados, exhaustividad y precisión, se considera obligada (Oppenheim, Morris, Mcknight y Lowley, 2000). Así, la práctica totalidad de los trabajos agrupados por Martínez Méndez en el apartado de “estudios experimentales” de la recuperación de información en Internet emplean el concepto de relevancia y sus indicadores (Martínez Méndez y Rodríguez Múñoz, 2003).

Tan importante como la determinación de los indicadores y medidas a emplear es el ajuste de las condiciones experimentales. Michael Gordon y Praveen Pathak, autores de uno de los más sólidos trabajos de evaluación (Gordon y Pathak, 1999) enumeran siete requisitos básicos:

1. Las búsquedas se deben basar en necesidades de información genuinas. El que los experimentadores

determinen los temas puede introducir errores que, por ejemplo, favorezcan a un sistema en perjuicio de otros. Además, es necesario ajustar el experimento a la increíble diversidad de las necesidades informativas de los usuarios finales y casuales.

2. Cuando se trata de identificar información relevante a una necesidad, es necesario que la persona que formula la necesidad exprese también cuanta información contextual sea posible: listas de palabras clave o expresiones formales (por ejemplo booleanas) sólo pueden reflejar el tema de forma aproximada. Si acompañan a la expresión natural del tema de búsqueda, en cambio, pueden contribuir a despejar ambigüedades.
3. Se deben de realizar un número suficiente de búsquedas para obtener resultados significativos.
4. La prueba debe incluir los sistemas principales.
5. Se debe explotar cada sistema empleando sus características distintivas. No hay por qué emplear el mismo perfil en todos los sistemas analizados.
6. El juicio de relevancia debe correr a cargo del propio usuario. Si lo realiza el experimentador, puede originar errores debidos a su falta de familiaridad con el tema o a su desconocimiento de las necesidades y conocimiento previo reales.
7. El buen desarrollo experimental exige la obtención de medidas significativas de rendimiento mediante el empleo de un buen diseño (por ejemplo, la presentación de los resultados en orden aleatorio); mediante el ajuste a indicadores habituales (como los de exhaustividad y precisión) y mediante el empleo de análisis estadístico que informe de la significación de las diferencias halladas entre los sistemas.

A lo largo del presente experimento se han seguido estos requisitos básicos y otros adicionales revisados por Olvera (Olvera Lobo, 2000b) quien, sin embargo, no incluyó la evaluación de los resultados por los propios usuarios (Olvera Lobo, 2000a).

En el siguiente apartado se introducen los conceptos e indicadores empleados en este estudio. Posteriormente se detallan

las condiciones del experimento: expresión de las necesidades informativas, sistemas seleccionados, desarrollo de las búsquedas y evaluación de los usuarios. A continuación se presentan los resultados agrupados en cinco series.

5.1 Relevancia, exhaustividad y precisión como criterios de rendimiento

En una reciente revisión que destaca por su claridad, Raquel Gómez resume el procedimiento tradicional de evaluación del rendimiento de la recuperación, sus supuestos básicos, sus conceptos subyacentes, sus indicadores principales y su metodología (Gómez Díaz, 2003). El tratamiento de Francisca Abad es anterior, pero sus grandes líneas son coincidentes (Abad García, 1997a). Este apartado se apoya en ambas aportaciones y matiza la aplicación de los indicadores al estudio de los sistemas. La imprecisión terminológica y la ambigüedad de algunos conceptos se han limitado atendiendo a la discusión de los mismos realizada por Lancaster y Warner (Lancaster y Warner, 1993a)

A grandes rasgos, la relevancia mide la proximidad entre un documento y la formulación de una petición o expresión de la necesidad informativa. Una valoración meramente mecánica de la relevancia se obtiene mediante comparación entre un perfil de búsqueda (la expresión sintáctica o lógica que refleja la demanda) y los términos que reflejan el contenido informativo de los documentos. Sin embargo, “para evaluar un sistema de información real, con usuarios reales que formulan peticiones reales basadas en genuinas necesidades de información, es imperativo determinar la medida en que el servicio satisface las necesidades informativas de los usuarios” (Lancaster y Warner, 1993b). La influencia de los modelos probabilísticos de recuperación se ha extendido hasta considerar la relevancia una variable continua, que puede adoptar valores intermedios entre el 1 y el 0. Ello no obsta para que, en muchos casos, se emplee una escala discreta o se haga en otros un manejo binario del concepto.

Basándose en el cálculo de relevancia, es posible determinar el valor de los índices de precisión y exhaustividad, indicadores de rendimiento de la recuperación:

La exhaustividad mide la proporción de documentos relevantes que son recuperados. Corresponde al cociente entre el número de documentos relevantes recuperados y el total de documentos relevantes existentes en la colección.

La precisión mide la proporción de documentos recuperados que son relevantes. Refleja la eficacia de las búsquedas y pone en relación el número de documentos relevantes recuperados con el número total de documentos recuperados.

El cálculo de la exhaustividad presenta algunas dificultades en los sistemas tradicionales. Más aún en los sistemas de recuperación en Internet. En ambos casos, la determinación del denominador de la expresión (número total de documentos relevantes existentes en la colección) se estima de forma indirecta. Se habla entonces de exhaustividad relativa.

Las más recientes críticas al empleo de indicadores tradicionales proponen ciertas modificaciones, además de añadir nuevas variables de análisis (Vaughan, 2004) relacionadas con la estabilidad de las colecciones y el procedimiento de ordenación de resultados.

El modelo tradicional de evaluación de la recuperación de información parte de los experimentos realizados en los años 60 en Cranfield y se prolonga a partir de 1992 en la serie de conferencias anuales sobre recuperación textual TREC (Text Retrieval Conferences). En la actualidad, el National Institute of Standards and Technology estadounidense proporciona colecciones documentales y temas de búsqueda que cada sistema experimental procesa. Luego los resultados se comparan con controles (colecciones de referencia), constituidos por listas de documentos ordenados por relevancia. No es posible emplear este modelo, puesto que hasta hace muy poco (Bailey, Craswell y Hawking, 2003) no se ha podido contar con uno de los elementos esenciales: la colección de evaluación o colección de referencia que, por ende, contara con relaciones estructurales similares a las de los documentos hipertextuales del espacio Web (Gurrin y Smeaton, 2004). Una reciente alternativa, que propone el análisis automático del rendimiento de los sistemas (Can, Nuray y Sevdik, 2004) parece demasiado incipiente, aunque no esté exenta de interés.

5.2 Otras medidas

Aunque exhaustividad y precisión son los indicadores fundamentales en la determinación del rendimiento de los sistemas de recuperación, existen otros que permiten compararlos. De ellos, interesa la cobertura, el grado en que los sistemas incorporan el espacio Web y algunos indicadores relacionados: solapamiento y aporte específico (Abad García, 1997b). Una medida complementaria es la accesibilidad de los resultados, es decir, la presencia de los documentos hallados en todos los sistemas de búsqueda que, teóricamente, los incorporan. Otra, relacionada con el grado de actualización de los sistemas, es el número de resultados que cada uno devuelve de forma duplicada.

5.3 Método

5.3.1 Indicadores y medidas empleados

A lo largo del presente experimento, se determinan el solapamiento relativo y el aporte específico de cada sistema. El índice de solapamiento relativo se calcula mediante la relación entre el número de documentos recuperados simultáneamente por dos o más sistemas y el total de documentos recuperados por esos mismos sistemas (Abad García, 1997b). La expresión habitual es:

$$\text{Solapamiento entre A y B} = \frac{N(A \cap B)}{N(A \cup B)}$$

El aporte específico se obtiene mediante la relación entre el número de documentos recuperados exclusivamente por un sistema y el número total de documentos recuperados (Abad García, 1997b). Ambas medidas se han obtenido para cada sistema y cada búsqueda y, con posterioridad, se han promediado los valores calculados en cada búsqueda.

Igualmente, la determinación del rendimiento en términos de exhaustividad (E) y precisión (P) se ha realizado para cada una de los sistemas, búsquedas y niveles de resultados, siguiendo el

método aplicable a conjuntos de resultados ordenados (Salton y McGill, 1983).

Se ha calculado la exhaustividad relativa, tomando como denominador de la ecuación la suma de documentos juzgados relevantes en cada búsqueda por el total de sistemas.

Se ha empleado una determinación binaria de la relevancia, es decir, los documentos se han considerado relevantes o no relevantes. El juicio de relevancia ha sido subjetivo, expresado por los usuarios finales.

5.3.2 Expresión de las necesidades informativas

20 estudiantes de la asignatura “Sistemas de información en red” de la Diplomatura en Biblioteconomía y Documentación de la Universitat de València cumplimentaron un formulario de búsqueda diseñado para que expresaran la información que solicitaban. Las peticiones eran muy variadas, abarcando desde “la aplicación del método Nordoff-Robins en dificultades motrices” hasta “biografías de compositores clásicos”, “homosexualidad en la antigua Grecia”, “el mundo de la Tierra Media de Tolkien: mapas y planos”, los “tipos de conexión a Internet que se ofrecen en todo el mundo”, etc. Junto al tema de búsqueda, expresado en lenguaje natural, también se pidió a los participantes que sugirieran expresiones equivalentes, que ofrecieran la posible traducción a una expresión booleana de la combinación de términos, que anotaran los términos que se deberían eliminar, para evitar resultados erróneos y, además, se les requirió para que enumeraran los sistemas que empleaban habitualmente en sus búsquedas en Internet.

El anejo 5.1 reproduce el formulario tal y como fue cumplimentado por uno de los participantes.

5.3.3 Selección de los sistemas

Siete sistemas de recopilación automática y un directorio temático fueron objeto de estudio. Los sistemas fueron AltaVista, Hispavista, Lycos, Terra/OLE, Ozú, Sol y a y Yahoo; el directorio, Enlaweb. Los sistemas fueron seleccionados atendiendo al criterio expresado en el estudio anterior (véase 4.1 Selección de los sistemas). Además se desecharon todos aquellos que, en una

primera aproximación, ofrecieron menos de 5 resultados y no se contó con aquellos otros que habían cesado o modificado su actividad.

Yahoo cuenta con una base de datos de recopilación automática además del directorio clasificatorio que inicialmente constituía. Las búsquedas se realizaron en el primero de los componentes.

5.3.4 Desarrollo de las búsquedas

Todas las búsquedas fueron realizadas por el autor en las últimas semanas de mayo de 2002. La expresión textual de la necesidad de información y el resto de las expresiones contenidas en los formularios podían resolverse de diversas formas. Se empleó la técnica de perfil mejorado (*best search*) de forma que cada petición se sometió a cada sistema repetidamente, empleando diversos perfiles y formulaciones alternativas, hasta hallar resultados óptimos.

En el momento en que no se pudo mejorar los resultados de cada búsqueda en cada sistema, se registraron, en su correspondiente orden, las URLs de los 20 primeros resultados. Posteriormente, se trasladaron a hojas de cálculo y se compactaron. El objetivo era evitar que los usuarios se vieran obligados a examinar repetidamente un mismo resultado procedente de dos o más sistemas diferentes. La figura 5.1 presenta un fragmento de los resultados de búsqueda tal y como se presentaron al usuario. Las URLs de cada conjunto de resultados se ordenaron aleatoriamente y se ocultó el sistema o sistemas del que procedía cada resultado.

Figura 5.1: Algunos resultados de la búsqueda nº 4 (“el mundo de la Tierra Media”)

Resultados para MHR/04	**Valoración**

http://angelcities.com/members/sda/	
http://www.tinet.org/~axpa/spanish/to-cri1.htm	
http://www.cyberdark.net/ver.php3/fn/100/	
http://tv.ociojoven.com/article/tienda/6008/1/101/	
http://www.geocities.com/area51/8876/mapas.html	
http://www.submarino.com/books_productdetails.asp?Query=yProdTypeId=1yCatId=121yPrevCatId=121yProdId=2413871	
http://www.euskalnet.net/lordoftherings/contenido/imagenes/mapas.htm	
http://espanol.geocities.com/bfm_mitologico/Las_tierras_de_Mordor.html	
http://astrored.virtualave.net/nueveplanetas/solarsystem/earth.html	
http://www.anillounico.net	
http://www.tugueb.com/cine/2001/05/reportaje/tolkien.html	
http://www.visualmap.com/	
http://buscador.ya.com/indice/Ocio/Juegos/Rol/cat1.html	
http://tinet.fut.es/~jvega/mepbm.html	
http://www.libroadicto.com/anillos.htm	
https://www.terra.es/foro/portada.cfm?s=MEypCat=4ypForo=12624ypConv=628494ypExpan=1	

5.3.5 Evaluación por los usuarios

Cada alumno recibió un archivo de hoja de cálculo con la lista de resultados desordenados. El número de URLs que contenía rondaba en todos los casos el número de 100. Los participantes fueron instruídos sobre la expresión de su juicio de relevancia. En concreto, se hizo especial énfasis en que el juicio se ciñera a la página visualizada y que no se extendiera a páginas conexas o a la sede que las albergaba. Conectaron con cada URL y, tras su examen, determinaron si su contenido era pertinente (1) o irrelevante (0) a su necesidad de información. La imposibilidad de visualizar un documento se anotó como error (E) y, a efectos de cálculo, su valor se consideró nulo. Así mismo, se anotaron como valores nulos las repeticiones de resultados en un mismo sistema. La figura 5.2 muestra el juicio de relevancia expresado por uno de los participantes en el experimento, una vez tranferidas sus puntuaciones a la hoja de resultados originales. Se puede apreciar, en las columnas iniciales, los sistemas que devuelven cada

resultado con el número de orden en que aparecen. A continuación, la URL de cada documento recuperado y, finalmente, la puntuación adjudicada por el usuario y la indicación de error de conexión (E).

Figura 5.2: Evaluación de los resultados de búsqueda realizada por un usuario (fragmento)

Lyc01	Ozu01	Sol01	Tel01	Yah01	Yco01	http://w3.arrakis.es/iea/drogas/anaboli.htm	1
Lyc02						http://www.farmaweb.com/ffff/messages/96.html	E
Lyc03	Ole05	Ozu20	Tel19	Yah14		http://www.voraus.com/articulos/articulos/a000126.html	0
Lyc04	Ozu16					http://www.verdemente.com/anabolizantes.htm	E
Lyc05	Ozu14	Yah12				http://www.neogym-online.com/farmedana.htm	1
Lyc06	Yco03					http://www.depal.com.ve/productos/g2art.htm	0
Lyc07						http://www.llibreriapedagogica.com/informaciodrogues7.htm	1
Lyc08						http://lightning.prohosting.com/~ron90/doping.htm	1
Lyc09	Ozu12	Tel13				http://www.adcd.org/sp/drogas/droga13.html	1
Lyc10						http://copsa.cop.es/psicothema/AbsVol1136es.htm	1
Lyc11						http://club2.telepolis.com/luky5/anabolizantes.html	0

5.4 Resultados y discusión

Los cambios en los sistemas de recuperación de información en Internet son continuos. Afectan a su interfaz, a sus mecanismos de cobertura, a la representación de los documentos y su indización y también a sus procedimientos de recuperación y ordenación. Los hallazgos aquí expuestos son, en consecuencia, limitados y provisionales, aunque válidos por cierto tiempo.

Aunque se realizaron las búsquedas correspondientes a los 20 temas propuestos en los 8 sistemas seleccionados (un total de 160), sólo se dispuso de los juicios de relevancia de 11 usuarios, u 88 búsquedas evaluadas.

El experimento arrojó cinco series de resultados. En primer lugar, el número de errores contabilizados al intentar acceder a las páginas resultantes. En segundo lugar, una nueva estimación del grado de cobertura y solapamiento de los sistemas. En tercer lugar, una medición complementaria de la anterior, que determina la accesibilidad de los resultados. En cuarto lugar, los relativos al rendimiento de los sistemas estudiados, en términos de exhaustividad y precisión. Por último, algunos datos adicionales sobre el tipo de los documentos recuperados y su relación con los errores.

5.4.1 Errores de conexión

La tabla 5.1 resume los errores de conexión hallados. Los resultados se distribuyen entre los sistemas analizados (filas) y entre las búsquedas evaluadas (columnas B14 a B05). Globalmente, el 12,39% de los resultados de búsqueda condujeron a páginas erróneas. Sólo en una ocasión el acceso se vio imposibilitado por la exigencia de identificación y contraseña de usuario. En el resto, los intentos de acceso devolvieron un error 404.

Tabla 5.1: Distribución por sistemas y por búsquedas evaluadas de los errores de conexión hallados

	B14	B16	B04	B12	B01	B09	B19	Bem	B11	B08	B05	**Por sistema**	**%**
AltaVista	0	4	2	4	1	7	5	3	2	5	4	37	16,97
EnlaWeb	1	2	3	2	3	0	3	7	0	0	0	21	9,63
Lycos	0	0	1	1	1	0	7	2	1	1	1	15	6,88
Ole	2	0	4	2	6	1	3	2	5	2	1	28	12,84
Ozu	0	1	3	0	3	4	0	2	1	2	4	20	9,17
Sol	1	2	9	5	3	10	2	2	6	9	0	49	22,48
Ya	1	2	1	2	1	2	7	3	2	1	1	23	10,55
Yahoo	1	0	2	1	4	4	4	1	3	1	4	25	11,47
Por búsqueda	6	11	25	17	22	28	28	18	27	21	15	218	

Para averiguar si existía relación entre el número de resultados erróneos y la naturaleza de los documentos recuperados se compararon las proporciones respectivas. Las páginas dinámicas se identificaron por la extensión de los documentos (asp, php y otras) o por la inclusión en la URL de una función CGI. Por término medio, se halló una proporción del 10,78% en las búsquedas evaluadas. En este mismo conjunto de búsquedas (ver Tabla 5.1) la proporción de errores superaba el 12%.
La figura 5.3 ilustra la falta de correlación existente entre ambas proporciones.

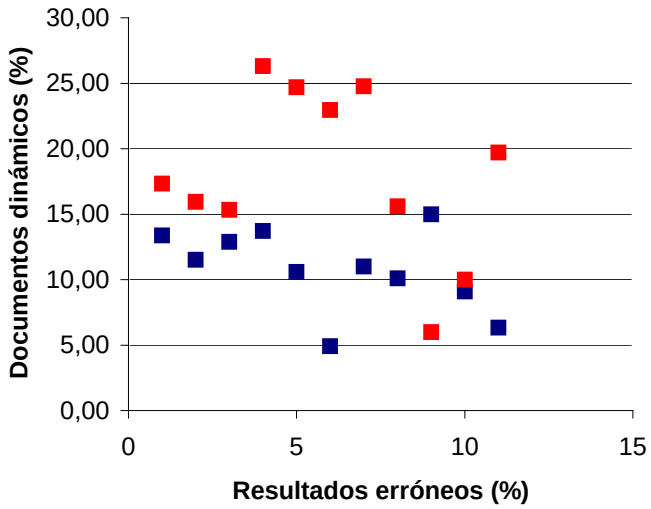


Figura 5.3: Correlación entre las proporciones de documentos dinámicos obtenidos en cada búsqueda y de errores devueltos.

Esta tasa de errores no se relaciona con el dinamismo de los documentos recuperados y, así, más parece atribuirse a un ritmo inadecuado de actualización de las bases de datos de cada sistema. La proporción de resultados duplicados podrían avalar esta conclusión: el sistema con más errores (Sol) también muestra la mayor proporción de duplicados (10%), junto con Yahoo. Sin embargo, no existe correlación significativa entre la proporción de errores y de resultados duplicados.

5.4.2 Accesibilidad de las páginas halladas

La tabla 5.2 ofrece algunas cifras absolutas sobre la accesibilidad de los resultados de las búsquedas realizadas. Se han tenido en cuenta todas las búsquedas realizadas, no sólo las evaluadas. Para cada una de las búsquedas (b01 a b20) se ofrece el número de documentos resultantes y, en las columnas siguientes, el número de sistemas (entre 1 y un máximo de 7) que devolvieron esos resultados. Puede apreciarse que sólo en un caso (búsqueda

14) un resultado fue devuelto por siete de lo sistemas estudiados y en ningún caso todos los sistemas cubrían un mismo recurso.

Tabla 5.2 Número de sistemas de procedencia de los resultados de búsqueda

	Unicos	1	2	3	4	5	6	7
b01	127	100	21	6	0	0	0	0
b02	91	58	18	11	1	1	2	0
b03	113	90	13	4	4	1	1	0
b04	163	138	23	2	0	0	0	0
b05	57	25	15	7	6	3	1	0
b06	123	104	11	7	1	0	0	0
b07	98	79	9	8	2	0	0	0
b08	85	51	17	14	1	2	0	0
b09	122	94	18	10	0	0	0	0
b10	138	124	10	1	2	1	0	0
b11	109	83	13	7	2	4	0	0
b12	109	80	18	10	0	1	0	0
b13	97	60	29	5	2	1	0	0
b14	100	69	14	9	6	1	0	1
b15	109	83	13	7	2	4	0	0
b16	110	83	20	6	1	0	0	0
b17	102	70	16	9	4	3	0	0
b18	114	94	16	2	2	0	0	0
b19	142	109	24	8	1	0	0	0
b20	88	68	10	5	4	1	0	0
Total	2197	1662	328	138	41	23	4	1

La figura 5.4, desplegada a continuación, representa gráficamente los datos, empleando una escala porcentual en el eje de valores. A pesar de que los sistemas analizados parten con el objetivo común de proporcionar acceso a los documentos del espacio Web en España, casi el 76% de los resultados obtenidos se encuentran cubiertos de forma exclusiva por un único sistema entre ocho.

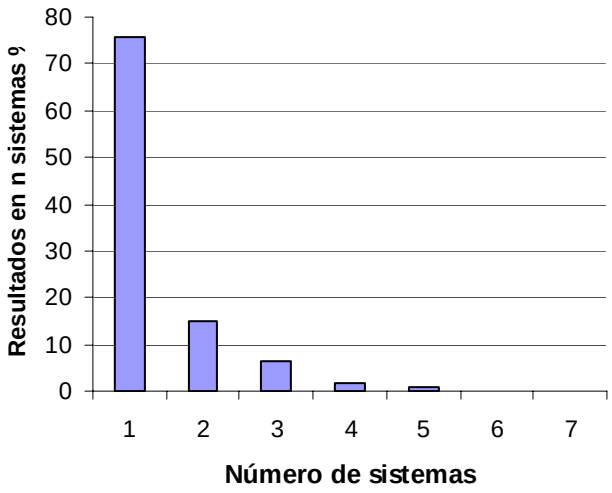


Figura 5.4 Accesibilidad de los documentos resultantes de las búsquedas

La tabla 5.3 refleja la duplicidad de contenidos de las bases de datos de los sistemas estudiados (un indicador relacionado con su grado de actualización). En ella se distribuyen las ocasiones en que los sistemas han devuelto más de una vez un resultado idéntico para determinada búsqueda. Las columnas de totales contienen la suma de resultados coincidentes (en un mismo sistema) para cada conjunto de resultados de búsqueda (Total b) y para cada sistema a lo largo de las 20 búsquedas (Total s).

Tabla 5.3 Duplicidad de contenidos en los sistemas expresada por el número de resultados idénticos

	Altavista	Enlaweb	Lycos	Ole	Ozu	Sol	Ya	Yahoo	Total b
b1	8	10	9	4	10	11	4	4	60
b2	0	0	0	0	0	0	0	0	0
b3	0	0	0	2	0	0	3	0	5
b4	0	0	0	0	0	0	0	0	0
b5	0	0	0	0	0	0	0	0	0
b6	0	0	0	0	0	0	0	0	0
b7	0	0	0	0	0	0	0	0	0
b8	0	0	0	0	0	0	0	0	0
b9	0	0	0	0	1	0	0	0	1
b10	0	0	0	0	0	0	0	0	0
b11	0	0	0	0	0	0	1	0	1
b12	0	0	0	0	0	0	0	0	0
b13	2	1	2	1	1	0	0	5	12
b14	0	0	0	2	0	4	0	6	12
b15	2	0	0	0	0	0	1	0	3
b16	0	0	0	0	0	0	0	0	0
b17	0	0	0	0	0	0	0	0	0
b18	0	0	0	0	0	0	0	0	0
b19	0	0	0	0	0	1	1	0	2
b20	0	0	0	0	0	0	0	1	1
Totals	12	11	11	9	12	16	10	16	

5.4.3 Solapamiento e índice de aporte específico

El solapamiento de los sistemas se refiere al número de resultados coincidentes procedentes de los ocho sistemas estudiados. Se ha determinado búsqueda a búsqueda entre todos los sistemas tomados dos a dos y se ha promediado entre las 20 búsquedas realizadas.

Así, por ejemplo, la búsqueda número 2 contiene 16 coincidencias entre los sistemas Lycos y OLE/Terra, lo que arroja un grado de solapamiento de 16/40 = 0,40 o, en términos porcentuales, el 40%. Los valores promedios de solapamiento entre todos los sistemas, acompañados de la desviación típica de cada distribución, se expresan en la tabla 5.4.

Tabla 5.4: Solapamiento entre los sistemas estudiados*

	Altavista	**Enlaweb**	**Lycos**	**OLE**	**Ozu**	**Sol**	**Ya**	**Yahoo**
Altavista		0,001 ± 0,001	0,03 ± 0,09	0,04 ± 0,04	0,04 ± 0,06	0,01 ± 0,03	0,13 ± 0,44	0,06 ± 0,07
Enlaweb			0,01 ± 0,02	0,01 ± 0,02	0,01 ± 0,03	0,01 ± 0,02	0,01 ± 0,02	0,00 ± 0,01
Lycos				0,20 ± 0,18	0,03 ± 0,04	0,01 ± 0,02	0,20 ± 0,13	0,05 ± 0,06
OLE					0,06 ± 0,06	0,01 ± 0,02	0,21 ± 0,15	0,07 ± 0,07
Ozu						0,01 ± 0,02	0,05 ± 0,06	0,19 ± 0,13
Sol							0,01 ± 0,02	0,02 ± 0,05
Ya								0,05 ± 0,06
Yahoo								0,05 ± 0,06

*Datos expresados com media ± desviación típica

El valor máximo corresponde a los sistemas Ya y OLE, cuyos resultados aparecen solapados en un 21%. En general, es el sistema Ya quien presenta mayor grado de solapamiento con el resto, mientras que Enlaweb (el único directorio evaluado) y Sol muestran los menores valores en conjunto.

El indicador de aporte específico, complementario al de solapamiento, se obtiene mediante la relación porcentual entre el número de documentos recuperados por un sistema exclusivamente y el total de documentos recuperados en una búsqueda determinada. Se ha calculado para cada uno de los ocho sistemas y para cada una de las 20 búsquedas. En la tabla 5.5 se expresan los valores promediados y sus desviaciones. La figura 5.5 expresa gráficamente los promedios hallados.

Tabla 5.5: Valores promedio de aportación específica de cada sistema

	media	d.e.
Altavista	10,65	2,959
Enlaweb	9,478	4,915
Lycos	6,095	4,561
OLE	3,805	3,308
Ozu	6,066	3,497
Sol	9,082	4,79
Ya	5,957	4,632
Yahoo	6,853	3,756

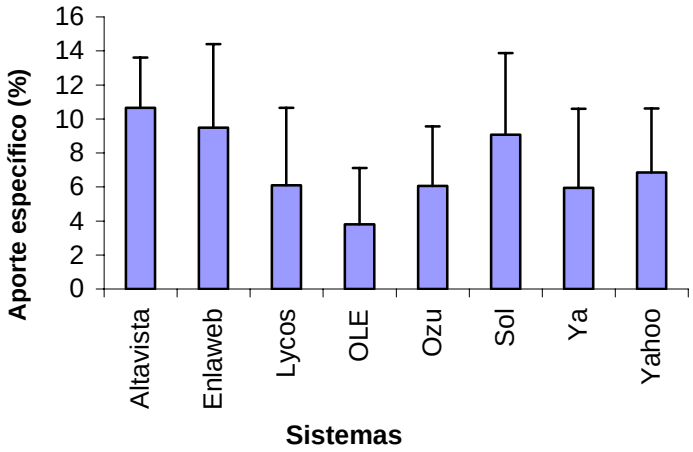


Figura 5.5: Aporte específico de cada sistema expresado en promedio y desviación típica

Del total de 3200 resultados de búsqueda, el número de documentos resultantes únicos, descontadas las apariciones repetidas, fue de 2197. De ellos, algo más del 75% fueron obtenidos por sólo un sistema. Otra observación que, junto con las anteriores, permite calificar de deficitarios los procedimientos de recopilación de los sistemas y, en consecuencia, su cobertura. En el hecho de que ningún resultado de búsqueda haya sido cubierto por los 8 sistemas estudiados ha de contrastarse con la especial naturaleza del sistema EnlaWeb: se trata de un directorio que incorpora sedes, no páginas de niveles inferiores, lo que dificulta que las URLs ofrecidas coincidan con las de los sistemas de recopilación automática. No es de extrañar que su aporte específico, la proporción de documentos recuperados exclusivamente por este directorio, sea superior al de todos los sistemas exceptuando Altavista. También su bajo nivel de solapamiento es indicativo.

En 3.3.6 se ofrecía una estimación del número de enlaces entre las páginas Web del dominio .es analizadas. En 4.1.5 se aportaron datos sobre el nivel de cobertura de las sedes por parte de los sistemas analizados. Los datos sobre la cobertura de las sedes y la accesibilidad de los resultados aportados aquí son coherentes con los hallazgos de los apartados mencionados.

5.4.4 Exhaustividad y precisión de los sistemas

Las tablas 5.6 a 5.13 muestran los valores de exhaustividad (E) y precisión (P) de las búsquedas realizadas en cada sistema. Los resultados se detallan para cada nivel de respuesta, del 1 al 20. El cálculo de la exhaustividad toma como denominador el número total de documentos relevantes identificados por los usuarios en el contexto global de las búsquedas en los 7 sistemas. Las cifras son B14=22, B16=9, B04=28, B12=30, B01=33, B09=28, B19=4, Bem=46, B11=72, B08=3 y B05=49.

Tabla 5.6: Valores de exhaustividad y precisión para las búsquedas realizadas en el sistema AltaVista

B14		B16		B04		B12		B01		B09		B19		Bem		B11		B08		B05	
E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P
0,00	0,00	0,00	0,00	0,00	0,00	0,03	1,00	0,00	0,00	0,04	1,00	0,00	0,00	0,02	1,00	0,01	1,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,07	1,00	0,03	0,50	0,04	0,50	0,00	0,00	0,02	0,50	0,01	0,50	0,00	0,00	0,00	0,00
0,05	0,33	0,00	0,00	0,00	0,00	0,07	0,67	0,03	0,33	0,07	0,67	0,00	0,00	0,04	0,67	0,03	0,67	0,00	0,00	0,02	0,33
0,05	0,25	0,00	0,00	0,00	0,00	0,10	0,75	0,06	0,50	0,11	0,75	0,00	0,00	0,07	0,75	0,04	0,75	0,00	0,00	0,04	0,50
0,05	0,20	0,00	0,00	0,00	0,00	0,10	0,60	0,06	0,40	0,11	0,60	0,00	0,00	0,09	0,80	0,06	0,80	0,00	0,00	0,04	0,40
0,05	0,16	0,00	0,00	0,00	0,00	0,10	0,50	0,06	0,33	0,14	0,67	0,00	0,00	0,09	0,67	0,06	0,67	0,00	0,00	0,06	0,50
0,05	0,14	0,00	0,00	0,00	0,00	0,10	0,43	0,06	0,29	0,14	0,57	0,00	0,00	0,09	0,57	0,06	0,57	0,00	0,00	0,08	0,67
0,05	0,13	0,00	0,00	0,00	0,00	0,10	0,38	0,06	0,25	0,14	0,50	0,00	0,00	0,09	0,50	0,06	0,50	0,00	0,00	0,10	0,83
0,09	0,22	0,00	0,00	0,04	0,11	0,10	0,33	0,06	0,22	0,14	0,44	0,00	0,00	0,09	0,44	0,06	0,44	0,00	0,00	0,10	0,56
0,14	0,30	0,11	0,10	0,04	0,00	0,10	0,30	0,08	0,30	0,14	0,40	0,00	0,00	0,09	0,40	0,06	0,40	0,00	0,00	0,10	0,50
0,14	0,27	0,11	0,09	0,04	0,09	0,10	0,27	0,08	0,27	0,14	0,36	0,00	0,00	0,09	0,36	0,06	0,36	0,00	0,00	0,10	0,45
0,14	0,25	0,11	0,08	0,04	0,08	0,10	0,25	0,08	0,25	0,14	0,33	0,00	0,00	0,09	0,33	0,07	0,42	0,00	0,00	0,10	0,42
0,14	0,23	0,11	0,08	0,04	0,08	0,10	0,23	0,08	0,23	0,18	0,38	0,00	0,00	0,09	0,31	0,07	0,38	0,00	0,00	0,10	0,38
0,14	0,21	0,11	0,07	0,07	0,14	0,10	0,21	0,08	0,14	0,18	0,36	0,00	0,00	0,11	0,36	0,13	0,43	0,00	0,00	0,12	0,43
0,14	0,20	0,11	0,07	0,13	0,30	0,10	0,20	0,08	0,20	0,18	0,33	0,00	0,00	0,13	0,40	0,13	0,47	0,00	0,00	0,12	0,40
0,14	0,19	0,11	0,06	0,07	0,13	0,10	0,19	0,08	0,19	0,18	0,31	0,00	0,00	0,13	0,38	0,14	0,50	0,00	0,00	0,12	0,38
0,1	0,1	0,1	0,0	0,1	0,1	0,1	0,1	0,0	0,1	0,2	0,3	0,0	0,0	0,1	0,4	0,1	0,5	0,0	0,0	0,1	0,3

4	8	1	6	1	8	0	8	8	8	1	5	0	0	5	1	5	3	0	0	2	5
0,18	0,22	0,22	0,11	0,11	0,17	0,10	0,17	0,08	0,17	0,21	0,33	0,00	0,00	0,17	0,44	0,17	0,56	0,00	0,00	0,12	0,33
0,18	0,21	0,22	0,11	0,14	0,21	0,10	0,16	0,08	0,16	0,25	0,37	0,00	0,00	0,20	0,47	0,18	0,58	0,00	0,00	0,14	0,39
0,18	0,20	0,22	0,10	0,14	0,20	0,10	0,15	0,08	0,15	0,25	0,35	0,00	0,00	0,22	0,50	0,20	0,60	0,00	0,00	0,14	0,35

Tabla 5.7: Valores de exhaustividad y precisión para las búsquedas realizadas en el sistema Enlaweb

B14		B16		B04		B12		B01		B09		B19		Bem		B11		B08		B05	
E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	1,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,50	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,33	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,33	0,03	0,67	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,25	0,03	0,25	0,00	0,00	0,00	0,00	0,04	0,50	0,04	0,75	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,20	0,03	0,20	0,00	0,00	0,00	0,00	0,04	0,40	0,06	0,80	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,17	0,03	0,17	0,00	0,00	0,00	0,00	0,07	0,50	0,07	0,83	0,00	0,00	0,00	0,00
0,05	0,14	0,00	0,00	0,00	0,00	0,03	0,14	0,03	0,14	0,00	0,00	0,00	0,00	0,07	0,43	0,08	0,86	0,00	0,00	0,00	0,00
0,09	0,25	0,00	0,00	0,00	0,00	0,07	0,25	0,03	0,13	0,00	0,00	0,00	0,00	0,09	0,50	0,08	0,75	0,00	0,00	0,00	0,00
0,09	0,22	0,00	0,00	0,00	0,00	0,07	0,22	0,03	0,11	0,00	0,00	0,00	0,00	0,11	0,56	0,08	0,67	0,00	0,00	0,00	0,00
0,09	0,20	0,00	0,00	0,00	0,00	0,07	0,20	0,03	0,10	0,00	0,00	0,00	0,00	0,11	0,50	0,08	0,60	0,00	0,00	0,00	0,00
0,09	0,18	0,00	0,00	0,00	0,00	0,07	0,18	0,03	0,09	0,00	0,00	0,00	0,00	0,13	0,55	0,08	0,55	0,00	0,00	0,00	0,00
0,09	0,17	0,00	0,00	0,00	0,00	0,07	0,17	0,03	0,08	0,00	0,00	0,00	0,00	0,13	0,50	0,08	0,50	0,00	0,00	0,00	0,00
0,14	0,23	0,00	0,00	0,00	0,00	0,07	0,15	0,03	0,08	0,00	0,00	0,00	0,00	0,15	0,54	0,10	0,54	0,00	0,00	0,00	0,00
0,18	0,29	0,00	0,00	0,00	0,00	0,07	0,14	0,03	0,07	0,00	0,00	0,00	0,00	0,17	0,57	0,11	0,57	0,00	0,00	0,00	0,00
0,18	0,27	0,00	0,00	0,00	0,00	0,07	0,13	0,03	0,07	0,00	0,00	0,00	0,00	0,20	0,60	0,11	0,53	0,00	0,00	0,00	0,00
0,18	0,25	0,00	0,00	0,00	0,00	0,07	0,13	0,03	0,06	0,00	0,00	0,00	0,00	0,22	0,63	0,13	0,56	0,00	0,00	0,00	0,00
0,18	0,24	0,00	0,00	0,00	0,00	0,07	0,12	0,03	0,06	0,00	0,00	0,00	0,00	0,22	0,59	0,13	0,53	0,00	0,00	0,00	0,00

0,18	0,22	0,00	0,00	0,00	0,00	0,07	0,11	0,03	0,06	0,00	0,00	0,00	0,00	0,24	0,61	0,14	0,56	0,00	0,00	0,00	0,00
0,18	0,21	0,00	0,00	0,00	0,00	0,10	0,16	0,03	0,05	0,00	0,00	0,00	0,00	0,24	0,58	0,15	0,58	0,00	0,00	0,00	0,00
0,18	0,20	0,00	0,00	0,00	0,00	0,10	0,15	0,03	0,05	0,00	0,00	0,00	0,00	0,24	0,55	0,17	0,60	0,00	0,00	0,00	0,00

Tabla 5.8: Valores de exhaustividad y precisión para las búsquedas realizadas en el sistema Lycos

B14		B16		B04		B12		B01		B09		B19		Bem		B11		B08		B05	
E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P
0,00	0,00	0,00	0,00	0,00	0,00	0,03	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	1,00
0,00	0,00	0,00	0,00	0,00	0,00	0,07	1,00	0,03	0,50	0,00	0,00	0,00	0,00	0,02	0,50	0,01	0,50	0,00	0,00	0,02	0,50
0,00	0,00	0,00	0,00	0,00	0,00	0,10	1,00	0,03	0,33	0,00	0,00	0,00	0,00	0,02	0,33	0,01	0,33	0,00	0,00	0,02	0,33
0,00	0,00	0,00	0,00	0,00	0,25	0,10	0,75	0,03	0,25	0,00	0,00	0,00	0,00	0,02	0,25	0,01	0,25	0,00	0,00	0,02	0,25
0,00	0,20	0,00	0,00	0,00	0,20	0,13	0,80	0,03	0,20	0,00	0,00	0,00	0,00	0,04	0,40	0,03	0,40	0,00	0,00	0,04	0,40
0,05	0,17	0,00	0,00	0,04	0,17	0,13	0,67	0,03	0,17	0,00	0,00	0,00	0,00	0,07	0,50	0,04	0,50	0,00	0,00	0,04	0,33
0,05	0,14	0,00	0,00	0,04	0,14	0,13	0,57	0,03	0,14	0,00	0,00	0,00	0,00	0,09	0,57	0,06	0,67	0,00	0,00	0,06	0,43
0,05	0,13	0,00	0,00	0,04	0,13	0,13	0,50	0,03	0,13	0,00	0,00	0,00	0,00	0,11	0,63	0,07	0,73	0,00	0,00	0,08	0,50
0,05	0,11	0,00	0,00	0,04	0,11	0,13	0,44	0,06	0,22	0,00	0,00	0,00	0,00	0,13	0,67	0,08	0,87	0,00	0,00	0,08	0,44
0,05	0,10	0,11	0,10	0,04	0,10	0,13	0,40	0,06	0,20	0,00	0,00	0,00	0,00	0,13	0,60	0,10	0,70	0,00	0,00	0,08	0,40
0,05	0,09	0,11	0,09	0,04	0,09	0,13	0,36	0,06	0,18	0,00	0,00	0,00	0,00	0,13	0,55	0,10	0,64	0,00	0,00	0,08	0,36
0,05	0,08	0,11	0,08	0,04	0,08	0,17	0,42	0,09	0,25	0,00	0,00	0,00	0,00	0,13	0,50	0,11	0,67	0,00	0,00	0,08	0,33
0,05	0,08	0,11	0,08	0,04	0,08	0,17	0,38	0,13	0,31	0,00	0,00	0,00	0,00	0,13	0,46	0,13	0,69	0,00	0,00	0,10	0,38
0,05	0,07	0,11	0,07	0,04	0,07	0,17	0,36	0,13	0,29	0,00	0,00	0,00	0,00	0,13	0,43	0,13	0,64	0,00	0,00	0,10	0,36
0,05	0,07	0,11	0,07	0,04	0,07	0,17	0,33	0,13	0,27	0,00	0,00	0,00	0,00	0,13	0,40	0,14	0,67	0,00	0,00	0,10	0,33
0,05	0,06	0,11	0,06	0,04	0,06	0,17	0,31	0,13	0,25	0,00	0,00	0,00	0,00	0,13	0,38	0,15	0,69	0,00	0,00	0,10	0,31
0,05	0,06	0,11	0,06	0,04	0,06	0,17	0,29	0,13	0,24	0,00	0,00	0,00	0,00	0,13	0,35	0,15	0,65	0,00	0,00	0,12	0,35

0,05	0,06	0,11	0,06	0,04	0,06	0,17	0,28	0,13	0,22	0,00	0,00	0,00	0,00	0,13	0,33	0,17	0,67	0,00	0,00	0,14	0,39
0,05	0,05	0,11	0,05	0,04	0,05	0,17	0,26	0,16	0,26	0,00	0,00	0,00	0,00	0,13	0,32	0,17	0,63	0,00	0,00	0,14	0,37
0,05	0,05	0,11	0,05	0,04	0,05	0,17	0,25	0,19	0,30	0,00	0,00	0,00	0,00	0,13	0,30	0,18	0,65	0,00	0,00	0,16	0,40

Tabla 5.9: Valores de exhaustividad y precisión para las búsquedas realizadas en el sistema OLE/Terra

B14		B16		B04		B12		B01		B09		B19		Bem		B11		B08		B05	
E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	1,00
0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,04	1,00
0,05	0,33	0,00	0,00	0,00	0,00	0,07	0,67	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,33	0,01	0,33	0,00	0,00	0,04	0,67
0,05	0,25	0,00	0,00	0,00	0,00	0,10	0,75	0,03	0,25	0,00	0,00	0,00	0,00	0,06	0,25	0,03	0,50	0,00	0,00	0,04	0,50
0,05	0,20	0,00	0,00	0,00	0,00	0,10	0,60	0,06	0,40	0,00	0,00	0,00	0,00	0,06	0,20	0,04	0,60	0,00	0,00	0,06	0,60
0,05	0,17	0,00	0,00	0,00	0,00	0,10	0,50	0,06	0,33	0,00	0,00	0,00	0,00	0,06	0,17	0,04	0,50	0,00	0,00	0,06	0,50
0,05	0,14	0,00	0,00	0,00	0,00	0,13	0,57	0,06	0,29	0,00	0,00	0,00	0,00	0,06	0,14	0,04	0,43	0,00	0,00	0,06	0,43
0,05	0,13	0,00	0,00	0,00	0,00	0,13	0,50	0,06	0,25	0,00	0,00	0,00	0,00	0,06	0,13	0,04	0,38	0,00	0,00	0,08	0,50
0,05	0,11	0,00	0,00	0,00	0,00	0,13	0,44	0,06	0,22	0,00	0,00	0,00	0,00	0,06	0,11	0,06	0,44	0,00	0,00	0,10	0,56
0,05	0,10	0,11	0,10	0,00	0,00	0,13	0,40	0,06	0,20	0,00	0,00	0,00	0,00	0,06	0,10	0,07	0,50	0,00	0,00	0,10	0,50
0,05	0,09	0,11	0,09	0,04	0,09	0,13	0,36	0,06	0,18	0,00	0,00	0,00	0,00	0,06	0,09	0,08	0,55	0,00	0,00	0,10	0,45
0,05	0,08	0,11	0,08	0,04	0,08	0,13	0,33	0,09	0,25	0,00	0,00	0,00	0,00	0,06	0,08	0,08	0,50	0,00	0,00	0,10	0,42
0,05	0,08	0,11	0,08	0,04	0,08	0,13	0,31	0,09	0,23	0,00	0,00	0,00	0,00	0,06	0,08	0,08	0,46	0,00	0,00	0,10	0,38
0,05	0,07	0,11	0,07	0,04	0,07	0,13	0,29	0,09	0,21	0,00	0,00	0,00	0,00	0,06	0,07	0,08	0,43	0,00	0,00	0,12	0,43
0,05	0,07	0,11	0,07	0,04	0,07	0,13	0,27	0,09	0,20	0,00	0,00	0,00	0,00	0,06	0,07	0,08	0,40	0,00	0,00	0,12	0,40
0,09	0,13	0,11	0,06	0,07	0,13	0,13	0,25	0,09	0,19	0,04	0,06	0,00	0,00	0,06	0,06	0,08	0,38	0,00	0,00	0,14	0,44
0,0	0,1	0,1	0,0	0,0	0,1	0,1	0,2	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,3	0,0	0,0	0,1	0,4

9	2	1	6	7	2	3	4	9	8	4	6	0	0	6	6	8	5	0	0	4	1
0,0	0,1	0,1	0,0	0,0	0,1	0,1	0,2	0,1	0,2	0,0	0,1	0,0	0,0	0,1	0,1	0,0	0,3	0,0	0,0	0,1	0,4
9	1	1	6	7	1	3	2	3	2	7	1	0	0	1	1	8	3	0	0	6	4
0,0	0,1	0,1	0,0	0,0	0,1	0,1	0,2	0,1	0,2	0,0	0,1	0,0	0,0	0,1	0,1	0,1	0,3	0,0	0,0	0,1	0,4
9	1	1	5	7	1	3	1	3	1	7	1	0	0	7	6	0	7	0	0	8	7
0,0	0,1	0,1	0,0	0,0	0,1	0,1	0,2	0,1	0,2	0,1	0,1	0,0	0,0	0,2	0,2	0,1	0,4	0,0	0,0	0,1	0,4
9	0	1	5	7	0	3	0	3	0	1	5	0	0	2	0	1	0	0	0	8	5

Tabla 5.10: Valores de exhaustividad y precisión para las búsquedas realizadas en el sistema Ozu

B14		B16		B04		B12		B01		B09		B19		Bem		B11		B08		B05	
E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	1,00	0,00	0,00	0,00	0,00	0,02	1,00	0,00	0,00	0,00	0,00	0,02	1,00
0,00	0,00	0,00	0,00	0,04	0,50	0,00	0,00	0,06	1,00	0,04	0,50	0,00	0,00	0,04	1,00	0,01	0,50	0,00	0,00	0,04	1,00
0,05	0,33	0,00	0,00	0,07	0,67	0,00	0,00	0,06	0,67	0,04	0,33	0,00	0,00	0,04	0,67	0,03	0,67	0,00	0,00	0,06	1,00
0,05	0,25	0,00	0,00	0,11	0,75	0,00	0,00	0,06	0,50	0,04	0,25	0,00	0,00	0,04	0,50	0,04	0,75	0,00	0,00	0,06	0,75
0,05	0,20	0,00	0,00	0,14	0,80	0,00	0,00	0,06	0,40	0,04	0,20	0,00	0,00	0,04	0,40	0,04	0,60	0,33	0,20	0,06	0,60
0,05	0,17	0,00	0,00	0,18	0,83	0,00	0,00	0,06	0,33	0,07	0,33	0,00	0,00	0,04	0,33	0,04	0,50	0,33	0,17	0,06	0,50
0,09	0,29	0,11	0,14	0,21	0,86	0,00	0,00	0,06	0,29	0,11	0,43	0,00	0,00	0,07	0,43	0,04	0,43	0,33	0,14	0,08	0,57
0,14	0,38	0,11	0,13	0,25	0,88	0,00	0,00	0,06	0,25	0,11	0,38	0,25	0,13	0,07	0,38	0,06	0,50	0,33	0,13	0,08	0,50
0,14	0,33	0,11	0,11	0,29	0,89	0,03	0,11	0,06	0,22	0,11	0,33	0,50	0,22	0,09	0,44	0,06	0,44	0,33	0,11	0,08	0,44
0,14	0,30	0,11	0,10	0,32	0,90	0,07	0,20	0,09	0,30	0,11	0,30	0,50	0,20	0,09	0,40	0,06	0,40	0,33	0,10	0,08	0,40
0,14	0,27	0,11	0,09	0,32	0,82	0,20	0,27	0,09	0,27	0,11	0,27	0,50	0,18	0,09	0,36	0,06	0,36	0,33	0,09	0,10	0,45
0,14	0,25	0,11	0,08	0,32	0,75	0,10	0,25	0,09	0,25	0,11	0,25	0,50	0,17	0,09	0,33	0,06	0,33	0,33	0,08	0,10	0,42
0,14	0,23	0,22	0,15	0,32	0,69	0,10	0,21	0,09	0,23	0,14	0,31	0,50	0,15	0,09	0,31	0,06	0,31	0,33	0,08	0,10	0,38
0,14	0,21	0,22	0,14	0,36	0,71	0,10	0,21	0,09	0,21	0,18	0,36	0,50	0,14	0,09	0,29	0,06	0,29	0,33	0,07	0,12	0,43
0,14	0,20	0,22	0,13	0,36	1,27	0,10	0,20	0,09	0,20	0,21	0,40	0,50	0,13	0,09	0,27	0,06	0,27	0,33	0,07	0,12	0,40
0,14	0,19	0,33	0,19	0,36	0,63	0,10	0,19	0,09	0,19	0,21	0,38	0,50	0,13	0,09	0,25	0,06	0,25	0,33	0,06	0,12	0,38
0,14	0,18	0,33	0,18	0,36	0,59	0,13	0,24	0,09	0,18	0,25	0,41	0,50	0,12	0,09	0,24	0,06	0,24	0,33	0,06	0,12	0,35

0,14	0,17	0,33	0,17	0,39	0,61	0,13	0,22	0,09	0,17	0,25	0,39	0,50	0,11	0,09	0,22	0,06	0,22	0,33	0,06	0,12	0,39
0,18	0,21	0,33	0,16	0,43	0,63	0,13	0,21	0,09	0,16	0,29	0,37	0,50	0,11	0,09	0,21	0,06	0,21	0,33	0,05	0,14	0,42
0,18	0,20	0,33	0,15	0,43	0,60	0,13	0,20	0,09	0,15	0,29	0,40	0,50	0,10	0,09	0,20	0,06	0,20	0,33	0,05	0,14	0,40

Tabla 5.11: Valores de exhaustividad y precisión para las búsquedas realizadas en el sistema Sol

B14		B16		B04		B12		B01		B09		B19		Bem		B11		B08		B05	
E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	P	E	P	E
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,50	0,00	0,00	0,00	0,00	0,00	0,50	0,00	0,50	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,33	0,00	0,00	0,00	0,00	0,00	0,33	0,00	0,33	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,25	0,00	0,00	0,25	0,25	0,00	0,25	0,00	0,25	0,00	0,00	0,00	0,00
0,05	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,20	0,00	0,00	0,25	0,20	0,00	0,20	0,00	0,20	0,00	0,00	0,00	0,00
0,05	0,17	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,17	0,00	0,00	0,25	0,17	0,00	0,17	0,00	0,33	0,00	0,00	0,00	0,00
0,05	0,14	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,14	0,00	0,00	0,25	0,14	0,00	0,14	0,00	0,29	0,00	0,00	0,00	0,00
0,05	0,13	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,25	0,00	0,00	0,25	0,13	0,00	0,13	0,00	0,25	0,00	0,00	0,00	0,00
0,09	0,22	0,00	0,00	0,00	0,00	0,00	0,00	0,09	0,33	0,00	0,00	0,25	0,11	0,00	0,11	0,00	0,22	0,00	0,00	0,00	0,00
0,14	0,30	0,00	0,00	0,00	0,00	0,00	0,00	0,09	0,30	0,00	0,00	0,25	0,10	0,00	0,10	0,00	0,20	0,00	0,00	0,00	0,00
0,14	0,27	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,36	0,00	0,00	0,25	0,09	0,00	0,09	0,00	0,18	0,00	0,00	0,00	0,00
0,14	0,25	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,33	0,00	0,00	0,25	0,08	0,00	0,08	0,00	0,17	0,00	0,00	0,00	0,00
0,14	0,23	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,31	0,00	0,00	0,25	0,08	0,00	0,15	0,00	0,23	0,00	0,00	0,00	0,00
0,14	0,21	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,29	0,00	0,00	0,25	0,07	0,00	0,14	0,00	0,29	0,00	0,00	0,00	0,00
0,14	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,27	0,00	0,00	0,25	0,07	0,00	0,13	0,00	0,33	0,00	0,00	0,00	0,00
0,14	0,19	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,25	0,00	0,00	0,25	0,06	0,00	0,13	0,00	0,38	0,00	0,00	0,00	0,00
0,14	0,18	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,24	0,00	0,00	0,25	0,06	0,00	0,12	0,10	0,41	0,00	0,00	0,00	0,00

0,14	0,17	0,00	0,00	0,00	0,00	0,00	0,00	0,15	0,28	0,00	0,00	0,25	0,06	0,04	0,11	0,10	0,39	0,00	0,00	0,00	0,00
0,14	0,16	0,00	0,00	0,00	0,00	0,00	0,00	0,15	0,26	0,00	0,00	0,25	0,05	0,04	0,11	0,11	0,42	0,00	0,00	0,00	0,00
0,14	0,15	0,00	0,00	0,00	0,00	0,00	0,00	0,15	0,25	0,00	0,00	0,25	0,05	0,04	0,10	0,13	0,45	0,00	0,00	0,00	0,00

Tabla 5.12: Valores de exhaustividad y precisión para las búsquedas realizadas en el sistema Ya

B14		B16		B04		B12		B01		B09		B19		Bem		B11		B08		B05	
E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	1,00
0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,50	0,03	0,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,04	1,00
0,00	0,00	0,00	0,00	0,00	0,00	0,07	0,67	0,03	0,33	0,00	0,00	0,00	0,00	0,02	0,33	0,01	0,33	0,00	0,00	0,06	1,00
0,00	0,00	0,00	0,00	0,04	0,25	0,10	0,75	0,03	0,25	0,00	0,00	0,00	0,00	0,02	0,25	0,03	0,50	0,00	0,00	0,08	1,00
0,00	0,00	0,00	0,00	0,04	0,20	0,10	0,60	0,06	0,40	0,00	0,00	0,00	0,00	0,02	0,20	0,03	0,40	0,00	0,00	0,08	0,80
0,00	0,00	0,00	0,00	0,04	0,17	0,10	0,50	0,06	0,33	0,00	0,00	0,00	0,00	0,02	0,17	0,03	0,33	0,00	0,00	0,08	0,67
0,00	0,00	0,00	0,00	0,04	0,14	0,13	0,57	0,06	0,29	0,00	0,00	0,00	0,00	0,02	0,14	0,03	0,29	0,00	0,00	0,08	0,57
0,00	0,00	0,00	0,00	0,04	0,13	0,13	0,50	0,06	0,25	0,00	0,00	0,00	0,00	0,04	0,25	0,04	0,38	0,00	0,00	0,08	0,50
0,00	0,00	0,00	0,00	0,04	0,11	0,13	0,43	0,06	0,22	0,00	0,00	0,00	0,00	0,04	0,22	0,04	0,33	0,00	0,00	0,08	0,44
0,00	0,00	0,00	0,00	0,04	0,10	0,13	0,43	0,06	0,20	0,00	0,00	0,00	0,00	0,07	0,30	0,06	0,40	0,03	0,10	0,10	0,50
0,00	0,00	0,00	0,00	0,04	0,09	0,13	0,36	0,06	0,18	0,00	0,00	0,00	0,00	0,07	0,27	0,06	0,36	0,03	0,09	0,10	0,45
0,00	0,00	0,00	0,00	0,04	0,08	0,13	0,33	0,09	0,25	0,00	0,00	0,00	0,00	0,07	0,25	0,06	0,33	0,03	0,08	0,12	0,50
0,00	0,00	0,00	0,00	0,04	0,08	0,13	0,31	0,09	0,23	0,00	0,00	0,00	0,00	0,07	0,23	0,07	0,38	0,03	0,08	0,14	0,54
0,00	0,00	0,00	0,00	0,04	0,07	0,13	0,29	0,09	0,21	0,00	0,00	0,00	0,00	0,07	0,21	0,08	0,43	0,03	0,07	0,14	0,50
0,00	0,00	0,00	0,00	0,04	0,07	0,13	0,27	0,09	0,20	0,00	0,00	0,00	0,00	0,07	0,20	0,10	0,47	0,03	0,07	0,14	0,47
0,00	0,00	0,00	0,00	0,04	0,06	0,13	0,25	0,09	0,19	0,00	0,00	0,00	0,00	0,09	0,25	0,11	0,50	0,03	0,06	0,14	0,44
0,00	0,00	0,00	0,00	0,04	0,06	0,13	0,24	0,12	0,24	0,00	0,00	0,00	0,00	0,11	0,29	0,13	0,53	0,03	0,06	0,14	0,41

0,00	0,00	0,00	0,00	0,04	0,06	0,13	0,22	0,15	0,28	0,04	0,06	0,00	0,00	0,13	0,33	0,14	0,56	0,33	0,06	0,16	0,44
0,00	0,00	0,00	0,00	0,04	0,05	0,13	0,21	0,15	0,26	0,04	0,05	0,00	0,00	0,15	0,37	0,15	0,58	0,33	0,05	0,16	0,42
0,00	0,00	0,00	0,00	0,04	0,05	0,13	0,20	0,15	0,25	0,04	0,05	0,00	0,00	0,17	0,40	0,15	0,55	0,33	0,05	0,18	0,45

Tabla 5.13: Valores de exhaustividad y precisión para las búsquedas realizadas en el sistema Yahoo

B14		B16		B04		B12		B01		B09		B19		Bem		B11		B08		B05	
E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P
0,05	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,25	1,00	0,02	1,00	0,01	1,00	0,00	0,00	0,02	1,00
0,05	0,50	0,11	0,50	0,04	0,50	0,00	0,00	0,00	0,00	0,00	0,00	0,25	0,50	0,02	0,50	0,01	0,50	0,00	0,00	0,02	0,50
0,05	0,33	0,11	0,33	0,07	0,67	0,00	0,00	0,00	0,00	0,00	0,00	0,25	0,33	0,02	0,33	0,03	0,67	0,00	0,00	0,04	0,67
0,09	0,50	0,11	0,25	0,11	0,75	0,00	0,00	0,03	0,25	0,00	0,00	0,25	0,25	0,02	0,25	0,03	0,50	0,00	0,00	0,06	0,75
0,09	0,40	0,11	0,20	0,14	0,80	0,00	0,00	0,06	0,40	0,04	0,20	0,25	0,20	0,02	0,20	0,03	0,40	0,00	0,00	0,06	0,60
0,09	0,33	0,11	0,17	0,14	0,67	0,00	0,00	0,06	0,33	0,07	0,33	0,25	0,17	0,02	0,17	0,03	0,33	0,00	0,00	0,06	0,50
0,09	0,29	0,11	0,14	0,18	0,86	0,03	0,14	0,09	0,43	0,07	0,29	0,25	0,14	0,02	0,14	0,03	0,29	0,00	0,00	0,06	0,43
0,14	0,38	0,11	0,13	0,18	0,75	0,07	0,29	0,09	0,38	0,11	0,38	0,25	0,13	0,02	0,13	0,04	0,48	0,00	0,00	0,06	0,38
0,14	0,33	0,11	0,11	0,18	0,67	0,03	0,43	0,09	0,33	0,14	0,44	0,25	0,11	0,02	0,11	0,04	0,43	0,00	0,00	0,08	0,44
0,14	0,30	0,11	0,10	0,21	0,70	0,03	0,57	0,12	0,40	0,14	0,40	0,25	0,10	0,02	0,10	0,04	0,40	0,00	0,00	0,08	0,40
0,14	0,27	0,11	0,09	0,25	0,73	0,33	0,50	0,15	0,45	0,18	0,45	0,25	0,09	0,02	0,09	0,04	0,47	0,00	0,00	0,08	0,36
0,14	0,25	0,11	0,08	0,25	0,67	0,33	0,63	0,15	0,42	0,18	0,42	0,25	0,08	0,02	0,08	0,04	0,45	0,00	0,00	0,10	0,42
0,14	0,23	0,11	0,08	0,25	0,62	0,33	0,75	0,15	0,38	0,21	0,46	0,25	0,08	0,02	0,08	0,04	0,43	0,00	0,00	0,12	0,46
0,14	0,21	0,22	0,14	0,25	0,57	0,33	0,88	0,15	0,36	0,25	0,50	0,25	0,07	0,02	0,07	0,04	0,41	0,00	0,00	0,12	0,43
0,18	0,27	0,22	0,13	0,25	0,53	0,33	0,27	0,15	0,33	0,29	0,53	0,25	0,07	0,02	0,07	0,04	0,40	0,00	0,00	0,14	0,47
0,18	0,25	0,22	0,13	0,25	0,50	0,17	0,31	0,15	0,31	0,29	0,50	0,25	0,06	0,02	0,06	0,04	0,19	0,00	0,00	0,14	0,44
0,18	0,24	0,22	0,12	0,29	0,53	0,17	0,29	0,15	0,29	0,32	0,53	0,25	0,06	0,02	0,06	0,04	0,18	0,00	0,00	0,14	0,41

Rendimiento

0,18	0,22	0,22	0,11	0,29	0,50	0,20	0,28	0,15	0,28	0,32	0,50	0,25	0,06	0,02	0,06	0,04	0,17	0,00	0,00	0,14	0,39
0,18	0,21	0,22	0,11	0,29	0,47	0,23	0,37	0,15	0,26	0,32	0,47	0,25	0,05	0,02	0,05	0,04	0,16	0,33	0,05	0,14	0,37
0,18	0,20	0,22	0,10	0,29	0,45	0,23	0,35	0,18	0,30	0,32	0,45	0,25	0,05	0,02	0,05	0,04	0,15	0,33	0,05	0,16	0,40

La tabla 5.14, a continuación, resume los valores hallados promediando las cifras de exhaustividad y precisión para los siete sistemas y cada nivel de resultados en cada uno de los 11 temas de búsqueda.

Tabla 5.14: Valores promedio de E-P de los sistemas analizados

Altavista		Enlaweb		Lycos		OLE		Ozu		Sol		Ya		Yahoo	
E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P
0,01	0,36	0,00	0,09	0,00	0,18	0,00	0,09	0,01	0,27	0,00	0,09	0,00	0,18	0,03	0,45
0,02	0,27	0,00	0,05	0,01	0,27	0,01	0,14	0,02	0,41	0,01	0,14	0,01	0,18	0,05	0,32
0,03	0,33	0,01	0,12	0,02	0,21	0,02	0,21	0,03	0,39	0,01	0,09	0,02	0,24	0,05	0,30
0,04	0,39	0,01	0,16	0,02	0,18	0,03	0,23	0,04	0,34	0,02	0,09	0,03	0,27	0,06	0,32
0,05	0,35	0,01	0,15	0,03	0,24	0,03	0,24	0,07	0,31	0,03	0,09	0,03	0,24	0,07	0,31
0,05	0,32	0,02	0,15	0,04	0,23	0,03	0,20	0,08	0,29	0,04	0,09	0,03	0,20	0,08	0,27
0,05	0,29	0,02	0,16	0,04	0,23	0,04	0,18	0,10	0,32	0,04	0,08	0,03	0,18	0,09	0,29
0,05	0,28	0,03	0,17	0,05	0,24	0,04	0,17	0,13	0,33	0,04	0,08	0,04	0,18	0,10	0,30
0,06	0,25	0,03	0,16	0,05	0,24	0,04	0,17	0,16	0,33	0,04	0,09	0,04	0,16	0,10	0,30
0,08	0,25	0,03	0,15	0,06	0,24	0,05	0,17	0,17	0,33	0,04	0,09	0,07	0,18	0,11	0,31
0,08	0,23	0,04	0,14	0,06	0,21	0,06	0,17	0,18	0,31	0,04	0,09	0,07	0,17	0,12	0,30
0,08	0,22	0,04	0,13	0,07	0,22	0,06	0,17	0,18	0,29	0,04	0,08	0,08	0,17	0,13	0,30
0,08	0,21	0,04	0,14	0,08	0,22	0,06	0,15	0,19	0,28	0,06	0,09	0,08	0,17	0,13	0,31
0,09	0,21	0,05	0,15	0,08	0,21	0,06	0,15	0,20	0,28	0,06	0,09	0,08	0,16	0,14	0,31
0,10	0,22	0,05	0,15	0,08	0,20	0,06	0,14	0,20	0,32	0,06	0,09	0,08	0,16	0,15	0,26
0,10	0,21	0,06	0,15	0,08	0,19	0,07	0,15	0,21	0,26	0,06	0,09	0,08	0,16	0,16	0,25
0,11	0,22	0,06	0,14	0,08	0,19	0,07	0,14	0,22	0,25	0,06	0,09	0,09	0,17	0,16	0,25
0,13	0,23	0,06	0,14	0,08	0,19	0,09	0,16	0,22	0,25	0,07	0,09	0,10	0,18	0,17	0,23
0,14	0,24	0,06	0,14	0,09	0,18	0,10	0,16	0,23	0,25	0,07	0,09	0,11	0,18	0,20	0,23
0,14	0,24	0,07	0,14	0,09	0,19	0,11	0,17	0,23	0,24	0,07	0,09	0,11	0,18	0,20	0,23

La figura 5.6, por su parte, compara los trazados correspondientes a estas distribuciones.

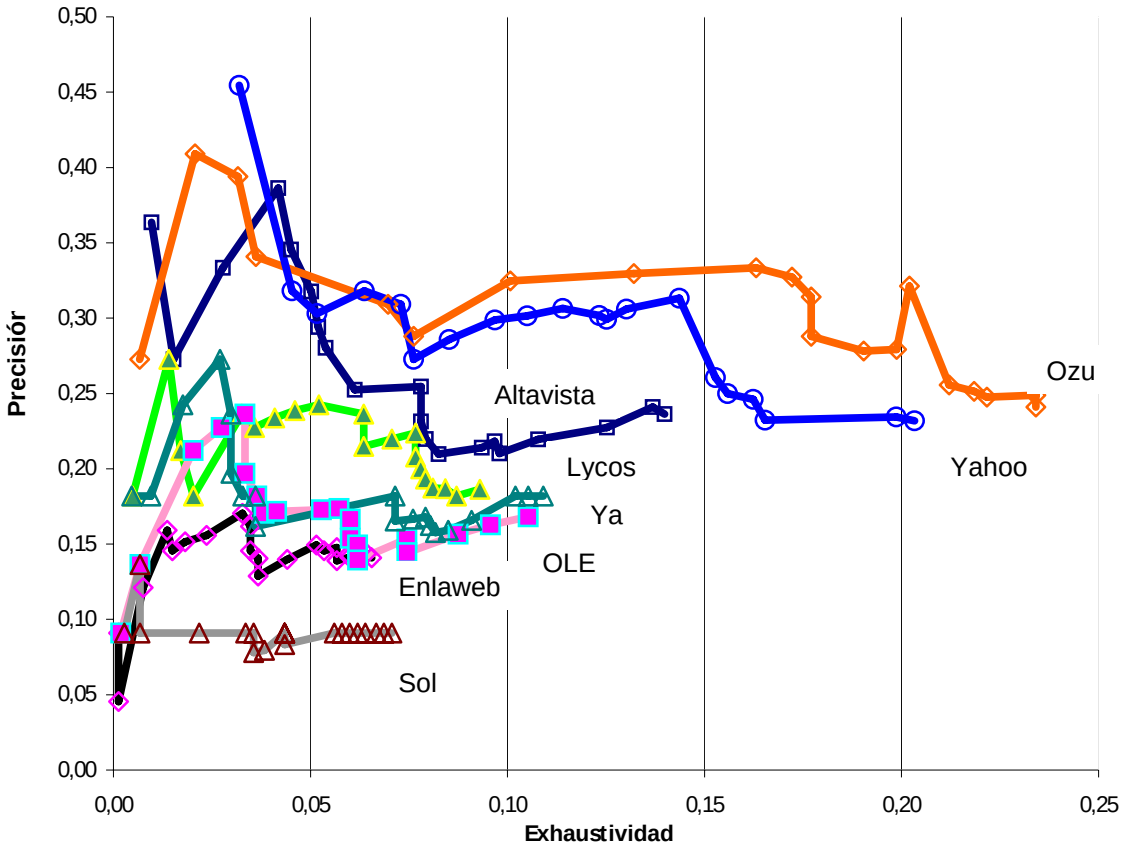


Figura 5.6: Diagrama general con los valores promedio de exhaustividad vs. precisión para los 8 sistemas analizados

El rendimiento de los sistemas, expresado en términos de ehxaustividad y precisión, es reducido, con valores promedio de exahustividad situados entre el 7 (Altavista) y el 14% (Ozú) y de precisión entre el 9 (Sol) y el 30% (Ozú). Sólo Yahoo muestra un comportamiento típico, con relación inversa entre los valores a lo largo de los 20 resultados de búsqueda. Los sistemas restantes muestran invariablemente un ascenso en las cifras de precisión a

partir del segundo o tercer resultado de búsqueda, lo que abunda en la idea de defectos en el algoritmo de ordenación de los resultados obtenidos. La apariencia “contraída” del trazado correspondiente a EnlaWeb se justifica por su naturaleza de directorio y su consecuente mecanismo clasificatorio: a partir del ascenso inicial en los valores de precisión, se produce una estabilización del cociente. En este sentido, la horizontalidad de extensos tramos en algunos trazados (Ozú y Yahoo), con estancamiento de los valores de precisión y el posterior remonte, vuelven a ser indicativos del funcionamiento anómalo del algoritmo de ordenación de resultados.

Para avanzar en el conocimiento de estos resultados, fue necesario analizar los componentes implicados y su interacción: el tema de búsqueda, aquí denominado indistintamente tema o búqueda; el sistema y el orden de aparición de los resultados.

Se llevó a cabo un análisis de la varianza (ANOVA) de estos tres factores. Los resultados se muestran en las tablas 5.15 y 5.16. Indican que las diferencias de rendimiento dependen en gran medida del sistema empleado en las búsquedas. Por lo que respecta a la exhaustividad, el orden de presentación de los resultados también influye. El factor denominado “tema”, que en realidad se refiere tanto al tema de búsqueda como al usuario que lo ha propuesto y ha juzgado los resultados, tiene cierta importancia en relación con la exhaustividad, pero sobre todo es determinante de la precisión alcanzada. Así mismo, tiene relevancia la interacción entre sistema y tema. En otras palabras: el juicio de relevancia y el hecho de que determinados temas de búsqueda se ajusten más a la especial mecánica de la recuperación de un sistema son factores determinantes del rendimiento global.

Tabla 5.15: Cuadro resumen de ANOVA de tres factores para los resultados sobre la exhaustividad de los sistemas analizados

	SS	GL	MS	F
Total	13,9511			
Sistema	2,1889	7	0,3127	144,1807

Orden	2,2269	19	0,172	54,0401
Tema	0,1973	10	0,0197	9,0966
Sistema/Orden	0,5906	133	0,0044	2,0475
Sistema/Tema	5,7218	70	0,0817	37,6892
Orden/Tema	0,1411	190	0,0007	0,3424
Residual	2,8845	1330	0,0022	

Tabla 5.16: Cuadro resumen de ANOVA de tres factores para los resultados sobre la precisión de los sistemas analizados

	SS	GL	MS	F
Total	98,0188			
Sistema	8,7992	7	1,257	74,0358
Orden	0,6856	19	0,036 1	2,1254
Tema	31,5033	10	3,150 3	185,547
Sistema/Orden	1,6527	133	0,012 4	0,7319
Sistema/Tema	29,588	70	0,422 7	24,8952
Orden/Tema	3,2094	190	0,016 9	0,9949
Residual	22,5815	1330	0,017	

En ambas tablas, MS representa los cuadrados medios de cada factor, SS su suma, y GL indica los grados de libertad. F, el cociente entre cada MS y el MS residual, se ha calculado para $p > 0,95$ (Hald, 1952).

La figura 5.7 presenta un perfil del rendimiento de los sistemas en términos de exhaustividad y precisión. Pone en evidencia que, tanto en términos de exhaustividad como de precisión, los sistemas

Ozú y Yahoo destacan sobre el resto, seguidos por las puntuaciones de Altavista. No existen grandes diferencias en la exhaustividad obtenida por los restantes sistemas, pero se puede observar que Sol presenta una precisión (0,092 = 9%) especialmente reducida.

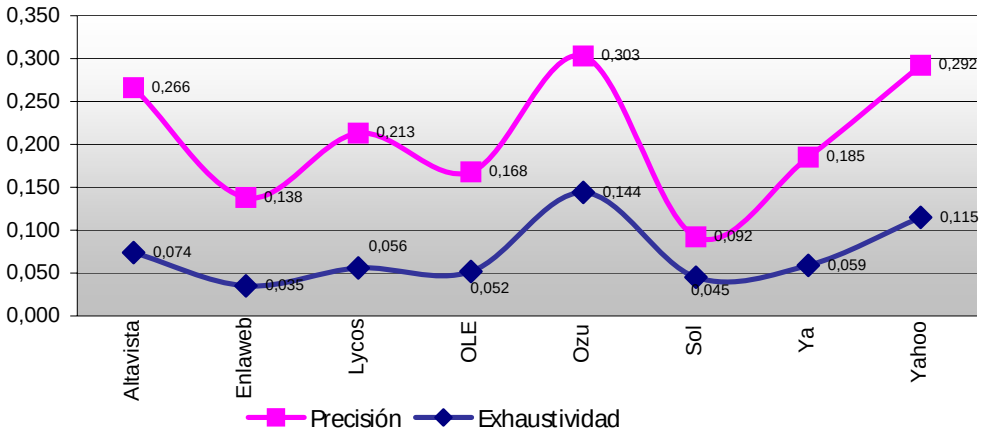


Figura 5.7: Valores promedio de exhaustividad y precisión de los 8 sistemas analizados

En relación con el tema de búsqueda, se observa que la mejor precisión se obtiene en las búsquedas 9 y 11, mientras que las 4 y 7, con cifras muy bajas en precisión, arrojan mejores resultados en exhaustividad (Figura 5.8). Los resultados de las búsquedas 2 y 10 son claramente inferiores tanto en precisión como en exhaustividad, aunque se aprecia que éste último indicador se mueve en un reducido rango, entre el 5 y el 10%.

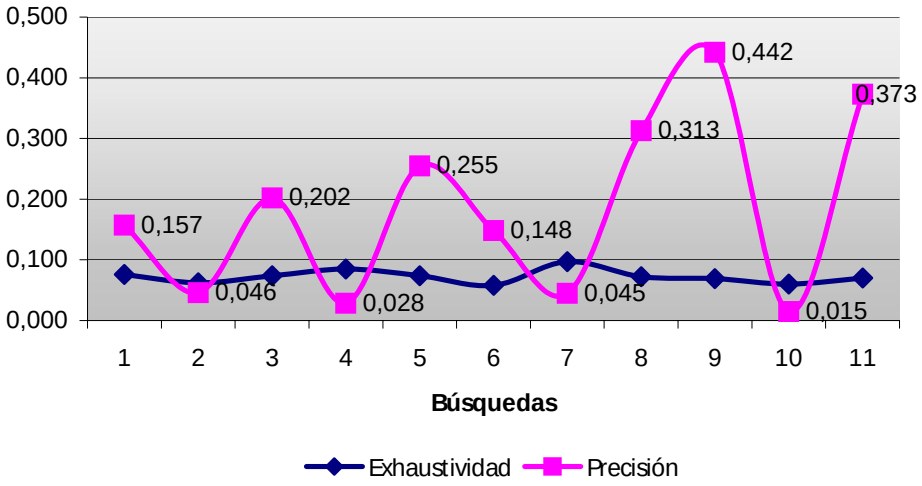


Figura 5.8: Valores promedio de exhaustividad y precisión en las búsquedas evaluadas

La interacción entre sistema y búsqueda es muy importante, tal y como revelaban los valores de F en las tablas 5.15 y 5.16. Los trazados de las figuras 5.9 y 5.10 muestran la exhaustividad y precisión media de los sistemas en función de las 11 búsquedas evaluadas. Los sistemas Ozú y Yahoo muestran los mejores resultados de precisión en las búsquedas 3 y 11. El directorio EnlaWeb sólo destaca en la búsqueda 9, donde los resultados de todos los sistemas experimentan una mejoría en su precisión. Podría referirse este hallazgo como dependiente de un juicio de relevancia especialmente generoso. Algo similar sucede en la búsqueda 11, pero en este caso se deben exceptuar el directorio y el sistema Sol.

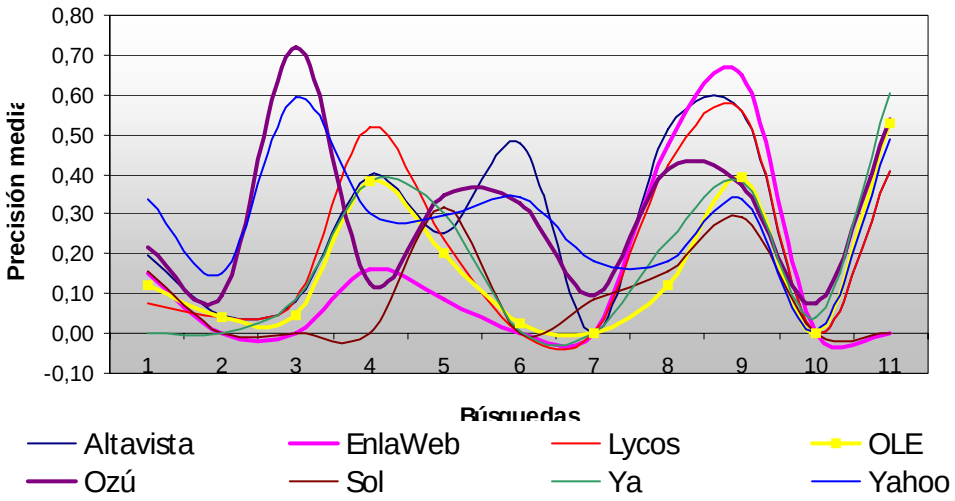


Figura 5.9: Interacción entre sistema y tema de búsqueda en los valores de precisión

En cuanto a la exhaustividad, también son destacables Yahoo y Ozú, con valores máximos en las búsquedas 3, 7 y 11.

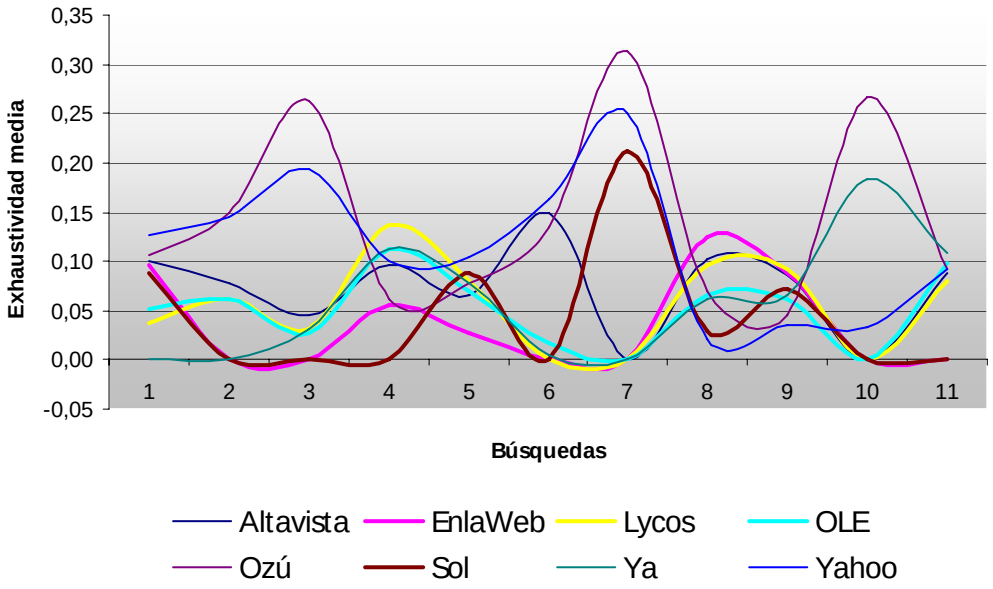


Figura 5.10: Interacción entre sistema y tema de búsqueda en los valores de exhaustividad

La prueba de comparación múltiple de Tukey (Winer, 1962) indica que existen diferencias significativas entre los sistemas.

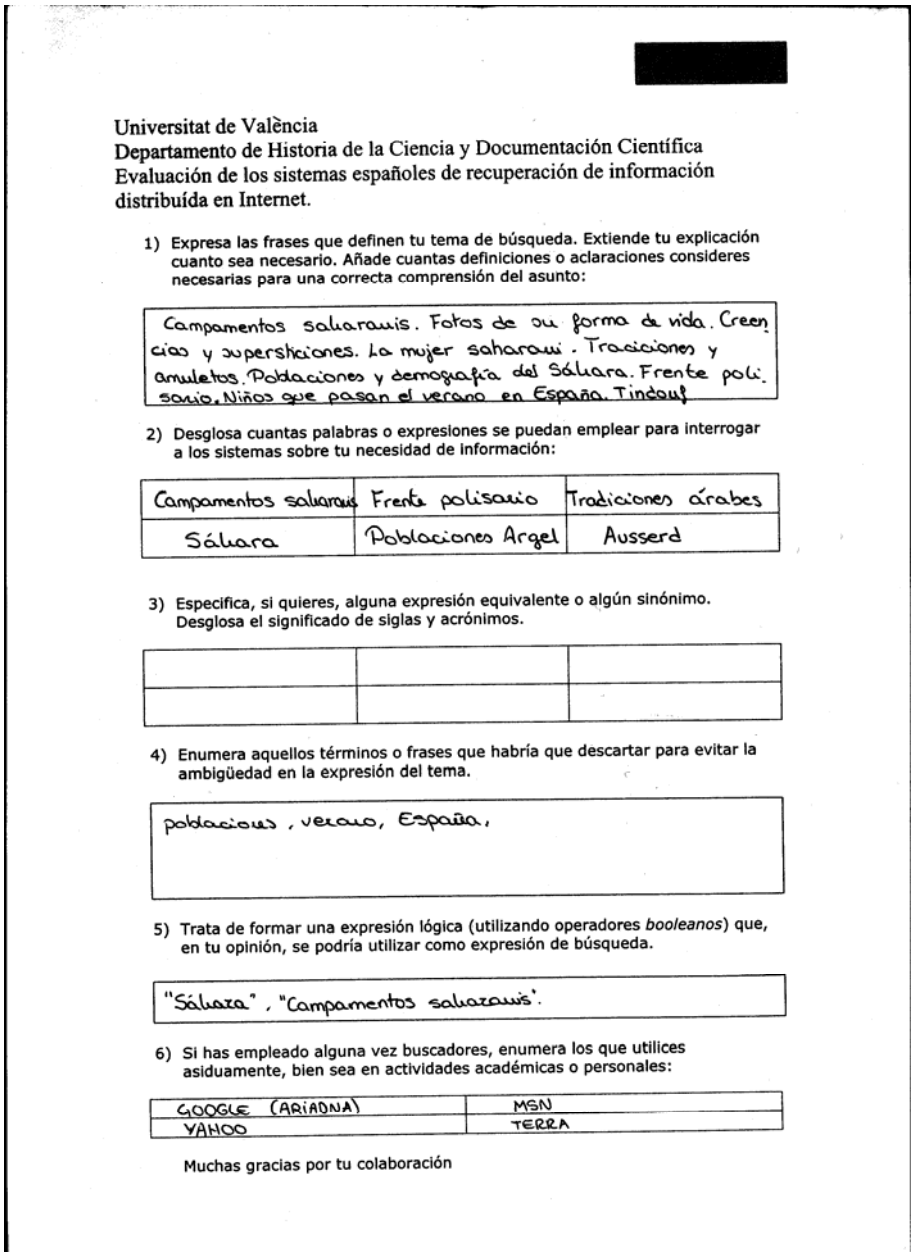
Los escasos valores de cobertura y aportación específica de los sistemas justificarían el diseño de sistemas de búsquedas simultáneas en paralelo (los que se vienen denominando popularmente metabuscadores). La desaparición de elindice.com supuso también la de metaindice, tal y como se comentó en 4.1.1.

La cortedad de la serie de búsquedas evaluadas no permite aventurar conclusiones sólidas sobre el rendimiento, habida cuenta de la interacción entre sistemas y búsquedas (o juicios de los usuarios) evidenciada por el análisis de varianza. A pesar de esto y según los resultados del experimento, se puede afirmar que la eficacia de la recuperación es reducida, aunque existen diferencias estadísticamente significativas en los valores de exhaustividad y precisión entre los 8 sistemas analizados. El rendimiento en la recuperación parece más estrechamente relacionado con la función

de similitud y el cálculo de la relevancia propios de cada sistema, pero también resulta evidente que la recuperación de determinados temas en determinados sistemas ofrece mejor rendimiento, aunque haya otros para los cuales resulte indiferente el sistema empleado.

En la formulación de objetivos se esgrimían una serie de argumentos para justificar el estudio del rendimiento de los sistemas de recuperación de información distribuída en Internet. De ellos, uno se refería a la posibilidad de integrar los sistemas de búsqueda en redes corporativas o institucionales, especialmente aquellas ligadas al desarrollo del conocimiento. No parece que los resultados del presente estudio favorezcan el empleo de los métodos de los sistemas analizados en esos ámbitos.

Anejo 5.1: Facsímil del formulario de búsqueda cumplimentado por uno de los participantes en el estudio.



[redacted]

Universitat de València
Departamento de Historia de la Ciencia y Documentación Científica
Evaluación de los sistemas españoles de recuperación de información distribuída en Internet.

1) Expresa las frases que definen tu tema de búsqueda. Extiende tu explicación cuanto sea necesario. Añade cuantas definiciones o aclaraciones consideres necesarias para una correcta comprensión del asunto:

Campamentos saharauis. Fotos de su forma de vida. Creencias y supersticiones. La mujer saharaui. Tradiciones y amuletos. Poblaciones y demografía del Sáhara. Frente polisario. Niños que pasan el verano en España. Tindouf

2) Desglosa cuantas palabras o expresiones se puedan emplear para interrogar a los sistemas sobre tu necesidad de información:

Campamentos saharauis	Frente polisario	Tradiciones árabes
Sáhara	Poblaciones Argel	Ausserd

3) Especifica, si quieres, alguna expresión equivalente o algún sinónimo. Desglosa el significado de siglas y acrónimos.

4) Enumera aquellos términos o frases que habría que descartar para evitar la ambigüedad en la expresión del tema.

poblaciones, verano, España.

5) Trata de formar una expresión lógica (utilizando operadores *booleanos*) que, en tu opinión, se podría utilizar como expresión de búsqueda.

"Sáhara", "Campamentos saharauis".

6) Si has empleado alguna vez buscadores, enumera los que utilices asiduamente, bien sea en actividades académicas o personales:

GOOGLE (ARIADNA)	MSN
YAHOO	TERRA

Muchas gracias por tu colaboración

6. Conclusiones y perspectivas

La propia evolución técnica y social de Internet ha originado su inestabilidad, su ingente volumen y su heterogeneidad informativa y formal. Las tres características han originado limitaciones en la accesibilidad de los recursos distribuidos.

Los sistemas que se han desarrollado para afrontar la búsqueda y recuperación de información en Internet, centrados inicialmente en espacios definidos, han evolucionado hasta intentar abarcar el espacio global. A las listas y esquemas clasificatorios iniciales se ha ido superponiendo sistemas de recopilación automática que, en sus últimas versiones, combinan la recuperación sintáctica y un tratamiento algorítmico de elementos contextuales.

La accesibilidad de los documentos distribuidos es muy variable; hasta el punto de que cabe sustituir la distribución de espacios informativos basada en las diferencias de protocolos y formatos por una distribución de espacios basada en la accesibilidad de la información de los documentos que contienen. Desde este punto de vista, los sistemas de recuperación basados en agentes pueden constituir la solución más adecuada para el acceso a la información y los documentos distribuidos en Internet.

El enfoque global y los objetivos de este trabajo se han basado en un esquema hasta cierto punto tradicional en el área de la investigación de la recuperación de información. En primer lugar se ha analizado la línea de los documentos, a través de la caracterización de una muestra del espacio Web español. En segundo lugar se ha estudiado el esquema de datos y las características operativas de los sistemas españoles de recuperación de información en Internet. En tercer lugar, se ha determinado el rendimiento de la recuperación a través de esos mismos sistemas.

La inestabilidad de los documentos y de los propios sistemas limita hasta cierto punto los resultados. Los resultados de cada serie de experimentos están limitados al plazo temporal en que se realizaron[1]. A pesar de esto, cabe formular conclusiones acerca de

[1] Una cronología aproximada sería la siguiente: 4 de enero de 2001: obtención de la muestra de sedes; marzo-abril de 2001: encuesta; mayo de 2001:

los tres apartados y, además, valorar si el conjunto de métodos empleados podría constituir un modelo analítico para posteriores evaluaciones.

El espacio Web español

El conjunto de sedes que constituyen el espacio Web de España está compuesto, desde el punto de vista de su accesibilidad directa y, por tanto, de su contenido informativo, por tres sectores. En primer lugar, un número de sedes de escaso o nulo contenido informativo, de formación incipiente o establecidas con el único propósito de reservar una denominación de segundo nivel. Un segundo grupo de sedes, que presenta un contenido predominantemente estático, poca estructuración de los documentos y un nivel técnico acorde. Un último grupo de sedes, minoritario, que concentra la mayor parte de los documentos y el contenido informativo del espacio Web español. Este último grupo presenta generación dinámica de sus contenidos, que están bien estructurados, y un alto nivel técnico. Esta distribución resulta de una mera "instantánea". Una segunda toma de datos, posterior en dos años a la primera, permite sin embargo confirmar esta configuración.

Las sedes analizadas emplean los enlaces hipertextuales con intención organizativa y alcance interno. La proporción de enlaces a otras sedes es mínima. La recopilación basada en el seguimiento de enlaces difícilmente podrá alcanzar una cobertura exhaustiva y muchas de las sedes permanecen invisibles para los sistemas de recuperación.

Todas las variables analizadas se distribuyen de forma exponencial. Así, tanto por el número de páginas como por el número de niveles, el volumen en bytes o el número de elementos, se destaca un núcleo de sedes sobre las restantes. Si se atiende a este carácter exponencial y al consiguiente "efecto Mateo", similar al de muchas distribuciones bibliométricas, es previsible un crecimiento mayor de las sedes con mayor componente dinámico en la

---

estudio del esquema de datos y de la cobertura; octubre de 2001: caracterización del espacio web español; 30 de agosto de 2002: popularidad de los sistemas; noviembre de 2002: primera fase de búsquedas; abril de 2003: segunda fase de búsquedas; septiembre de 2003: actualización de los datos de accesibilidad; febrero de 2004: determinación del rendimiento de los sistemas

generación de sus contenidos. La situación que puede resultar de este fenómeno es un futuro incremento de la proporción de contenidos generados por transacciones, de forma dinámica. Si a este desequilibrio se uniera la comercialización de los servicios y los accesos a la información, se incrementaría notablemente el componente “invisible” de la Web en España. Esta panorámica se enfrentaría con la naturaleza actual de los sistemas de recuperación, diseñados como sistemas de recopilación automática y recuperación sintáctica de páginas estáticas, y les restaría mucha utilidad como instrumentos de búsqueda y selección de recursos e información.

Características de los sistemas de recuperación

A excepción del directorio EnlaWeb, los sistemas aquí analizados son de recopilación automática y recuperación sintáctica. En relación con la primera de estas características, cabe afirmar que el nivel de cobertura de los sistemas analizados es muy bajo: sólo llegan a incorporar, de forma individual y en el mejor de los casos, la tercera parte del espacio Web español considerado. Además, situados en un entorno operativo, casi el 80% de los resultados de un conjunto de búsquedas se recuperaban a partir de un sistema.

La cortedad en la cobertura no obedece a la incapacidad para capturar contenidos de generación dinámica. Quizá sea más achacable al bajo nivel de interconexión de las sedes españolas, característica que dificulta el funcionamiento de los módulos de recopilación convencional. En cualquier caso, la visibilidad de un documento o una sede distribuidos en el espacio Web español difícilmente ha de mejorar si depende de su inclusión en los sistemas analizados.

A pesar de la mínima respuesta a los cuestionarios sobre las características operativas de los sistemas analizados, se han podido analizar algunas mediante observación directa. Este análisis revela que los sistemas analizados no representan adecuadamente los documentos que recopilan y, además, los indizan de forma limitada. En general, el esquema de datos de los sistemas contiene poca o mínima información sobre la responsabilidad de un documento, su creador y distribuidor, y sobre la relación entre el documento en

cuestión y otros. Estas limitaciones contrastan con la creciente necesidad de determinar la fiabilidad de los documentos distribuidos.

Situados frente a una escala de 1 a 19, todos los sistemas presentan puntuaciones inferiores a 10. Por otra parte, las relaciones hipertextuales entre unos y otros documentos se fían exclusivamente a la sintaxis de las respectivas URLs, y esto supone una limitación en la estructuración de resultados de búsqueda.

Los sistemas no siguen la dinámica del espacio Web español. La falta de ritmo de su módulo de recopilación explica su falta de actualización. De ella deriva el porcentaje significativo de resultados erróneos (es decir, sin conexión efectiva con los documentos fuente) obtenidos: entre un 22 y un 9 por ciento. También se puede atribuir a este defecto el número de resultados duplicados ofrecidos por cada sistema, que se situa entre el 5 y el 10 por ciento, según los sistemas.

Aunque parezca contradictorio, los elementos que se incluyen en los esquemas de datos no pasan a configurar índices en las respectivas bases de datos o, al menos, no constituyen elementos de acceso en el momento de la búsqueda. Por eso, cabe concluir que la relación entre los elementos estructurales de los documentos y las sedes, la recopilación de datos y las posibilidades de recuperación es ineficiente. El ejemplo más típico está representado por el elemento título, presente en una alta proporción de los documentos del espacio Web español analizado, contemplado en el esquema de la totalidad de los sistemas, pero descartado como elemento de acceso o método de refinamiento de resultados por muchos de ellos.

Los sistemas de recuperación analizados presentan una funcionalidad y unas posibilidades de manejo manifiestamente inferiores a los tradicionales sistemas de catalogación o recuperación de información bibliográfica.

Además, es patente la contradicción entre las cifras de sedes y páginas cubiertas esgrimidas por cada sistema y la cortedad de los datos de cobertura que apenas sobrepasan la tercera parte del espacio Web español en dos casos, con datos de aporte específico que rondan el 10% en el mejor de los cálculos.

El rendimiento de los sistemas

El rendimiento de la recuperación de los sistemas analizados es, en términos generales, reducido. Esto se refiere tanto a la recuperación como a la función de similitud en que se basa la ordenación de los resultados.

La primera conclusión deriva de las cifras de exhaustividad y precisión alcanzadas por los sistemas, con valores promedio máximos del 23 de precisión y del 45% respectivamente. Aunque se aprecian variaciones entre ellos. El cálculo de un valor discreto para la relevancia, del que deriva la ordenación de resultados, está basado en una mera función estadística que parece insuficiente. A este respecto, resulta revelador que sólo Yahoo muestre un trazado típico y de tendencia continuada en el diagrama exhaustividad-precisión. En los restantes sistemas se producen repuntes en la precisión tras un arranque que indica resultados no relevantes encabezando las listas. Aunque los valores máximos de precisión se alcanzan en los 5 primeros resultados y los valores de exhaustividad se estabilizan en el tramo 1-15, estos hallazgos son coherentes con una deficiente ordenación de los resultados de búsqueda.

La búsqueda de determinados temas ofrece mejores resultados en unos sistemas que en otros. Esta conclusión contradice la pretendida universalidad de los sistemas analizados. Por otra parte, este componente incorpora la mayor o menor generosidad en el juicio de relevancia de cada usuario individual, un factor que podría tener igual o mayor peso que la desigual orientación en la cobertura.

El censo actual (junio de 2004) de sistemas españoles de recuperación de información en Internet ofrece 36 servicios generales (no limitados regionalmente). Entre ellos no se incluye el directorio EnlaWeb, considerado inactivo desde el 22 de mayo de 2004 y ofrecen problemas de conexión Avispanet y VenyBusca. De los 34 servicios en operación, 15 se han analizado en alguna de las fases de este proyecto. Dos de los restantes son sistemas especializados en información turística o en tecnologías de Internet (SearchIberia y Navegalis, respectivamente). Se cuentan 7 sistemas de directorio. De los restantes 10 sistemas, sólo uno, la versión española del sistema Google (introducida el 25 de septiembre de 2003) emplea la recopilación automática y la recuperación contextual.

Si se tienen en cuenta el conjunto de variables estudiadas, Altavista es el sistema que se destaca sobre los demás. Presenta las

mejores cifras de aporte específico (sobre un conjunto de sistemas del que cabe recordar que 9, con nulo aporte específico, se pueden considerar prescindibles en la búsqueda de información en el espacio Web español). Además, presenta la menor proporción de solapamiento si se descuenta Enlaweb que, por su naturaleza de directorio, no incluye páginas sino sedes en su cobertura. Altavista presenta, por otra parte, una gran coherencia entre su nivel de representación de los documentos y las opciones de recuperación que ofrece a sus usuarios, y buenas cifras de rendimiento.

Los sistemas dominantes en el cálculo del rendimento, Ozu y Yahoo, emplean tecnología de recuperación de Google. Sus bases de datos son muy limitadas, según revelan las cifras de aporte específico y de solapamiento, y muestran claras divergencias entre el nivel de representación de los documento y las opciones ofrecidas a sus usuarios. Una mayor atención a la cobertura premitiría mejorar en mucho su utilidad en la recuperación de información de la web española. El sistema Ya está basado también en Google y presenta datos de rendimiento ligeramente inferiores a los anteriores y una base de datos con gran solapamiento con los restantes sistemas.

Estos tres casos permiten concluir que no basta el empleo de programas de recuperación de reconocido prestigio y probado rendimiento para garantizar un nivel de recuperación de información adecuado en el espacio Web.

El único directorio evaluado, Enlaweb, ofrece datos coherentes con su propia naturaleza. Unicamente es destacable por su bajo nivel de solapamiento, aunque este dato resulta de un artificio: el hecho de que sólo incluya en su cobertura sedes, y no páginas individualizadas. Ni su rendimiento, ni su aporte específico ni su nivel de representación justifican su empleo.

Del mismo modo, los sistemas Ole y Lycos parecen prescindibles, si se tiene en cuenta su alto grado de solapamiento. El sistema Sol añade a su mínima aportación específica el peor rendimiento en términos de exhaustividad y precisión.

Un modelo de evaluación

Desde el punto de vista de la constitución de un modelo evaluativo de los sistemas de recuperación de información en Internet, las tres fases en que se ha estructurado la presente investigación constituyen un mínimo o sólo un punto de partida. El esquema se ajusta a los requisitos demandados por la mayoría de los trabajos del área. Sin embargo, se debe completar con estudios de la línea de los usuarios. Los métodos cualitativos y el trabajo con grupos de enfoque, sólo esbozados aquí, han de completarse con análisis de transacciones que puedan evidenciar necesidades, peticiones y juicios de los usuarios de los sistemas. Este tipo de análisis se puede considerar como el paso previo a un estudio de la conducta de recuperación (*information seeking behavior*) de los usuarios. Pero ello requiere el acuerdo de los sistemas, que deberían poner los datos a disposición de las comunidades de investigadores del área. La alternativa, es decir, el análisis automático del rendimiento de los sistemas parece demasiado incipiente, aunque no exenta de interés. Igualmente incipiente resulta el empleo de colecciones de referencia (*test collections*) necesarias para proceder a estudios de evaluación que sigan el modelo de Cranfield.

Por otro lado, no se debe olvidar el carácter estático, sincrónico, del estudio de caracterización que se ha realizado. El número de “observatorios de la sociedad de la información” implantados en España y en otros ámbitos es suficiente como para que se planteen la realización de estudios diacrónicos, de seguimiento del espacio Web desde el punto de vista informativo.

Perspectivas

Las recientes intenciones de flexibilización del registro de dominios españoles (Ministerio de Ciencia y Tecnología, 2003) pueden sentar las bases para facilitar la distribución online de información en el espacio Web español. Pero para que esta información resultara realmente visible y accesible, se necesitaría un acuerdo de mayor alcance: la sindicación de contenidos entre los 24

agentes registradores españoles y los propietarios de los sistemas de recuperación. La cobertura del espacio bajo dominio .es o alguna de sus combinaciones con subdominios genéricos se vería enormemente mejorada.

Un procedimiento similar permitiría a los sistemas de recuperación ofrecer información procedente de algunas bases de datos cuyo contenido se genera automáticamente en respuesta a transacciones. El protocolo ICE, operativo desde 2000, y el desarrollo de funciones CGI y otras pasarelas y traductores posibilitarían el acceso (sin necesidad de innecesarias recargas de los índices).

Naturalmente, nada de ello ha de resultar en mejor cobertura si los módulos de recopilación no se programan al ritmo suficiente y si no se dota a los algoritmos de detección de duplicados de una mayor eficacia.

Se debería valorar la posibilidad de emplear agentes en lugar de almacenar en bases de datos locales la información procedente de páginas (obligatoriamente estáticas). Esta alternativa permitiría dotar de una tasa de refresco inmediata a los sistemas. Acaso convendría la combinación con programas de visualización de la información que ofrezcan los resultados estructurados gráficamente, en lugar de ordenarlos en una mera lista unidimensional. El empleo de este tipo de programas o de soluciones basadas en la tecnología de mapas semánticos requiere, sin embargo, del cálculo de las relaciones entre unos y otros documentos recuperados, y conduciría, en su caso, a nuevos conceptos y procedimientos en la recuperación, muy próximos a los que emplean los sistemas de recuperación contextual, e incluso más avanzados.

Ya son apreciables, por otra parte, los intentos de dotar a los módulos de indización de algoritmos que permitan incorporar el contenido de documentos textuales o mixtos en formatos propietarios, sobre todo MS Word y Adobe Acrobat.

Por lo que respecta a las condiciones operativas de la indización y la recuperación, se debería obtener un mejor ajuste entre la estructura implícita de los documentos “planos”, es decir, aquellos basados únicamente en HTML y la correspondiente arquitectura de índices. De esta forma, se ofrecería una gama mayor de puntos de acceso al usuario, al tiempo que se posibilitaba el refinamiento de búsqueda y la ganancia en relevancia.

La recuperación en los sistemas de información en Internet no puede depender, ni en la selección ni en la ordenación de resultados, de funciones basadas en meros recuentos estadísticos de términos. Una mayor atención a aspectos estructurales implícitos en los documentos, con tratamiento diferenciado de los términos de los títulos, por ejemplo, podría añadir variables de utilidad al cálculo de la función de similitud y a la ordenación de resultados. Si es cierto que el número de enlaces externos de las sedes sigue modelos exponenciales de crecimiento, existe la posibilidad de que las funciones contextuales intervengan en la recuperación o, al menos, en la ordenación de los conjuntos de documentos recuperados. Los grupos de investigación nacionales comienzan a ofrecer avances de gran aplicación a la recuperación de información distribuida. La clasificación automática, la recuperación de pasajes, la visualización y otros temas son de aplicación directa a la mejora del rendimiento en la recuperación.

El estudio de los sistemas de recuperación de información distribuida es deseable y necesario, ya sea para su aplicación a los sistemas propietarios o a las intranets, en el contexto de lo que se ha dado en llamar *Enterprise Content Management*, ya sea a nivel general. En este sentido, cabe finalizar con una apelación dirigida por Manuel Castells desde la perspectiva global de lo que viene en denominarse la “sociedad red”:

“Una vez que toda la información está en la red, una vez que el conocimiento está en la red, el conocimiento codificado, pero no el conocimiento que se necesita para lo que se quiere hacer, de lo que se trata es de saber dónde está la información, cómo buscarla, cómo procesarla, cómo transformarla en conocimiento específico para lo que se quiere hacer (...) Es ahí donde está, empíricamente hablando, la divisoria [brecha] digital en estos momentos” (Castells, 2001).

Por lo que a este proyecto respecta, cabe la posibilidad de suscribir la frase con que Cyril Cleverdon, el gran clásico de la evaluación, cierra su comunicación a la no menos clásica International Conference on Scientific Information:

> The experiment in which the researcher knows for certain that all the variables have been identified and whether each will significantly affect the results is the exception rather than the rule.
> It is most unlikely that this project is “the exception.”

7. Glosario y siglas

Aunque a lo largo del texto se han presentado las definiciones de algunos conceptos, se incluyen a continución otras, con objeto de completar las anteriores y aclarar algunos términos y el significado de algunas siglas. Para su elaboración se han empleado fuentes tanto electrónicas (Wales, 2004; Rouse, 2004; Webopedia, 2004) como impresas (Corbalán y Amat, 2003).

Accesibilidad (Accesibility)
Probabilidad de que se pueda utilizar el contenido de un documento o el resultante de un servicio Web. La accesibilidad absoluta no discrimina entre grupos de usuarios.

Análisis de transacciones (Log analysis)
Análisis cuantitativo de los ficheros (log files) que registran la actividad de un servidor Web. Su objetivo es determinar qué archivos se solicitan, cuándo, quién los solicita y de dónde proceden las peticiones. Por extensión, el análisis de transacciones de un sistema de recuperación permite identificar y analizar los perfiles empleados y los enlaces resultantes seguidos por los usuarios, así como su frecuencia de aceso y otras características.

ARPAnet
La red de ordenadores que constituyó la base de Internet. De patrocinio militar, consistía en varios ordenadores individuales conectados por líneas dedicadas y que empleaban un método de transmisión común. En los años 80 se desglosó en una red estrictamente militar y la red de la National Science Foundation (NSFnet) que, a su vez, dio paso en 1995 a las redes comerciales actuales.

ASCII
American Standard Code for Information Interchange. Las siglas se aplican a determinados archivos para indicar su carácter textual.

Asignación múltiple
Registro de un mismo dominio de segundo nivel bajo diversos dominios de primer nivel (.net, .org, .com, etc).

ASP (Active Server Pages)
Páginas que incluyen algunos guiones que se procesan en un servidor de Microsoft antes de su envío al usuario. Similares a aplicaciones CGI, permiten adaptar los contenidos al perfil o a las demandas de cada usuario.

BMP
Véase Mapa de bits.

CGI (Common Gateway Interface)
Un procedimiento generalizado mediante la cual se transfiere una petición de un usuario a una aplicación determinada que, tras ejecutar un proceso, devuelve la información al usuario en sintaxis HTML.

Conectividad (Connectiveness)
Véase Accesibilidad.

DCMI
Véase Iniciativa de Metadatos de Dublin.

DESIRE
Development of a European Service for Information on Research and Education. Un proyecto multinacional auspiciado por la Unión Europea que, entre 1998 y 2000, trataba de diseñar grandes sistemas de recuperación de información para las comunidades de investigación.

Dirección IP (IP Address)
Número de 32 bits (en la versión 6 del protocolo de internet) que identifica a cada transmisor o receptor de paquetes de información. Identifica tanto las redes como los dispositivos (servidores o estaciones) dentro de cada una.

Directorios (Directories)
Sistemas de recuperación de información en Internet que organizan sedes (y no páginas o documentos individuales) siguiendo un esquema clasificatorio que se ofrece a los usuarios para su consulta.

Documentos compuestos
Documentos, generalmente electrónicos, que integran unidades de información de varios formatos y cuyo acceso requiere de aplicaciones informáticas igualmente diferentes.

Dominio (Domain)
Conjunto de direcciones de Internet. Los dominios se organizan en niveles jerárquicos de los cuales el más general (de primer nivel) identifica un espacio geográfico (.es por ejemplo) o una categoría (.com). El segundo nivel identifica un lugar determinado dentro de la división más amplia y, de hecho, equivale a una dirección Internet única. Estrictamente hablando, un dominio de segundo nivel (subdominio) puede tener varias denominaciones que serefieren a un servidor o host.

Enlaces de llegada (Links out)
Número de hipernelaces que una página o sede Web emite. Cuando los enlaces se dirigen a documentos situados en otras sedes, son enlaces externos.

Enlaces de partida (Links in)
Número de hiperenlaces que se dirigen a una página o sede Web desde otra página o desde otras sedes.

ES-NIC
Capítulo español del Network Information Center, encargado del registro y asignaciónde las sedes españolas bajo el dominio geográfico .es y, desde 2003, de la asignación de los dominios mixtos combinados con el geográfico.

Espacio Web (WWW, W3, World Wide Web)

Espacio informativo de Internet, gobernado por el protocolo http, en que documentos ajustados a la sintaxis HTML se interrelacionan a través de (hiper)enlaces.

FNC (Federal Networking Council)
Organismo estadounidense, establecido por el National Science and Technology Council's Committee on Computing, Information and Communications, para actuar como foro de coordinación entre diversas agencias federales. El 1 de octubre de 1997, la mayor parte de sus funciones se transfirieron al Large Scale Networking Group.

GIF
Graphics Interchange Format o Formato de Intercambio de Gráficos. Diseñado por UNISYS sobre la base del algoritmo de compresión Lempel Ziv Welch, es el segundo formato más empleado en Internet, después de JPEG. El hecho de que se patentara ha impulsado el desarrollo del formato PNG como sustituto.

Granularidad (Granularity)
Grado de detalle informativo con que se expresa un recurso u objeto. En recuperación, el concepto es similar al de "especificidad de la indización". Las mayores diferencias en granularidad se observan entre los directorios y los sistemas de recopilación automática en función del número de páginas o documentos de una sede que recogen.

HTML (HyperText Markup Language)
Lenguage de marcado de hipertextos, utilizado para expresar, en los documentos transmitidos en la Web, los códigos y marcas utilizados para darles determinado formato.

ICE
Véase Sindicación de contenidos

Information Content Exchange
Véase Sindicación de contenidos

Iniciativa de Metadatos de Dublin (Dublin Core Metadata Initiative)
Esquema para la descrición de recursos electrónicos. Propuesto en 1995 por OCLC con apoyo de instituciones ligadas al desarrollo

de Internet, tenía por objetivo definir un conjunto básico de elementos para la descripción de los recursos de la red. En septiembre de 1998, la Internet Engineering Task Force emitió una RFC (Requets for Comments 2431) sobre el formato. Comprende 15 elementos: título, creador, palabras clave, descripción, editor, otros colaboradores, fecha, tipo, formato, identificador de la fuente, idioma, recursos relacionados, cobertura (geográfica o temporal) y derechos.

Internet invisible (Hidden Web, Invisible Web, Deep Web...)

Conjunto de espacios informativos cuyo acceso requiere interacción de parte del usuario. Habitualmente, se trata de bases de datos cuyos contenidos no pueden ser recopilados por los sistemas de recuperación de información en Internet habituales.

Javascript

JavaScript. Lenguaje de guiones desarrollado por NetScape y empleado en el diseño de sedes Web interactivas. Las páginas generadas pueden llevar la extensión .js

JPG

Ficheros gráficos que utilizan el procedimiento de compresión desarrollado por el Joint Photographic Experts Group.

JS

Véase JavaScript.

LC

Library of Congress, también se aplican las siglas a su esquema clasificatorio que, en ocasiones, se emplea en el Dublin Core para la clasificación de documentos de Internet.

Lenguajes de guiones (Scripting languages)

Lenguaje de programación específico de un programa determiado que lo soporta. Se utiliza habitualmente para automatizar características avanzadas o complejas en el seno de ese programa.

Mapa de bits (Bitmap)

Formato gráfico que define mediante funciones la posición y el color de cada pixel (punto de patalla) en un espacio dado o bien las variaciones entre puntos conjuntos. El término es genérico, aunque usualmente se asigne a ficheros con la extensión .bmp.

Módulo de búsqueda (Searcher)
Parte de un sistema de recuperación de información en Internet encargada de afrontar las solicitudes de búsqueda, procesarlas y presentar los resultados al usuario.

Módulo de indización (Indexer)
Parte de un sistema de recuperación de información en Internet encargada de la construcción de los índices de contenido de las páginas recopiladas.

Módulo de recopilación (Crawler)
Parte de un sistema de recuperación de información en Internet encargada de la identificación de nuevas sedes y páginas y de la transferencia de sus contenidos al módulo de indización.

NFSnet
Véase ARPAnet.

Open Directory Project
Un directorio multilingüe y abierto en que editores voluntarios organizan a través de un esquema clasificatorio ad hoc sedes web según su contenido.

Página Web (Web page)
Documento preparado con sintaxis HTML que se puede visualizar mediante un programa llamado navegador.

Pago por posición (Pay per placement)
Mecanismo por el que un sistema de recuperación recibe una compensación económica por la colocación preferente de documentos en las listas de resultados de búsqueda.

PDF (Portable Document Format)

Formato de Intercambio de Documentos desarrollado por Adobe y convertido, de facto, en un estándar para la preparación y visualización de facsímiles electrónicos.

PHP (Hypertext PreProcessor)
Lenguaje de guiones que permite, entre otras aplicaciones, la generación de páginas dinámicas. Los documentos así generados incorporan la extensión .php.

PNG
Portable Network Graphics véase GIF.

RDF
Véase Resource Description Framework.

Redes de almacenamiento y distribución (Store-and-forward networks)
El almacenamiento y distribución es una técnica común de los servicios de mensajería en que la transmisión entre emisor y receptor pasa por un centro de mensajes, usualmente un servidor que se emplea para almacenar el mensaje entrante hasta que se localiza el dispositivo de destino. Una vez localizado, reenvía el mensaje al destinatario y lo borra del servidor. USEnet, FidoNet y otras son ejemplos de este tipo de redes.

Resource Description Framework
Marco general para la descripción de metadatos de una sede Web. Proporciona interoperatividad entre aplicaciones que intercambian información legible por máquina en Internet.

Sede Web (Web site)
Conjunto de páginas Web interrelacionadas encabezadas por la página principal o home page y alojadas, habitualmente, en un espacio físico común.

Servidor (Host)
Véase Dominio

Sindicación de contenidos
En general, la sindicación es la provisión de material para su reutilización y/o integración con otro, generalmente mediante el pago de una suscripción. Por lo que se refiere a los contenidos electrónicos u online, tras el acuerdo con el proveedor, el distribuidor incorpora a sus páginas o sedes la información contratada. Para asegurar la interoperatividad se ha desarrollado, sobre la base XML, el protocolo Information Content Exchange (ICE) que define una arquitectura y un lenguaje común para la sindicación, la publicación y la comercialización de información

SIRI
Sistemas de Recuperación de Información en Internet.

Sistemas de recopilación automática
Sistemas de recuperación de información en internet dotados de procedimientos para la localización de URL, la extracción de los contenidos de los documentos allí situados y su indización automática.

Subdominio (Subdomain)
Véase Dominio

Tasa de refresco
Frecuencia con que se actualiza la base de datos de un sistema de recopilación automática, tras recorrer las URLs recopiladas con anterioridad.

TCP/IP (Transmission Control Protocol/Internet Protocol)
El protocolo o lenguaje básico de comunicación en Internet, aunque pueda emplearse igualmente en redes internas. De sus dos capas, el TCP se encarga de fraccionar los mensajes en unidades básicas o paquetes que, tras ser transmitidos, vuelven a organizarse en el ordenador que los recibe para resaturar el mensaje original; el Internet Protocol se encarga de la correcta dirección de los paquetes. Como protocolo general, alberga otros como el HTTP, FTP, Tenet y SMTP.

URL (Uniform Resource Locator)

Localizador Uniforme (antes Universal) de Recurso. Dirección única de un documento u objeto en Internet. Contiene el nombre del protocolo de acceso al recurso, un nombre de dominio, que especifica el ordenador que lo alberga, el encaminamiento jerárquico hacia la localización del recurso y el codigo del puerto de acceso al equipo.

USEnet
Véase Redes de almacenamiento y distribución

W3
Espacio Web

World Wide Web
Espacio Web

WWW
Espacio Web

XML
Lenguaje extendido de marcado. Metalenguaje que supone una reducción o versión simplificada del lenguaje generalizado (SGML) para documentos electrónicos transmitidos a través del espacio Web, de manera que se puedan procesar igual que los documentos HTML.

8. Referencias

Diameter of the World-Wide Web (1999). *Nature, 401* (6749): 130-131.

20 Year Usenet Timeline (2003). Google, Inc [Online]. Accesible en: http://www.google.com/googlegroups/archive_announce_20.html (3 de Julio, 2003)

The Open Directory Project (2003). Wikipedia [Online]. Accesible en: http://www.wikipedia.org/wiki/Open_Directory_Project (6 de Agosto, 2003)

Abad García, M. (1997). Evaluación de los componentes de los sistemas de recuperación de la información. En: *Investigación Evaluativa en Documentación* (pp. 125-163). Valencia: Universitat de València.

Abad García, M. (1997). Evaluación de la eficacia de los SRI. En: *Investigación evaluativa en Documentación: Aplicación a la Documentación Médica* (pp. 85-122). Valencia: Universitat de València.

Abiteboul, S., Preda, M., Cobena, G(2003): Adaptive On-Line Page Importance Computation. *Twelfth International World Wide Web Conference*. 20 de mayoo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Adamic, L. Huberman, B. (2001). The Web's Hidden Order. *Communications of the ACM, 44* (9): 55-59.

Adell, J : WWW and gopher statistics? (Respuesta) [Online]. Accesible en: http://groups.google.com/groups?hl=es&lr=&ie=UTF-8&oe=UTF-

8&selm=jordi-150394110424%40bembo.edu.uji.es. (15 de Marzo, 1994)

Adell, J. (2002). Arqueología digital: Los primeros servidores web de España. Universitat Jaume I, Departament de Noves Tecnologies en Educació [Online]. Accesible en: http://nti.uji.es/~jordi/historia_spain_web/html/index.html (13 de Febrero, 2003)

Aguilló, I.(2000): Internet invisible o Infranet: definición, clasificación y evaluación. *Séptimas Jornadas Españolas de Documentación.*19 de octubre de 2000. Bilbao, FESABID.

Tsoi, A.S., Morini, G., Scarselli, F., Hagenbuchner, M., Maggini, M. (2003): Adaptive Ranking of Web Pages. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Aldana Montes, J., Gómez Lora, A., Moreno Vergara, N., Roldán García, MM (2002). Querying the Semantic Web: Feasibility Issues. *UPGrade, 3* (4).

Alonso Berrocal, J. (2000). *Cibermetría. Análisis de los dominios Web españoles: recuperación en internet.* Tesis doctoral. Universidad de Salamanca.

Amat, C. B. (1998). Sistemas de recuperación de información distribuida en Internet. Una revisión de su evolución, sus características y sus perpectivas. Primera parte. *Revista Española de Documentación Científica, 21* (4): 463-474.

Amat, C. B. (1999). Recuperación en Internet: Cuatro modelos complementarios y una agenda para su integración. *Boletín de RedIRIS,*(48).

Amat, C. B. (2003). Caracterización de una muestra de sedes Web españolas bajo dominio .es. *Boletín de RedIRIS,*(64): 33-40.

Andreesen, M : NCSA Nosaic for X 0.10 available [Online]. Accesible en: http://groups.google.com/groups?selm=MARCA.93Mar14225600%40wintermute.ncsa.uiuc.edu. (14 de Marzo, 1993)

Arasu, A., Cho, J., García-Molina, H., Paepcke, A., Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology, 1* (1): 2-43.

AT&T (1995). AT&T to include FrontPage in Easy World Wide Web Service. AT&T [Online]. Accesible en: http://www.att.com/news/1195/951121.bsa.html

Baeza-Yates, R. Ribeiro-Neto, B. (1999). Searching the Web. In *Modern Information Retrieval* (pp. 367-396). Harlow: Pearson Education.

Baeza-Yates, R. (2002). The Web of Spain . *UPGrade* [Online]. Accesible en: http://www.upgrade-cepis.org/issues/2002/3/upgrade-vIII-3.html (21 de Octubre, 2003)

Baeza-Yates, R. Saint-Jean, F. (2003). Análisis de consultas a un buscador y su aplicación a la jerarquización de páginas web. *BiD* [Online]. Accesible en: http://www2.ub.es/bid/consulta_articulos.php?fichero=10baeza.htm (23 de Septiembre, 2003)

Baeza-Yates, R. (2004). Excavando la Web. *El Profesional de la Información, 13* (1): 4-10.

Baeza-Yates, R. (2003). Information retrieval in the Web: beyond current search engines. *International Journal of Approximate Reasoning, 34* (2-3): 97-104.

Bailey, P., Craswell, N., Hawking, D. (2003). Engineering a multi-purpose test collection for Web retrieval experiments. *Information Processing and Management, 39* (6): 853-871.

Bar-Ilan, J. (1998). On the overlap, the precision and estimated recall of search engines. A case study of the query 'Erdos'. *Scientometrics, 42* (2): 207-228.

Bar-Ilan, J. (1999). Search Engine Results over Time: A Case Study on Search Engine Stability. *Cybermetrics, 2-3* (1): 1.

Bar-Ilan, J. (2003). How much information do search engines disclose on the links to a web page? A longitudinal case study of the 'cybermetrics' home page. *Journal of Information Science, 28* (6): 455-466.

Baró i Queralt, J.(1997): Cerca i recuperació d'informació al World Wide Web: una aproximació a les eines disponibles. *Sisenes Jornades Catalanes de Documentació*. 23 de Octubre de 1997. Barcelona: FESABID; SOCADI.

Bates, M. (2002). After the Dot-Bomb: Getting Web Information Retrieval Right This Time. *First Monday* [Online]. Accesible en www.firstmonday.dk/issues/issue7_7/bates/ (20 de septiembre, 2002)

Beaver, A. (1998). Evaluating Search Engine Models for Scholarly Purposes: A report from the Internet Applications Laboratory. *D-Lib Magazine* [Online]. Accesible en: http://www.dlib.org/dlib/diciembre98/12beavers.html (20 de septiembre, 2002)

Beckett, D.(1997): 30% Accessible - A Survey of The UK Wide Web. *6th World Wide Web Conference*. Santa Clara (California), International World Wide Web Consortium.

Behlendorf, B : MCC's EINet(TM) Introduces Galaxy, an Internet Directory Service [Online]. Accesible en: http://groups.google.com/groups?q=einet+galaxy&hl=es&lr=&ie=UTF-8&oe=UTF-8&selm=2i2l2f%24goc%40agate.berkeley.edu&rnum=1. (20 de enero, 1994)

Bellardo Hahn, T. (1998). Text Retrieval Online: Historical Perspective on Web Search Engines. *Bulletin of the American Society for Information Science, 24* (4): 7-10.

Bergman, M. (2001). The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing* [Online]. Accesible en: http://www.press.umich.edu/jep/07-01/bergman.html (11 de julio, 2003)

Bergonneau, M. (2002). The French Connection: Minitel meets the Web. *Onlie Journalism Review* [Online]. Accesible en: http://www.ojr.org/ojr/business/1017968245.php (11 de enero, 2004)

Berners-Lee, T. (1989). Information Management: A Proposal. W3 Archive [Online]. Accesible en: http://www.w3.org/History/1989/proposal.html (11 de julio, 2003)

Berners-Lee, T (1991): WorldWideWeb: Summary [Online]. Accesible en: http://groups.google.com/groups?selm=6487@cernvax.cern.ch. (6 de agosto, 2003)

Berners-Lee, T., Caillou, R., Groff, J., Pollermann, B. (1992). World-Wide Web: The Information Universe . *Electronic Networking: Research, Applications and Policy, 1* (2): 78-84.

Berners-Lee, T. (1996). The World Wide Web: Past, Present and Future. W3 Archive [Online]. Accesible en: http://www.w3.org/People/Berners-Lee/1996/ppf.html (15 de julio, 2003)

Berners-Lee, T. (1998). Semantic Web Road map. World Wide Web Consortium [Online]. Accesible en: http://www.w3.org/DesignIssues/Semantic.html (17 de septiembre, 2003)

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. *Scientific American* (mayo, 2001).

Berrocal, J., Figuerola, C., Zazo, A., Rodríguez, E.(2002): La Cibermetría en la recuperación de información en el Web. *Primeras Jornadas de Tratamiento y Recuperación de la Información*. 4 y 5 de julio de 2002, Valencia.

Berrocal, J., Figuerola, C., Zazo, A., Rodríguez, E. (2003). Agentes inteligentes: recuperación autónoma de la información en la Web. *Revista Española de Documentación Científica, 26* (1): 11-20.

Bharat, K. Broder, A.(1998): A technique for measuring the relative size and overlap of public Web search engines. *7th International WWW Conference*. 14 de abril de 1998. Brisbane.

Bharat, K (2001): Ranking search results by reranking the results based on local inter-connectivity. United States Patent 6,526,440

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation, 56* (1): 71-90.

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research* [Online]. Accesible en: http://informationr.net/ir/8-3/paper152.html (15 de enero, 2004).

Bowman, C., Danzig, P., Hardy, D., Manber, U., Schwartz, M.(1995): The Harvest Information Discovery and Access System.1 de Octubre de 1994. Chicago: National Center for Supercomputing Applications.

Bray, T. (1996). Measuring the Web. *Computer Networks and ISDN Systems, 28* (7-11): 993-1005.

Brewington, B. Cybenko, G. (2000). How dynamic is the Web ? *Computer Networks, 33* (1-6): 257-276.

Brin, S. Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30* (1-7): 107-117.

Broder, A. (2000). Graph structure in the Web. *Computer Networks, 33* (1-6).

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum, 36* (2).

Broder, A., Najork, M., Wiener, J.(2003): Efficient URL Caching for World Wide Web Crawling. *Twelfth International World Wide Web Conference*. 20 de mayoo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Brooks, T. (2003). Web search: how the Web has changed information retrieval. *Information Research* [Online]. Accesible en: http://informationr.net/ir/8-3/paper154.html (15 de enero, 2004).

Bruce, H. (1998). User satisfaction with information seeking on the Internet. *Journal of the American Society for Information Science, 49* (6): 541-556.

Bumgarner, J. (2002). The Great Renaming: 1985 - 1988. James Madison University [Online]. Accesible en: http://www.vrx.net/usenet/history/rename/ (3 de julio, 2003)

Burrows, M (1998): Method for statistically projecting the ranking of information. Unites States Patent 5,765,150

Bush, R. (1993). FidoNet: technology, tools, and history. *Communications of the ACM, 36* (8): 31-35.

Butler, D. (1999). The writing is on the Web for Science journals in print. *Nature, 397* (6716): 195-200.

Caillou, R. (2002). A Little History of the World Wide Web: from 1945 to 1995 Rev 1.39. Web Consortium [Online]. Accesible en: http://www.w3.org/History.html (14 de julio, 2003)

Calanag, M. L. (2003). Public libraries in the information society: what do information policies say*. World Library and Information Congress: 69th IFLA General Conference and Council* . 1 de agosto, 2003. Berlin, IFLA. [Online]. Accesible en http://www.ifla.org/IV/ifla69/papers/112e-Calanag.pdf (5 de febrero, 2004)

Can, F., Nuray, R., Sevdik, A. B. (2004). Automatic performance evaluation of Web search engines. *Information Processing Management, 40* (3): 495-514.

Castells, M. (2001). Internet y la sociedad red: Lección inaugural del programa de doctorado sobre la sociedad de la información y el conocimiento. Universitat Oberta de Catalunya [Online]. Accesible en: http://www.uoc.edu/web/esp/articles/castells/print.html (7 de abril, 2004)

Castillo Blasco, L., Martínez de Pablos, M., Server, G. (1999). Evaluación de la información contenida en seis sedes web de las Escuelas Universitarias y Facultades de Biblioteconomía y Documentación españolas. *Revista Española de Documentación Científica, 22* (3): 325-332.

Castillo Sobrino, M. d., Serrano Moreno, J., Sesmero Llorente, M.(2003): Arquitectura multiagente para la asignación de categorías a textos. *Segundas Jornadas de Tratramiento y Recuperación de la Información*. 8 de Septiembre de 2003. Leganés: Universidad Carlos III.

Cerf, V., Dalal, Y., Sunshine, C. (1974). RFC 675: Specification of Internet transmission control program. Network Information Center Network Working Group [Online]. Accesible en: http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc0675.html (29 de enero, 2003)

Chankhunthod, A., Danzig, P., Neerdaels, C., Schwartz, M., Worrel, K., c (1996). A Hierarchical Internet Object Cache. *Proceedings of the 1996 Usenix Technical Conference* [Online]. Accesible en:

http://www.usenix.org/publications/library/proceedings/sd96/full_papers/danzig-html/cache.html (22 de enero, 1996)

Cho, J. García-Molina, H.(2000): The Evolution of the Web and Implications for an Incremental Crawler. *VLDB Conference*.1 de Septiembre de 2000. El Cairo, Very Large Data Base Endowment Inc. [Online]. Accesible en: http://www.vldb.org/dblp/db/conf/vldb/ChoG00.html (15 de agosto, 2003)

Claffy, K. (2000). Measuring the Internet. *IEEE Internet Computing, 4* (1): 73-75.

Clarke, S. Willett, P. (1997). Estimating the recall perfomance of Web search engines. *ASLIB Proceedings, 49* (7): 184-189.

Clever Project (1999). Hypersearching the Web. *Scientific American,*(junio, 1999).

Codina, L. (2003). La Web semántica: una visión crítica. *El Profesional de la Información, 12* (2): 149-152.

Comisión del Mercado de las Telecomunicaciones (2001). Estudio sobre la presencia de las entidades españolas (.es) en Internet. *Novatica,*(152): 42-44.

Computer Museum History Center (2002). Timeline of Computer History. Computer Museum History Center [Online]. Accesible en: http://www.computerhistory.org/timeline/ (5 de febrero, 2003)

Corbalán, L. M. Amat, C. B. (2003). *Vocabulario de información y documentación automatizada*. Valencia: Universitat de València.

Corchuelo, R., Arjona, J., Toro, M. (2002). Automatic Extraction of Semantically-Meaningful Information from the Web. *UPGrade, 3* (3).

Corporation for Research and Educational Networking (1997). CREN History and Future. Corporation for Research and Educational Networking [Online]. Accesible en: http://www.cren.net/cren/cren-hist-fut.html (7 de febrero, 2003)

Courtois, M. Berry, M. (1999). Results ranking in Web search engines. *Online Magazine, 23* (3): 39.

Craven, T. C. (2004). Variations in use of meta tag descriptions by Web pages in different languages. *Information Processing Management, 40* (3): 479-493.

Crimmins, F., Smeaton, A., Dkaki, T., Mothe, J. (1999). TétraFusion: Information Discovery on the Internet. *IEEE Intelligent Systems, 14* (4): 55-62.

Croft, W. Turtle, H.(1989): A Retrieval Model Incorporating Hypertext Links. *Proceedings of the second annual ACM conference on Hypertext*. 1 de Noviembre de 1989. Pittsburgh, ACM.

Culliss, G (1999): Method for organizing information. United States Patent 6,006,222

Danzig, P., Obraczka, K., Li, S. (1993). Internet Resource Discovery Services. *IEEE Computer, 26* (9): 8-22.

Dasen, M. Wilde, E.(2001): Keeping Web indices up-to-date. *Tenth International World Wide Web Conference.*1 de Mayo de 2001. Hong Kong, International World Wide Web Consortium.

Davila, R. (2000). History and Development of the Internet. San Antonio Public Library: Government Documents [Online]. Accesible en: http://www.sat.lib.tx.us/Displays/itintro.htm (31 de Enero, 2003)

Dekkers, M. Weibel, S. (2003). State of the Dublin Core Metadata Initiative, Abril 2003. *D-Lib Magazine* [Online]. Accesible en: http://www.dlib.org/dlib/april03/weibel/04weibel.html (15 de enero, 2004).

Deutsch, P. Emtage, A.(1992): Archie: An Electronic Directory Service for the Internet. *Proceedings of Usenix*.1 de Enero de 1992. San Francisco, USENIX.

Dhyani, D, Keong Ng, W, Bhowmick, SS (2002). A survey of Web Metrics. *ACM Computing Surveys, 34* (4): 469-503.

Digital Equipment Corporation : Digital develops Internet's first "Super Spider" [Online]. Accesible en: http://groups.google.com/groups?selm=9512151806.AA02246%40raptor.pa.dec.com. (15 de diciembre, 2003)

Dill, S., Kumar, R., Mccurley, K., Rajagopalan, S., Sivakumar, D., Tomkins, A. (2002). Self-similarity in the web. *ACM Transactions on Internet Technology, 2* (3): 205-223.

Douglis, F., Feldmann, A., Krishnamurthy, B., Mogul, J. (1997). Rate of Change and other Metrics: a Live Study of the World Wide Web.

USENIX Symposium on Internet Technologies and System. 8 de diciembre, 1997. Monterrey, USENIX.

Dublin Core Metadata Initiative (2003). Dublin Core Metadata Element Set, Version 1.1: Reference Description. OCLC DCMI [Online]. Accesible en: http://www.dublincore.org/documents/dces/ (16 de septiembre, 2003)

Eckmann, J. Moses, E. (2002). Curvature of co-links uncovers hidden thematic layers in the World Wide Web. *Proceedings of the National Academy of Sciences USA, 99* (9): 5825-5829.

Eiron, N. Mccurley, K.(2003): Analysis of anchor text for web search. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. 28 de Julio de 2003. Toronto, ACM.

Emtage, A : Announcing "Archie 1.0": The Archive Server Server [Online]. Accesible en: http://groups.google.com/groups?q=archie+emtage&hl=es&lr=&ie=UTF-8&oe=UTF-8&selm=1990Nov15.045448.2861%40ox.com&rnum=1. (14 de noviembre, 1990)

Enos, L. (2001). Excite@Home is raising funds to improve its bottom line while at the same time taking steps to cut costs*. E-Commerce Times* [Online]. Accesible en: http://www.ecommercetimes.com/perl/story/11148.html (20 de agosto, 2002)

Escalona, M., Mejías, M., Torres, J. (2002). Methodologies to develop Web Information Systems and Comparative Analysis. *UPGrade, 3* (3).

ESNIC (2003). Estadísticas del ES-NIC: Dominios registrados en los últimos años. ESNIC [Online]. Accesible en: https://www.nic.es/documentacion/estadisticas.html (30 de julio, 2003)

Faloutsos, M., Faloutsos, P., Faloutsos, C. (1999). On Power-Law Relationships of the Internet Topology . *ACM SIGCOMM Computer Communication Review , Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication, 29* (4): 251-262.

Federal Networking Council (1995). FNC Resolution: Definition of "Internet" Federal Networking Council. [Online]. Accesible en http://www.hpcc.gov/fnc/Internet_res.html (11 de septiembre, 2002).

Fernández Beobide, C. González Obiol, A. (1992). Videotex e Ibertex: Experiencias y realizaciones. *Telos,*(29).

Fetterly, D., Manasse, M., Najork, M., Wiener, J.(2003): A Large-Scale Study of the Evolution of Web Pages. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Fichter, D. (2003). Exploiting intranet search engines for data discovery. *Online, 27* (6): 47.

Fidel, R., Davies, R., Douglas, M., Holder, J., Hopkins, C., Kushner, E. et al. (1999). A visit to the information mall: Web searching behavior of high school students. *Journal of the American Society for Information Science, 50* (1): 24-37.

Ford, G. (2001). Theory and Practice in the Networked Environment: A European Perspective. In C.McClure J. Bertot (Eds.), *Evaluating Networked Information Services. Techniques, Policy and Issues* (pp. 1-22). Melford: Information Today.

Ford, N. Miller, D. M. N. (2001). The role of individual differences in Internet searching: An empirical study. *Journal of the American Society for Information Science and Technology, 52* (12): 1049-1066.

Foster, S : Veronica: an Archie for Gopher [Online]. Accesible en: http://groups.google.com/groups?q=veronica+nevada+university+group:comp.infosystems.gopher&start=20&hl=es&lr=&ie=UTF-8&oe=UTF-8&scoring=d&selm=9211180514.AA01778%40pyramid&rnum=25. (17 de noviembre, 2003)

Fox, E. Urs, S. (2002). Digital Libraries. *Annual Review of Information Science and Technology, 36:* 503-589.

Fragoudis, D. Likothanassis, S.(1999): Retriever: an agent for intelligent information recovery. *Proceedings of the 20th*

International Conference on Information Systems.12 de Diciembre de 1999. Charlotte (NC).

García Barriocanal, H., Sicilia Urbán, M., Aedo Cuevas, I. (2003). Ontology-Based Annotation of Usability Evaluation-Related Resources: Design and Retrieval Mechanisms . *UPGrade, 4* (1): 12-17.

García Santiago, M. (2000). *Topología de la información en la World Wide Web: Modelo metodológico de visualización en una red hipertextual nacional.* Tesis doctoral. Universidad de Granada.

García, J. (1998). IRIS-NEWS: la aventura de la Usenet en RedIRIS. *Boletín de RedIRIS,*(44).

Garratt, A., Jakson, M., Burden, P., Wallis, J. (2001). A survey of alternative designs for a search engine storage structure. *Information and Storage Technology, 43* (11): 661-677.

Glover, E., Tsioutsiouliklis, K., Lawrence, S., Pennock, D., Flake, G.(2002): Using Web Structure for Classifying and Describing Web Pages. *Eleventh International World Wide Web Conference*. 7 de Mayo de 2002. Honolulu. International World Wide Web Consortium.

Google Groups Team (2001). Google Groups Archive Information. google.public.support.general [Online]. Accesible en: http://groups.google.com/groups?selm=90cbefb1.0112211728.4cfe9bb%40posting.google.com (8 de julio, 2003)

Gorbunov, A. (2002). Relevance of Web documents: Ghosts consensus method. *Journal of the American Society for Information Science and Technology, 53* (10): 783-788.

Gordon, M. Pathak, P. (1999). Finding information on the world wide web: the retrieval efectiveness of search engines. *Information Processing and Management, 35* (2): 144-180.

Gómez Díaz, R. (2003). La evaluación en recuperación de la información. *Hipertext.net* [Online]. Accesible en: http://www.hipertext.net/web/pag188.htm (5 de noviembre, 2003)

Gravano, L., Chang, K., García Molina, H., Lagoze, C., Paepcke, A. (1997). STARTS: Stanford Protocol Proposal for Internet Retrieval and Search. Digital Library Project Stanford University [Online].

Accesible en: http://www-db.stanford.edu/~gravano/starts.html (15 de septiembre, 2003)

Greco, G., Greco, S., Zumpano, E. (2001). A Probabilistic Approach for Distillation and Ranking of Web Pages. *World Wide Web, 4:* 189-207.

Griffiths, R. (2002). History of Internet, Internet for Historians (and just about everyone else). Leiden University [Online]. Accesible en: http://www.let.leidenuniv.nl/history/ivh/frame_theorie.html (2 de julio, 2003)

Guha, R., McCool, R., Miller, E.(2003): Semantic Search. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Gurrin, C. Smeaton, A. (2004). Replicating Web Structure in Small-Scale Test Collections. *Information Retrieval, 7* (3-4): 239-263.

GVU's WWW Surveying Team (1998). GVU's Tenth WWW User Survey (Conducted Octubre 1998). Georgia Institute of Technology [Online]. Accesible en: http://www.gvu.gatech.edu/user_surveys/survey-1998-10/ (30 de septiembre, 2003)

Haas, S. Grams, E. (2000). Readers, Authors, and Page Structure: A Discussion of Four Questions Arising from a Content Analysis of Web Pages. *Journal of the American Society for Information Science, 51* (2): 181-192.

Hald, A. (1952). *Statistical Tables and Formulas*. (s.l.): Wiley.

Han, Y., Loke, S., Sterling, L. (1996). *Agents for Citation Finding on the World Wide Web* . Technical Report 96/40. Parkville, University of Melbourne.

Hardy, D., Schwartz, M., Wessels, D. (1996). Harvest User's Manual Version 1.4 patchlevel 2. Internet Research Task Force Research Group on Resource Discovery [Online]. Accesible en: http://harvest.sourceforge.net/harvest-1.4.pl2-docs/user-manual.html (15 de septiembre, 2003)

Hardy, H. (1993). *The History of the Net v8.5.* Master Thesis. School of Communications, Grand Valley State University.

Harter, S. Hert, C. (1997). Evaluation of Information Retrieval Systems: Approaches, Issues and Methods. *Annual Review of Information Science and Technology, 32:* 3-94.

Hauben, M. Hauben, R. (1996). Netizens: On the History and Impact of the Net. Columbia University [Online]. Accesible en: http://www.columbia.edu/~rh120/ (2 de julio, 2003)

Hausherr, T. (2001). Xenu's Link Sleuth (Version 1.1c) [Programa informático]. Berlin.

Hawking, D., Craswell, N., Thistlewaite, P., Harman, D. (1999). Results and challenges in Web search evaluation. *Computer Networks, 31* 11-16.

Hawking, D., Craswell, N., Bailey, P., Griffiths, K. (2001). Measuring Search Engine Quality. *Information Retrieval, 4* (1): 33-59.

Hawking, D. Robertson, S. (2003). On Collection Size and Retrieval Effectiveness. *Information Retrieval, 6* (1): 99-105.

Heery, R. (1996). Review of Metadata Formats. *Program, 30* (4): 345-373.

Hendler, J. (1999). Web Matters: Is there an Intelligent Agent in Your Future ? Nature [Online]. Accesible en: http://www.nature.com/nature/webmatters/agents/agents.html (10 de diciembre, 2003)

Hendler, J. (2001). Agents and the Semantic Web. *IEEE Intelligent Systems, 16* (2): 30-37.

Henzinger, M., Heydon, A., MIzenmacher, M., Najork, M. (1999). Measuring index quality using random walks on the Web. *Computer Networks, 31* 1291-1303.

Henzinger, M., Bay-Wei Chang, Brian Milch, Sergey Brin(2003): Query-Free News Search. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Hermans, B. (1996). Intelligent Software Agents on the Internet: an inventory of currently offered functionality in the information society and a prediction of (near-)future developments. Doctoral Dissertation Tilburg University [Online]. Accesible en:

http://www.broadcatch.com/agent_thesis/ (24 de septiembre, 2003)

Hermans, B. (1997). Intelligent Software Agents on the Internet. *First Monday* [Online]. Accesible en: http://www.firstmonday.dk/issues/issue2_3/index.html (24 de septiembre, 2003)

Hermans, B. (1998). Desperately Seeking: Helping Hands and Human Touch. *First Monday* [Online]. Accesible en: http://www.firstmonday.dk/issues/issue3_11/index.html (24 de septiembre, 2003)

Herring, S. (2002). Computer-Mediated Communication on the Internet. *Annual Review of Information Science and Technology, 36:* 109-168.

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

Hölscher, C. Strube, G.(2000): Web Search Behavior of Internet Experts and Newbies. *Ninth International World Wide Web Conference*. 15 de Mayo de 2000. Amsterdam: Centre for Mathematics and Computer Science; International World Wide Web Consortium.

Hsieh-Yee, I. (1998). The retrieval power of selected search engines: how well do they address general reference questions and subject questions? *Reference Librarian, 60* 27-47.

Huberman, B., Pirolli, P., Pitkow, J., Lukose, R. (1998). Strong Regularities in World Wide Web Surfing. *Science, 280* (5630): 95-97.

Huberman, B. Adamic, L. (1999). Growth dynamics of the World-Wide Web. *Nature, 401* (6749): 131.

Huberman, B. (2002). Patterns in the World Wide Web. *Libray of Economics and Liberty* [Online]. Accesible en: http://www.econlib.org/library/Columns/Hubermanpatterns.html (5 de febrero, 2004)

Internet Society (2002). What is the Internet ? Internet Society [Online]. Accesible en: http://www.isoc.org/internet/index.shtml (3 de marzo, 2003)

Jansen, B. (1997). Using an intelligent agent to enhace search engine perfomance. *First Monday* [Online]. Accesible en: http://www.firstmonday.dk/issues/issue2_3/jansen/index.html (24 de septiembre, 2003)

Jansen, B. Pooch, U. (2001). A Review of Web Searching Studies and a Framework for Future Research. *Journal of the American Society for Information Science, 52* (3): 235-246.

Jansen, B., Spink, A., Saracevic, T. (2002). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management, 36* (2): 207-227.

Jansen, B. Spink, A. An analysis of Web searching by European AlltheWeb.com users. *Information Processing and Management,* (en prensa).

Delort, J.Y., Bouchon-Meunier, B., Rifqi, M. (2003): Web Document Summarization by Context. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Jenkins, C., Jackson, M., Burden, P., Wallis, J. (1998). Searching the World Wide Web: an evaluation of available tools and methodologies. *Information and Storage Technology, 39* (14-15): 985-994.

Johnson, F., Griffiths, J., Hartley, R. (2001). *DEVISE. A framework for the evaluation of Internet search engines* (Rep. No. 100). London: British Library.

Johnstone, B. Carlson, D. (2002). History of Electronic Publishing: Teletext and Videotext. Applied Interactive Newspapers Syllabus, Univ of Florida [Online]. Accesible en: http://iml.jou.ufl.edu/carlson/professional/new_media/history/ehistory.htm (17 de julio, 2003).

Kahle, B. (1989). Wide Area Information Server Concepts v4 Draft. Thinking Machines Corporation [Online]. Accesible en: http://nti.uji.es/software/Simple/docs/wais-concepts.txt (20 de julio, 2003).

Kahle, B. Medlar, A. (1991). An Information System for Corporate Users: Wide Area Information Servers v3. Universidad de Heidelberg [Online]. Accesible en: http://www.urz.uni-heidelberg.de/Netzdienste/internet/tools/info/wais/corporate.html (23 de julio, 2003).

Kannan, N : Qualifiers on Hypertext links... [Online]. Accesible en: http://groups.google.com/groups?selm=1991Aug2.115241@ardor.enet.dec.com. (2 de agosto, 2003).

Kantor, B. Lapsley, P. (1986). RFC 977: Network News Transfer Protocol: A Proposed Standard for the Stream-Based Transmission of News. Network Working Group [Online]. Accesible en: ftp://ftp.isi.edu/in-notes/rfc977.txt (8 de julio, 2003).

Kessler, J. (1995). The French Minitel: Is There Digital Life Outside of the "US ASCII" Internet? A Challenge or Convergence? *D-Lib Magazine* [Online]. Accesible en: http://www.dlib.org/dlib/diciembre95/12kessler.html (5 de septiembre, 2002).

Khan, M. Khor, S. (2004). Enhanced Web document retrieval using automatic query expansion. *Journal of the American Society for Information Science and Technology, 55* (1): 29-40.

Khare, R. Rifkin, A.(1998): The origin of (document) species. *7th International World Wide Web Conference*. 14 de abril de 1998. Brisbane.

Kirsch, ST (1997): Document retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents. United States Patent 5,659,732.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46* (5): 604-632.

Kleinberg, J (2000): Method and system for identifying authoritative information resources in an environment with content-based links between information resources. United States Patent 6,112,202.

Kleinberg, J. Lawrence, S. (2001). The Structure of the Web. *Science, 294* 1849-1850.

Kobayashi, M. Takeda, K. (2000). Information Retrieval on the Web. *ACM Computing Surveys, 32* (2): 144-173.

Koch, T., Ardo, A., Brümer, A., Lundberg, S. (1996). The building and maintenance of robot based internet search services: A review of current indexing and data collection methods. NetLab Lund University Library [Online]. Accesible en: http://www.lub.lu.se/desire/radar/reports/D3.11/ (1 de septiembre, 2001).

Koehler, W. (1999). An Analysis of Web Page and Web Site Constancy and Perfomance. *Journal of the American Society for Information Science, 50* (2): 162-180.

Koehler, W. (2002). Web page change and persistence: A four-year longitudinal study. *Journal of the American Society for Information Science and Technology, 53* (2): 162-171.

Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence. *Information Research* [Online]. Accesible en: http://informationr.net/ir/9-2/paper174.html (5 de febrero, 2004)

Koster, M : ALIWEB (Archie-Like Indexing for the Web) [Online]. Accesible en: http://groups.google.com/groups?q=koster+aliweb+group:comp.infosystems.www+author:koster&hl=es&lr=&ie=UTF-8&oe=UTF-8&selm=1993Nov30.093536.28554%40cs.nott.ac.uk&rnum=1. (30 de noviembre, 2003).

Koster, M.(1994): ALIWEB - Archie-Like Indexing in the WEB. *First International Conference on the World-Wide Web*. 25 de mayoo, 1994. Geneva: CERN.

Kwong, L. Ng, Y. (2003). Performing Binary-Categorization on Multiple-Record Web Documents Using Information Retrieval Models and Application Ontologies. *World Wide Web, 6* (3): 281-303.

Lamas, C. (2002). La investigación de Internet. *Telos,*(52).

Lancaster, F. Warner, A. (1993). Evaluation Criteria and Evaluation Procedures. In *Information Retrieval Today* (pp. 159-202). Arlington: Information Resources Press.

Lancaster, F. Warner, A. (1993). Subject Access: Problems and Perfomance Criteria. In *Information Retrieval Today* (pp. 43-63). Arlington: Information Resources Press.

Landoni, M. Bell, S. (2000). Information retrieval techniques for evaluating search engines: a critical overview. *ASLIB Proceedings, 52* (3): 124-129.

Lavoie, B. Frystyk Nielsen, H. (2003). Web Characterization Terminology Definitions Sheet. World Wide Web Consortium [Online]. Accesible en: http://www.w3.org/1999/05/WCA-terms/ (3 de mayo, 2002)

Lawrence, S. Giles, C. (1998). Searching the World Wide Web. *Science, 280* (5630): 98-100.

Lawrence, S. Giles, C. (1999). Accesibility of Information on the Web. *Nature, 400* (6740): 107-109.

Lawrence, S. Giles, C. (1999). Searching the Web: General and Scientific Information Access. *IEEE Communications, 37* (1): 116-122.

Leighton, V. Srivastava, J. (1997). Precision among World Wide Web Search Services. Winona State University [Online]. Accesible en: http://www.winona.msus.edu/library/webind2/webind2.htm (24 de septiembre, 2003)

Leighton, V. Srivastava, J. (1999). First 20 Precision among World Wide Web search services (Search Engines). *Journal of the American Society for Information Science, 50* (10): 870-881.

Leiner, B., Cerf, V., Clark, D., Kahn, R., Kleinrock, L., Lynch, D. et al. (1997). The past and future history of the Internet. *Communications of the ACM, 40* (2): 102-108.

Leiner, B., Cerf, V., Clark, D., Kahn, D., Kleinrock, L., Lynch, D. et al. (2000). *A Brief History of the Internet* Internet Society.

Li, L. Shang, Y. (2000). A new method for automatic performance comparison of search engines. *World Wide Web, 3* 241-247.

Li, L., Shang, Y., Zhang, W.(2002): Improvement of HITS-based Algorithms on Web Documents. *International World Wide Web Conference*. 7 de Mayo de 2002. Honolulu, International World Wide Web Consortium.

Liaw, S. S. Huang, H. M. (2003). An investigation of user attitudes toward search engines as an information retrieval tool. *Computers in Human Behavior, 19* (6): 751-765.

Licklider, J. Clark, W. (1962). On-line Man Computer Interactions. MIT. Publicado también como Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*: 4-11, 1960 [Online]. Accesible en: http://medg.lcs.mit.edu/people/psz/Licklider.html (28 de abril, 2004).

Lieberman, H., Fry, C., Weitzman, L. (2003). Exploring the Web with Reconnaissance Agents. *Communications of the ACM, 44* (8): 69-75.

Lindner, P (1991) : Internet Gopher v0.2 Curses Client and Server is available. [Online]. Accesible en: http://groups.google.com/groups?selm=1991Sep10.020238.4751%40cs.umn.edu. (10 de septiembre, 2002)

Lim, L., Wang, M., Padmanabha, S., Vitte, J.S., Agarwa, R (2003): Dynamic Maintenance of Web Indexes Using Landmarks. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Loeber, S. Cristea, A. (2003). A WWW Information Seeking Process Model. *Educational Technology Society, 6* (3): 43-52.

López Alonso, M. Mares Marín, J.(1996): El futuro de la identificación de la información en Internet. *Quintas Jornadas Españolas de Documentación Automatizada*. 17 de Octubre de 1996. Caceres: FESABID.

López, D. Massa, J. (1998). Dando forma al envase y, con ello, al contenido: Webber. *Boletín de RedIRIS,*(45).

Lyman, P. Varian, H. (2000). How Much Information? *Journal of Electronic Publishing* [Online]. Accesible en: http://www.press.umich.edu/jep/06-02/lyman.html (3 de Marzo, 2002).

MacMurdo, G. (1995). How the Internet was indexed. *Journal of Information Science, 21* (6): 479-489.

Maes, P. (2003). Agents that reduce work and information overload. *Communications of the ACM, 37* (7): 30-40.

Maldonado Martínez, A. Fernández Sánchez, E.(1998): Evaluación de los principales "buscadores" desde un punto de vista documental: recogida, análisis y recuperación de recursos de información. *Sextas Jornadas Españolas de Documentación*. 29 de Octubre de 1998. Valencia: FESABID.

Mañas, J. (1994). Búsqueda y recuperación de información en Internet. *Novatica,*(110): 75-81.

Marable, L. (2003). *False Oracles: Consumer Reaction to Learning the Truth About How Search Engines Work: Results of an Ethnographic Study.* Baltimore, Consumer WebWacht Research [Online]. Accesible en : http://www.consumerwebwatch.org/news/searchengines/ (28 de abril, 2004)

Marzoiori, M.(1998): The limits of Web metadata, and beyond. *Seventh International World Wide Web Conference*. 14 de Abrilde 1998. Brisbane. International World Wide Web Consortium.

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

Martínez de Lejarza Esparducer, I. (1999). Una aproximación al análisis regional del mercado de la información digital: Distribución, concentración y difusión regional de Internet en las comunidades autónomas españolas. Departamento de Economía Aplicada.Universidad de Valencia [Online]. Accesible en: http://www.uv.es/~econinfo/presentacion/mercinter/mercinter.html (4 de agosto, 2003)

Martínez Méndez, F. (2001). *Propuesta y desarrollo de una metodología para la evaluación de la recuperación de información en Internet.* Tesis doctoral. Universidad de Murcia.

Martínez Méndez, F. Rodríguez Múñoz, J. (2003). Síntesis y crítica de las evaluaciones de la efectividad de los motores de búsqueda en la Web. *Information Research* [Online]. Accesible en: http://informationr.net/ir/8-2/paper148.html (15 de enero, 2004).

Masashi Toyoda Masaru Kitsuregawa(2003): Analyzing Global Behavior of Web Community Evolution. *Twelfth International World Wide Web Conference*. Budapest: Computer and

Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Massa, J. (2003). Metainformación Dublin Core: Elementos del conjunto de metadatas de Dublin Core: Descripción de Referencia. CSIC Red IRIS [Online]. Accesible en: http://www.rediris.es/metadata/dublin_core_elements.es.html (16 de septiembre, 2003)

Mauldin, M. Leavitt, J.(1994): Web-agent related research at the Center for Machine Traslation. *Proceedings of the ACM Special Interest Group on Networked Information Discovery and Retrieval*. 1 de Agosto de 1994. McLean, ACM.

Mauldin, M. (1995). Measuring the Web with Lycos. Third International WWW Conference [Online]. Accesible en: http://www.lazytoad.com/lti/pub/lycos-websize.html (31 de enero, 2001)

Mauldin, M. (1997). Lycos: Design choices in an Internet search service. *IEEE Expert,*(Janjuary February): 8-11.

Mauldin, ML (1998): Method for searching a queued and ranked constructed catalog of files stored on a network. United States Patent 5,748,954

McCahill, M. Anklesaria, F. (1995). Evolution of Internet Gopher. *Journal of Universal Computer Science, 1* (4): 235-246.

Meghabghab, G. (2001). Google's Web Page Ranking Applied to Different Topological Web Graph Structures. *Journal of the American Society for Information Science and Technology, 52* (9): 736-747.

Menczer, F. (2003). Complementing search engines with online web mining agents. *Decision Support Systems, 35* (2): 195-212.

Menczer.F (2002). Growing and navigating the small world Web by local content. *Proceedings of the National Academy of Sciences USA, 99* (22): 14014-14019.

Méndez Rodríguez, E. (2002). *Metadatos y recuperación de información . Estándares , problemas y aplicabilidad en bibliotecas digitales*. Gijón, Trea.

Microsoft (1996). Microsoft Acquires Vermeer Technologies Inc.: Critically Acclaimed Visual Client-Server Web Publishing Tool to

Complement Internet Offerings From Microsoft Desktop Applications Division. [Online]. Accesible en: http://www.microsoft.com/presspass/press/1996/jan96/vrmeerpr.asp (21 de julio, 2003)

Milne, J (1995). Vermeer Technologies Gives Birth to FrontPage. *Network Computing, 6*.

Ministerio de Ciencia y Tecnología (2003). ORDEN CTE/662/2003, de 18 de marzo, por la que se aprueba el Plan Nacional de nombres de dominio de Internet bajo el código de país correspondiente a España («.es»). *Boletín Oficial del Estado,*(73): 11917-11924.

Mladenic, D. (1999). Text-Learning and related Intelligent Agents: A Survey. *IEEE Intelligent Systems, 14* (4): 44-54.

Moise, G., Sander, J., Rafiei, D.(2003): Focused Co-citation: Improving the Retrieval of Related Pages on the Web. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Monk, T. Claffy, K. (2002). A survey of Internet Statistics/Metrics Activities. Technical Report. National Laboratory for Applied Network Research [Online]. Accesible en: http://www.caida.org/outreach/papers/1996/metricsurvey/metricsurvey.html (11 de abril, 2002)

Montaner, M., López, B., Rosa, J. d. l. (2003). A Taxonomy of Recommender Agents on the Internet. *Artificial Intelligence Review, 19* (4): 285-330.

Montebello, M.(1998): Optimizing recall/precision scores in IR over the WWW. *Proceedings of the 21st Annual International ACN SIGIR Conference on Research and Development in Information Retrieval*. 1 de Agosto de 1998. Melbourne: ACM.

Mowshowitz, A. Kawaguchi, A. (2002). Assessing bias in search engines. *Information Processing and Management, 38* (1): 141-156.

Muylle, S., Moenaert, R., Despontin, M. (2004). The conceptualization and empirical validation of web site user satisfaction. *Information Management, 41* (5): 543-560.

Najork, M. Wiener, J.(2001): Breadth-First Search Crawling Yields High-Quality Pages. *Tenth International World Wide Web Conference.*1 de Mayo de 2001. Hong Kong, International World Wide Web Consortium.

Netscape Communications Corporation (1997). NetScape works with W3C and leading content providers to drive new specification for organizing, describing and navigating information on internet, intranets and desktops. NetScape [Online]. Accesible en: http://wp.netscape.com/flash1/newsref/pr/newsrelease488.html (16 de septiembre, 2003)

Newby GB (2002). The necessity for information space mapping for information retrieval on the semantic web. *Information Research* [Online]. Accesible en: http://informationr.net/ir/7-4/paper137.html (15 de enero, 2003).

Nogales Flores, J. (1999). Los usos básicos de Internet. Servicios y aplicaciones. En Caridad Sebastián, M (Ed.): *La Sociedad de la Información. Política, Tecnología e Industria de los contenidos* (pp. 143-173). Madrid: Centro de Estudios Ramón Areces.

Noh, Y.-H. (2003). A study on the estimation of performance of the concept-based information retrieval model for searching the Web. *Journal of Information Science, 28* (5): 407-415.

O'Neill, E., McClain, P., Lavoie, B. (1998). A Methodology for Samplig the World Wide Web. *Annual Review of OCLC Research* [Online]. Accesible en: http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003447 (20 de octubre, 2003)

O'Neill, E., Lavoie, B., McClain, P. (1999). Web Characterization Project: An Analysis of Metadata Usage on the Web. *Annual Review of OCLC Research* [Online]. Accesible en: http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003486 (21 de octubre, 2003)

O'Neill, E., Lavoie, B., Bennett, R. (2003). Trends in the Evolution of the Public Web: 1998-2002. *D-Lib Magazine* [Online]. Accesible

en: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html (21 de octubre, 2003).

Olvera Lobo, M. (1999). Métodos y técnicas para la indización y la recuperación de los recursos de la World Wide Web. *Boletín de la Asociación Andaluza de Bibliotecarios, 14* (57): 11-22.

Olvera Lobo, M. (1999). *Evaluación de la recuperación de información en internet : un modelo experimental.* Tesis doctoral. Universidad de Granada.

Olvera Lobo, M. (2000). Rendimiento de los sistemas de recuperación de información en la world wide web: revisión metodológica. *Revista Española de Documentación Científica, 23* (1): 63-78.

Olvera Lobo, M. (2000). Rendimiento de los sistemas de recuperación de información en la web: evaluación de servicios de búsqueda (search engines). *Revista Española de Documentación Científica, 23* (3): 302-316.

Oppenheim, C., Morris, A., Mcknight, C., Lowley, S. (2000). The evaluation of WWW search engines. *Journal of Documentation, 52* (2): 190-211.

Page, L (2001): Method for node ranking in a linked database. United States Patent 6,285,999

Peiró, C. (1996). El fenómeno Internet en España: ayer, hoy y mañana. Exposición Universal Internet'96 [Online]. Accesible en: http://personales.mundivia.es/astruc/doctxt53.htm (17 de julio, 2003)

Peterson, R. (1997). Eight Internet Search Engines Compared. *First Monday* [Online]. Accesible en: http://www.firstmonday.dk/issues/issue2_2/peterson/index.html (1 de Marzo, 2000).

Pettigrew, K., Durrance, J., Unruh, K. (2002). Facilitating community information seeking using the Internet: Findings from three public library-community network systems. *Journal of the American Society for Information Science and Technology, 53* (11): 894-903.

Picard, J. Savoy, J. (2003). Enhancing retrieval with hyperlinks: A general model based on propositional argumentation systems.

Journal of the American Society for Information Science and Technology, 54 (4): 347-355.

Pinkerton, B (1994). The WebCrawler Index: A content-based Web index [Online]. Accesible en: http://groups.google.com/groups?selm=2r0rnm%24ftj%40news.u.washington.edu. (11 de June, 1994)

Pinkerton, B.(1994): Finding What People Want: Experiences with the WebCrawler.1 de Octubre de 1994. Chicago: National Center for Supercomputing Applications.

Pinkerton, B. (2000). *WebCrawler: Finding What People Want.* Doctor in Philosophy Doctoral Dissertation, Computer Science Department, University of Washington.

Pinkerton, B. (2002). WebCrawler Timeline. WebCrawler [Online]. Accesible en: http://www.thinkpink.com/bp/WebCrawler/History.html (4 de septiembre, 2003)

Pollock, A. Hockley, A. (1997). What's Wrong with Internet Searching. *D-Lib Magazine* [Online]. Accesible en: http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/marzo97/bt/03pollock.html (15 de abril, 2001).

PricewaterhouseCoopers (2004). Estudio de la industria de contenidos digitales en España. Price Waterhouse Coopers España [Online]. Accesible en: http://www.pwc.com/es/esp/ins-sol/spec-int/ind_contenidos.html (7 de abril, 2004)

Quaterman, J. Hoskins, J. (1986). Notable Computer Networks. *Communications of the ACM, 29* (10): 932-971.

Quaterman, J. (1996). User Growth of the Internet and of the Matrix. *Matrix News, 6* (5).

Raghavan, P. (2002). Information Retrieval for Enterprise Content . *UPGrade, 3* (3).

Rasmussen, E. (2003). Indexing and retrieval for the Web. *Annual Review of Information Science and Technology, 37:* 91-124.

Rhind-Tutt, S. (2003). Semantic indexing: a case study. *Library Collections, Adquisitions and Technical Services, 27* (2): 243-248.

Rieh, S. Y. (2004). On the Web at home: Information seeking and Web searching in the home environment. *Journal of the American Society for Information Science and Technology, 55* (8): 743-753.

Risvik, K. Michelsen, R. (2002). Search Engines and Web Dynamics. *Computer Networks, 39* (3): 289-302.

Rouse, M. E. (2004). Whatis.com. Tech Target [Online]. Accesible en: http://whatis.techtarget.com/ (25 de febrero, 2004)

Ruthfield, S. (2002). The Internet's History and Development: From Wartime Tool to the Fish-Cam. *ACM Crossroads* [Online]. Accesible en: http://www.acm.org/crossroads/xrds2-1/inet-history.html (5 de marzo, 2003).

White, R.W., Jose, J.M., Ruthven, I. (2003): Using Top-Ranking Sentences for Web Search Result Presentation. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Salazar García, I.(2002): La Red profunda. Lo que los buscadores convencionales no encuentran. *1er Congreso ONLINE del Observatorio para la CiberSociedad*. 9 de Septiembre de 2002. Barcelona.

Salton, G. McGill, M. (1983). Text Analysis and Automatic Indexing. In *Introduction to Modern Information Retrieval* (pp. 52-117). New York: McGraw Hill.

Salton, G. McGill, M. (1983). Retrieval Evaluation. In *Introduction to Modern Information Retrieval* (pp. 157-197). New York: McGraw-Hill.

Sanchez, J., Sandra, N., Fernández, L., Chevalier, G. (2002). Distributed Information Retrieval from Web-Accessible Digital Libraries using Mobile Agents. *UPGrade, 3* (3).

Sandeep Pandey Krithi Ramamritham(2003): Monitoring the Dynamic Web to respond to Continuous Queries. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Sanz, M. (1998). Fundamentos históricos de la Internet en Europa y en España. *Boletín de RedIRIS, 45* 22-36.

Savoy, J. (2002). Information Retrieval on the Web: A New Paradigm. *UPGrade, 3* (3).

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

Schwartz, M., Emtage, A., Kahle, B., Neumann, B. (1992). A Comparison of Internet Resource Discovery Approaches. *Computing Systems, 5* (4).

Schwartz, MF (1994). Harvest Software Available [Online]. Accesible en: http://groups.google.com/groups?q=harvest&start=20&hl=es&lr=&ie=UTF-8&oe=UTF-8&as_drrb=b&as_mind=12&as_minm=5&as_miny=1990&as_maxd=8&as_maxm=8&as_maxy=1999&selm=Pine.3.89.9411080806.N21666-0100000%40plains&rnum=30. (5 de noviembre, 2003)

Selberg, E (1995). MetaCrawler, a parallel meta-search engine [Online]. Accesible en: http://www.google.com/groups?q=metacrawler&hl=es&lr=&ie=UTF-8&oe=UTF-8&as_drrb=b&as_mind=12&as_minm=5&as_miny=1994&as_maxd=12&as_maxm=8&as_maxy=1997&selm=3u0qo7%24u0k%40big.aa.net&rnum=3. (12 de septiembre, 2003).

Selberg, E. Etzioni, O.(1995): Multi-Service Search and Comparison Using the MetaCrawler. *Fourth International World Wide Web Conference*. 11 de Diciembre de 1995. Boston, International World Wide Web Consortium.

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

Kamvar, S.D., Haveliwala, T.H., Manning, C.D., Golub, G.H. (2003): Extrapolation Methods for Accelerating PageRank Computations. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of

the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Shaw, R (1995). Crawlers, Spider s and Worms. *Web Week* (July, 1)

Shipeng Yu, Deng Cai, Ji-Rong Wen, Wei-Ying Ma(2003): Improving Pseudo-Relevance Feedback in Web Information Retrieval Using Web Page Segmentation. *Twelfth International World Wide Web Conference*. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Shiu, J. K. H., Chan, S. C. F., Chung, K. F. L. (2003). Accessing hidden web documents by metasearching a directory of specialty search engines. *Databases in Networked Information Systems, Proceedings, 2822:* 27-41.

Silverstein, C., Henzinger, M., Marais, H., Moricz, M (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum, 33* (1).

Simon Lok Min-Yen Kan(2003): Employing Natural Language Summarization and Automated Layout for Effective Presentation and Navigation of Information Retrieval Result. *Twelfth International World Wide Web Conference*. 20 de Mayo de 2003. Budapest: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Slone, D. (2002). The influence of mental models and goals on search patterns during Web interaction. *Journal of the American Society for Information Science and Technology, 53* (13): 1152-1169.

Smith, A. (2003). Testing the Surf: Criteria for Evaluating Internet Information Resources. *Public Access Computer Systems Review, 8* (3).

Spicer, D., Bell, G., Zimmerman, J., Boas, J., Boas, B. (2002). Internet History and Microprocessor Timeline. Computer History Museum [Online]. Accesible en: http://www.computerhistory.org/exhibits/internet_history/ (31 de enero, 2003)

Spink, A., Wolfram, D., Jansen, B., Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology, 52* (3): 226-234.

Spink, A., Jansen, B., Wolfram, D., Saracevic, T. (2002). From E-sex to E-commerce: Web search changes. *IEEE Computer, 35* (3): 107-109.

Spink, A., Ozmutlu, S., Ozmutlu, H., Jansen, B. (2002). US versus European Web searching trends. *SIGIR Forum, 36* (2).

Su, L. Chen, H.(1999): User evaluation of Web search engines. *3rd Conceptions of Library and Information Science Conference*. 23 de Mayo de 1999. Dubrovnik.

Su, L. (2003). A comprehensive and systematic model of user evaluation of Web search engines: I. Theory and background. *Journal of the American Society for Information Science and Technology, 54* (13): 1175-1192.

Su, L. (2003). A comprehensive and systematic model of user evaluation of Web search engines: II. An evaluation by undergraduates. *Journal of the American Society for Information Science and Technology, 54* (13): 1193-1223.

Sullivan, D (1998). Open Text Repositions Its Web Index. *Search Engine Report*. (March, 31).

Sullivan, D (2002). Death Of A Meta Tag. *Search Engine Report*. (October, 1)

Térmens Graells, R., Ribera Turró, M., Sulé Duesa, A. (2003). Nivel de accesibilidad de las sedes Web de las universidades españolas. *Revista Española de Documentación Científica, 26* (1): 21-39.

Thelwall, M. (2001). The Responsiveness of Search Engine Indexes. *Cybermetrics, 5* (1).

Thelwall, M. (2003). Can Google's PageRank be used to find the most important academic Web pages? *Journal of Documentation, 59* (2): 205-217.

Thomas, C. Griffin, L. (1999). Who will create the metadata for the Internet? *First Monday* [Online]. Accesible en: http://www.firstmonday.dk/issues/issue3_12/thomas/index.html (18 de abril, 2004)

Tomlin, J.(2003): A New Paradigm for Ranking Pages on the World Wide Web. *Twelfth International World Wide Web Conference*. 24 de Mayo de 2003. Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium: Computer and Automation Research Institute of the Hungarian Academy of Sciences ; International World Wide Web Consortium.

Tramullas Saz, J. Olvera Lobo, M. (2001). *Recuperación de la Información en Internet*. Madrid: Ra-Ma.

Travis, I. (1998). From "Storage and Retrieval Systems" to "Search Engines": Text Retrieval in Evolution. *Bulletin of the American Society for Information Science, 24* (4): 1.

Vaughan, L. New measurements for search engine evaluation proposed and tested. *Information Processing and Management,* (in press).

Vellucci, S. (1998). Metadata. *Annual Review of Information Science and Technology, 33:* 187-222.

Voorbij, H. (1999). Searching scientific information on the Internet: A Dutch academic user survey. *Journal of the American Society for Information Science, 50* (7): 598-615.

Wales, J. F. (2004). Wikipedia: The Free Encyclopedia. Wikipedia [Online]. Accesible en: http://en.wikipedia.org/wiki/Main_Page (25 de febrero, 2004)

Walton, B (1994). WWW and Gopher Statistics ? (Respuesta) [Online]. Accesible en: http://groups.google.com/groups?hl=es&lr=&ie=UTF-8&oe=UTF-8&selm=%25brucew.42.0%40sas-aux.byu.edu. (11 de marzo, 2003)

Wang, P., Wawk, W., Tenopir, C. (2000). Users' interaction with World Wide Web resources: An exploratory study using a holistic approach. *Information Processing and Management, 36* (2): 229 251.

Web Characterization Project (2003). Web Sites: Concepts, Issues and Definitions (First Draft). *OCLC Research* [Online]. Accesible en: http://wcp.oclc.org/pubs/rn1-websites.html (6 de mayo, 2003)

Webopedia (2004). Webopedia: Online Dictionary for Computer and Internet Terms. Jupitermedia Corporation [Online]. Accesible en: http://www.pcwebopedia.com/ (23 de febrero, 2004)

Weibel, S. (1995). Metadata: The Foundations of Resource Description. *D-Lib Magazine* [Online]. Accesible en : http://www.dlib.org/dlib/July95/07weibel.html. (2 de marzo, 2000).

Weibel, S., Ianella, R., Cathro, W. (1997). The 4th Dublin Core Metadata Workshop Report. *D-Lib Magazine* [Online]. Accesible en: http://www.dlib.org/dlib/june97/metadata/06weibel.html (3 de marzo, 2000).

White, R., Joemon, M., Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing and Management, 39* (5): 707-733.

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

Wiederhold, G. Genesereth, M. (1997). The Conceptual Basis for Mediation Services. *IEEE Expert, 12* (5): 38-47.

Wiggins, RW (1994). Statistics on growth in WAIS databases ? [Online]. Accesible en: http://groups.google.com/groups?q=wais+statistics&hl=es&lr=&ie=UTF-8&oe=UTF-8&selm=2p4vql%24e8d%40msuinfo.cl.msu.edu&rnum=4. (20 de octubre, 2003)

Winer, B. (1962). Design and Analysis of Single-factor Experiments. In *Statistical Principles in Experimental Design* (2nd ed ed., pp. 149-260). New York: McGraw-Hill.

Wolfram, D., Spink, A., Jansen, B., Saracevic, T. (2001). Vox populi: The public searching of the web. *Journal of the American Society for Information Science and Technology, 59* (12): 1073-1074.

Woodruff, A., Aoki, P., Brewer, E., Gauthier, P., Rowe, L. (1996). An Investigation of Documents from the World Wide Web. *Computer Networks and ISDN Systems, 28* (7-11): 963-980.

Wooldridge, M. Jennings, N. (1995). Intelligent Agents: Theory and Practice. *Knowledge Engineering Review, 10* (2): 115-152.

World Wide Web Consortium (1997). World Wide Web Consortium Publishes Public Draft of Resource Description Framework (RDF). World Wide Web Consortium [Online]. Accesible en: http://www.w3.org/Press/RDF (16 de septiembre, 2003)

World Wide Web Consortium (2001). Metadata at W3C. World Wide Web Consortium [Online]. Accesible en: http://www.w3.org/Metadata/ (15 de septiembre, 2003)

Yok, S.-H., Hawoong J, Barabasi, A. (2002). Modeling the Internet's large-scale topology. *Proceedings of the National Academy of Sciences USA, 99* (21): 13382-13386.

Zadeh, L. A. (2003). From search engines to question-answering systems - the need for new tools. *Advances in Web Intelligence, 2663* 15-17.

Zakon, R. (2003). Hobbes' Internet Timeline v6.0. Zakon Group [Online]. Accesible en: http://www.zakon.org/robert/internet/timeline/ (8 de febrero, 2003)

Zien, J., Meyer, J., Tomlin, J., Liu, J.(2001): Web Query Characteristics and their Implications on Search Engines.*Tenth International World Wide Web Conference*. 1 de Mayo de 2001. Hong Kong, International World Wide Web Consortium.

Zook, M. (2000). Internet Metrics: Using Host and Domain Counts to Map the Internet Globally. *Telecommunications Policy Online* [Online]. Accesible en: http://www.tpeditor.com/contents/2000/zook.htm (7 de abril, 2004)

Agradecimientos

Alejandro de la Cueva y **Josep Lluis Canet** (Universitat de València) no sólo aprobaron el plan de trabajo de este proyecto y lo han seguido atentamente hasta su conclusión. También están en su origen, cuando me encargaron el desarrollo de las sesiones sobre recuperación de información distribuida del curso de postgrado sobre "Organización, acceso y recuperación de información electrónica". **Víctor Castelo** y **Diego López** (RedIRIS) me facilitaron la muestra de sedes web españolas analizadas, después de que **Luis López** y **Arrate Baquedano** (Sección de Informática de RTVV) demostraran que no era viable el método de muestreo de IPs para la selección de sedes españolas y de que **Antonio Ortiz** (Fujitsu ICL España) me ofreciera acceso a la lista de dominios alojados en sus servidores. Como profesores de la Diplomatura en Biblioteconomía y Documentación, **Juan Vicente Giménez** y **Enrique Bonet** (Universtat de València) me pusieron en contacto con los alumnos participantes en el estudio de rendimiento y facilitaron las sesiones de expresión de búsquedas y de evaluación de los resultados. La preparación de esta prueba exigió la realización de una ronda inicial de búsquedas y evaluaciones, que me facilitó José Rodolfo Hernández-Carrión (Universtat de València). El tratamiento estadístico de los datos de este apartado lo realizó **Luis Izquierdo** (Instituto de Agroquímica y Tecnología de Alimentos).

Al elaborar su *Vocabulario*, **Luis Corbalán** (Unidad de documentación, RTVV) me facilitó sobremanera la expresión de los conceptos barajados a lo largo del proyecto y la elaboración del glosario. **Steve Lawrence** (NEC Research Institute, actualmente en Google) me facilitó las versiones electrónicas de algunos de sus ineludibles trabajos. **Lourdes Castillo** (Universitat de València) y **Ricardo Fornás** (metodosdebusca.com y buscopio.net) ha realizado una continua y muy atenta revisión de los textos desde sus primeras versiones. **Francisco Rico** (Instituto de Agroquímica y Tecnología de Alimentos) e **Inmaculada Mesa** (Servicio de Publicaciones, Universitat de València) han contribuido al diseño de esta versión impresa.

Anexo 1: Formulario remitido a los servicios de recuperación para recabar datos sobre sus características funcionales.

Descripción de servicios españoles de recuperación de información en Internet

La información recabada por este documento es de exclusiva aplicación en investigación. Las cuestiones se encuadran en los siguientes epígrafes

1. Información de identificación y contacto
2. Cobertura, volumen y evolución del servicio
3. Recopilación de documentos
4. Indización y organización
5. Recuperación
6. Presentación de resultados
7. Interface de usuario
8. Referencias y estadísticas

1. Identificación y contacto

1.1. Nombre del servicio

1.2. URL de la portada del servicio

1.3. Otras URLs y mirrors

1.3. Entidad propietaria o responsable del servicio

1.4.URL de la entidad

1.5. Dirección postal de la entidad

1.3. Persona que cumplimenta el formulario e-Mail

---

2. Cobertura, volumen y evolución del servicio

2.1. Número de documentos indizados por el servicio

< de 100.000

2.2. Número de URLs incluídas en el servicio < de 300.000

2.3. Fecha de inicio del servicio

2.4. Número de documentos incorporados semanalmente

2.5. Frecuencia de actualización (reindización) de la base de datos completa

Diaria

2.6. Cobertura geográfica del servicio España

2.7. Tipo de documentos recuperables:

- Páginas Web
- News
- Listas de distribución
- Elementos multimedia (MP3, MPEG y otros ficheros de audio o video)

---

3. Recopilación de documentos

3.1. Métodos de recopilación

- Automática mediante robots u otros programas
- Manual a cargo de los productores del servicio
- Manual por alta de los distribuidores de los documentos
- Existe revisión manual de las altas
- Existe examen de los documentos y recomendación de algunos

- Las bajas o eliminaciones dependen del servicio ☐
- Las bajas o eliminaciones dependen de los distribuidores de los documentos ☐

3.2. Nombre de los programas o robots empleados, en su caso, en la recopilación

3.3. Método de recopilación automática [Depth first]

3.4. ¿ Existe alguna base de datos externa que alimente las propias del servicio ?

4. Indización y organización

4.1. La indización se produce de forma [Automática]

4.2. ¿ Qué programas se emplean para indizar ?

4.3. ¿ Cuántas páginas se indizan de cada sede ?

[Sólo la portada o página principal]

4.4. ¿ Cuáles de los siguientes elementos se indizan en la base de datos ?

Título del documento ☐	Fecha del documento ☐
Autor del documento ☐	Tamaño del fichero ☐
Enlaces del documento ☐	Textos de los enlaces ☐
Autor del documento ☐	Textos de las imágenes ☐
Etiquetas meta del header ☐	Texto completo del documento ☐

4.5. ¿ Qué otra información descriptiva añade el servicio ?

4.6. ¿ Cómo se generan, en su caso, los resúmenes o descripciones de los documentos ?

[De forma automática a partir del propio documento]

---

5. Recuperación

5.1. Programa de recuperación

5.1.1. ¿ Qué programa de recuperación emplea el servicio ?

5.1.2. El modelo de recuperación es

- Booleano
- Best match
- Combinado de los dos anteriores
- Vectorial
- No textual (por enlaces)
- Otros

5.1.3. Estructura de las peticiones y operaciones que soporta

- Frases en lenguaje natural
- Listas de palabras (sin asociación booleana)
- Combinaciones booleanas: and or not
- Anidamiento (paréntesis)
- Operadores de proximidad:
- Distancia numérica entre palabras
- Coincidencia en la estructura del texto
- Frase exacta

5.1.4. Algoritmo de ordenación de resultados

5.1.4.1. ¿ Qué factores intervienen ?

5.1.4.2. ¿ Cómo se asignan las puntuaciones ?

5.1.4.3. ¿ Existe ponderación por el usuario ? Sí

5.2. Expresión de peticiones

5.2.1 Truncación

- No soportada
- Automática: Mediante stemming (morfológica) Con comodines (mecánica)
- Manual

5.2.2. Búsqueda por cadenas de caracteres

- Expresiones regulares ☐
- Posibilidad de enmascaramiento ☐
- Sensibilidad a mayúsculas y minúsculas ☐

5.2.3. ¿ Cuantos documentos como máximo recupera el servicio ? ☐

5.3. Elementos recuperables

5.3.1. La búsqueda por defecto se produce en:

URL ☐ Título ☐ Palabras clave ☐

Descripción o resumen ☐ Texto completo ☐ Texto de los enlaces ☐

Otros elementos ☐

5.3.2 ¿ Qué campos de búsqueda son definibles por el usuario ?

URL ☐ Título ☐ Palabras clave ☐

Descripción o resumen ☐ Texto completo ☐ Texto de los enlaces ☐

5.3.3. Lista de palabras vacías

- ¿ Existe ? Sí
- ¿ Cómo se ha construído ? ☐
- ¿ Se puede ignorar en la búsqueda por frase ? Sí

5.4. Recuperación avanzada

- Existe un vocabulario controlado ☐
- Hay posibilidad de expansión del perfil ☐
- Existe búsqueda conceptual ☐
- Existe query by example ☐

---

6. Presentación de resultados

6.1. Información sobre el conjunto resultante

- Total de documentos recuperados ☐

- Número de documentos en respuesta a cada término de búsqueda ☐

6.2. ¿ Qué elementos se visualizan ?

- URLs ☐
- Enlace al documento original ☐
- Título ☐
- Palabras clave ☐
- Resumen descriptivo ☐
- Texto íntegro ☐
- URLs citadas o textos de enlaces ☐
- Aciertos (hits) en su contexto ☐
- Tamaño del documento ☐
- Fecha del documento ☐
- Fecha de incorporación del documento al sistema ☐

6.3. ¿ Existen formatos de visualización predefinidos que el usuario activa ? ?

Sí

6.4. Otros elementos

- Se visualizan las puntuaciones de relevancia ☐
- Se visualizan los términos coincidentes con el perfil ☐

6.5. Postprocesamiento de resultados:

- Control de duplicados ☐
- Agrupamiento según URL ☐

---

7. Interface de usuario

1. Posibilidad de interface exclusivamente textual ☐
2. Niveles adicionales de búsqueda ☐
3. Ayuda online ☐
4. Instrucciones de búsqueda ☐

5. Visualización de búsquedas paralelas

8. Referencias y estadísticas
Cite algunas referencias recientes, en medios técnicos o populares, que describan las características de su servicio

¿ Dispone el servicio de estadísticas de acceso públicas ?. ¿ En qué URL ?

¿ Dispone el servicio de control de logs (conexiones y peticiones, ver también 7.5) ?. ¿ Es posible obtener copia de una muestra del mismo ?

Enviar formulario | Restablecer formulario