CAPÍTULO 1

INTRODUCCIÓN

Hace aproximadamente cincuenta años que se comenzaron a utilizar ordenadores para almacenar y tratar información textual. Se esperaba que esta automatización pondría el conocimiento al alcance de todos e impulsaría enormemente el avance científico puesto que la información estaría, literalmente, en la punta de los dedos de los usuarios. La aparición primero de Internet y más tarde de la Web parecían acercar aún más ese ideal de una "biblioteca universal". Desafortunadamente, las enormes capacidades de almacenamiento de los soportes informáticos y el talento de la especie humana para producir nueva información a un ritmo increíble han hecho de la Web no sólo la biblioteca más grande que haya existido hasta el momento sino también la más anárquica. Se han investigado decenas de propuestas para llevar algo de orden a la Web. La última de ellas pretende que ésta dé un paso más en su evolución hasta convertirse en una Web Semántica permitiendo que agentes software sean capaces de localizar datos, realizar razonamientos y construir nuevo conocimiento de manera autónoma. No obstante, no parece que el texto no estructurado vaya a desaparecer de la Web, es más, probablemente no deje de aumentar. Por esa razón será necesario ofrecer mecanismos complementarios a la Web Semántica a fin de ayudar a los usuarios a lidiar con esa otra "Web no-tan-Semántica". El autor ha propuesto a ese fin la Web Cooperativa que precisará el uso de técnicas ya existentes y la investigación de algunas nuevas. Una de esas últimas es una técnica desarrollada por el autor, sencilla, independiente del idioma y susceptible de ser aplicada a diversas tareas de tratamiento de lenguaje natural. La motivación, fundamentos y viabilidad de dicha técnica así como su eventual aplicación al problema de la sobrecarga de información son objeto de estudio en este trabajo.

1 Almacenamiento y tratamiento automatizado de información

Los primeros ordenadores electrónicos eran poco más que calculadoras no requiriendo dispositivos de almacenamiento externo demasiado sofisticados. Bastaba, tan sólo, disponer de algún método para conservar el código y quizás también los datos de tal forma que no fuese necesario introducirlos "manualmente" cada vez que se desease ejecutar el correspondiente programa.

Cintas y tarjetas perforadas servían perfectamente a ese objetivo, siendo las últimas las más utilizadas en las máquinas comerciales aun cuando la cantidad de información que se podía codificar en cada una fuese limitada. El formato de *IBM*, por ejemplo, permitía almacenar en el denominado "modo texto" 80 caracteres por tarjeta. Así, un bloque de tarjetas de una pulgada de grosor, que contenía unas 143, almacenaría 11.440 caracteres, o lo que es lo mismo, un texto de unas 1.800 palabras. Dicho de otro modo, utilizando semejante soporte una comunicación a un congreso "típica", alrededor de 5.000 palabras, ocuparía 18,73 x 8,26 x 7,06cm, algo más de mil centímetros cúbicos.

Así pues, almacenar grandes cantidades de texto en tarjetas perforadas, aunque posible, resultaba indudablemente incómodo. La Seguridad Social de los EE.UU. debió experimentar semejantes problemas al acercarse la década de los 50 cuando mantenía toda la información sobre los trabajadores del país en tarjetas perforadas puesto que, aparentemente, presionó a *IBM* (2002) para solucionar esta situación; lo cual llevó a la empresa a establecer en 1952 el estándar *de facto* para almacenamiento en cinta magnética (aunque el *UNIVAC I* ya había utilizado ese soporte en 1951).

No obstante, a pesar de sus ventajas sobre las tarjetas perforadas (menor volumen, mayor velocidad de acceso, o posibilidad de reescritura de datos), las cintas presentaban el problema del acceso secuencial. En 1956 *IBM* presentó el *RAMAC 305* (*Random Access Method of Accounting and Control*, Método de Acceso Aleatorio para Contabilidad y Control) que incluía una unidad de almacenamiento en disco magnético, el *IBM 350*, con acceso aleatorio y capacidad para 5 millones de caracteres (o lo que es lo mismo, 62.500 tarjetas perforadas, 2 rollos de cinta para la unidad *IBM 726* de 1952, o 750.000 palabras).

El desarrollo de ambos tipos de soporte continuó, aumentando de manera progresiva tanto la densidad de almacenamiento como la velocidad de acceso, relegando a las tarjetas perforadas a tareas de introducción de datos hasta su práctica desaparición a mediados de los años 70. En aquel momento un rollo de cinta podía almacenar 180MB de datos y la unidad de disco *IBM 3340* 70MB (el equivalente a 230.000 y 90.000 tarjetas perforadas, respectivamente).

Por otro lado, aunque esta época estuvo caracterizada, desde el punto de vista del tratamiento de datos, por el uso de *mainframes* para almacenar y procesar básicamente registros y transacciones, esto es, información estructurada, existía también una cantidad enorme de información textual con poca o ninguna estructura que crecía de un modo continuo y debía ser consultada con frecuencia (patentes, jurisprudencia, informes técnicos, memorandos, etc.)

A partir de mediados de los 50 y en especial en los años 60 se implementaron múltiples sistemas de búsqueda en distintas organizaciones. Según Madeline M. Henderson (1998) existían en 1966 sólo en EE.UU. más de 150 sistemas automatizados para la consulta de información textual. No es de extrañar pues que algunos de los trabajos más influyentes en el campo del tratamiento y recuperación de información surgieran precisamente en esta época. No obstante, hacer una revisión exhaustiva de los primeros años de investigación en dicho área va más allá del objetivo de este trabajo. Para proseguir la línea argumental bastará con exponer algunos hitos fundamentales; el lector interesado en el tema puede consultar el interesante trabajo de Mary Elizabeth Stevens (1970).

Puede considerarse a Hans Peter Luhn el pionero del área. Describió un método estadístico para codificar y, posteriormente, recuperar información textual[1] de forma totalmente automática (Luhn 1957) y una técnica para obtener resúmenes automáticos (Luhn 1958). Luhn proponía utilizar la frecuencia de aparición en el texto de las distintas palabras, obviando las poco frecuentes y las demasiado comunes, introduciendo así el uso de la frecuencia de los términos en cada documento y de listas de ***stop words*** ("palabras vacías[2]"). Ambas técnicas siguen en uso.

Poco después, Melvin E. Maron y J.L. Kuhns (1960) propusieron una alternativa aritmética a la búsqueda booleana (los términos de la **consulta** están o no presentes en los documentos) que permitiría calcular para cada documento una cifra que indicase su mayor o menor grado de **relevancia**[3] en relación con la consulta planteada. Ellos son los primeros en señalar que la ponderación de los distintos términos tanto en la consulta como en los documentos de la **colección** es fundamental y que es posible asignar un simple "número" a un documento para indicar su mayor o menor relevancia para una consulta dada.

Aparentemente, los "pesos" de cada término debían ser asignados manualmente, por el usuario en el caso de las consultas y por un "bibliotecario" en el de los documentos. No obstante, señalan que la relevancia de los términos con que se etiqueta un documento será inversamente proporcional al número de documentos etiquetados, algo que, posteriormente, se revelaría muy importante.

El modelo propuesto presentaba otra dificultad más: para cada documento de la colección, D_i, y cada posible término empleado en una consulta, t_j, se debe conocer la probabilidad de que un usuario en busca de información del tipo contenido en el documento D_i emplease el término t_j en su consulta. Así pues, aun cuando en el experimento descrito por Maron y Kuhns se utilizaron palabras clave extraídas de los propios documentos, en colecciones realmente grandes este proceso sería muy difícil. No obstante, y a pesar de estos inconvenientes de índole práctica, la importancia de las ideas planteadas en ese trabajo es indudable.

Desde mediados de los 60 Gerald Salton y su equipo desarrollaron el sistema de recuperación de información *SMART* (*System for the Mechanical Analysis and Retrieval of Text*, Sistema para el Análisis y Recuperación Mecánica de Texto) introduciendo toda una serie de conceptos de gran influencia posterior: el modelo vectorial de documentos, la utilización de la función coseno para comparar consultas con documentos (Salton y Lesk 1965),

---

[1] El problema básico de la recuperación de información textual consiste en la forma de representar un conjunto (o colección) de documentos no estructurados (texto libre) para facilitar posteriormente la localización de aquellos que satisfagan una necesidad de información de un usuario formulada mediante una consulta también textual.

[2] Aquellas palabras que, a pesar de un uso frecuente, aportan por sí solas poco significado a un texto (se muestran subrayadas algunas palabras vacías del castellano).

[3] La relevancia es una medida de la "proximidad" entre los contenidos de un documento y la necesidad de información planteada por un usuario en forma de consulta. Está claro que se tratará no sólo de un valor subjetivo sino también cambiante, por lo que el término no suele hacer referencia al "juicio" que emitiría un usuario sino al valor que un sistema automático asigna a cada documento en relación con una consulta. El objetivo de los sistemas de recuperación de información es producir valores de relevancia próximos a los que asignaría el propio usuario.

algoritmos de ***stemming***[1] o el uso de diccionarios de sinónimos y co-ocurrencias (Salton 1968).

Posteriormente, Karen Spärck-Jones (1972) introdujo la idea de que un término no sólo es relevante si aparece frecuentemente en un texto sino que es más valioso cuanto más raro, esto es, cuanto menor es el número de documentos de la colección en que aparece. Esto es lo que se conoce como ***idf*** (*inverse document frequency*[2]) que, al combinarse con la frecuencia de aparición de los términos (Luhn 1957), ha dado lugar a uno de las formas de ponderación de términos más conocidas y utilizadas, *tf*idf*, según la cual la relevancia de un término es directamente proporcional a la frecuencia de aparición en un documento e inversamente proporcional al número de documentos en que aparece. Hay que decir, sin embargo, que Maron y Kuhns (1960, p. 230) ya apuntaron en esa dirección aunque no incidieron en la posibilidad de obtener esos valores de forma automática[3]. Poco después, Stephen E. Robertson y Spärck-Jones (1976) desarrollarían un modelo probabilista para ponderar la relevancia de los términos de una consulta.

Así pues, a finales de los 70, tras dos décadas de investigación, el campo contaba con unas bases teóricas sólidas que ofrecían diversas técnicas para el desarrollo de sistemas de recuperación de información con un rendimiento adecuado. Hay que decir que hasta entonces todos estos sistemas se habían diseñado y evaluado con colecciones de documentos de tamaño conocido[4]; sin embargo, eso iba a cambiar.

2 Internet y la sobrecarga de información

En 1969 comenzó a operar la red *ARPANET* que evolucionaría en los años 80 hasta convertirse en lo que hoy conocemos como Internet. Esta última, al acomodar otras redes de intercambio de información (como *USENET*[5]) y ofrecer soporte para la Web (que, finalmente, ha dado acceso a la práctica totalidad de servicios integrados en Internet) ha

---

[1] Un algoritmo de *stemming* o *stemmer* determina la raíz morfológica de una palabra colapsando múltiples formas de la raíz en un único término (`research`, `researcher`, `researching` y `researchers` colapsan en `research` empleando un *stemmer* para inglés). Un *stemmer* para castellano, por ejemplo, transformaría `andanzas` en `and`, `habitaciones` en `habit` o `juguéis` en `jug`.

[2] El término *inverse document frequency* se ha traducido como "frecuencia inversa del documento", "frecuencia documental inversa", "frecuencia inversa de documentos" o "frecuencia inversa en el documento". Sin embargo estas traducciones son incorrectas y, peor aún, confusas. La medida *idf* trata de ponderar el valor informativo de un término basándose en su frecuencia de aparición en distintos documentos de una colección, a mayor frecuencia de uso menor valor y viceversa. Por ejemplo, `una` aparece en 5 millones de páginas web españolas frente a las 3 que mencionan `folksonomía`; está claro, que el primer término es mucho menos informativo que el segundo. Dicho de otro modo, el valor de un término es inversamente proporcional al número de documentos en que aparece.

[3] *"Un término índice aplicado a cada documento de la biblioteca no tendrá significatividad, mientras que uno aplicado a sólo un documento será altamente significativo. Así pues, las medidas de la significatividad están relacionadas con un "indicador de extensión" para cada término, es decir, con el número de documentos etiquetados con el término –cuanto más pequeño sea este número, mayor será la significatividad del término índice.)"* (Maron y Kuhns 1960, p. 230)

[4] Para el proyecto *Cranfield II* se elaboró una colección con 1.400 documentos (Cleverdon y Keen 1966), la colección *NPL* contenía 11.429 (Spärck-Jones y Webster 1979) y la *CACM* 3.204 (Fox 1983).

[5] *USENET* permite a los usuarios publicar, leer y comentar mensajes de texto (artículos) sobre distintos temas organizados en grupos jerárquicos.

contribuido en enorme medida, junto con el tremendo abaratamiento del soporte en disco magnético[1], a la explosión de información textual que se vive hoy en día.

Maron y Kuhns (1960, p. 217) afirmaban que los datos documentales estaban siendo generados a un ritmo alarmante. Cuatro décadas después es difícil encontrar un superlativo para "alarmante" que dé idea de la tasa de producción de información que ha alcanzado la humanidad, tan sólo algunas cifras:

- Desde 1981 se han generado más de 845 millones de mensajes en *USENET*[2], lo cual supone una media de más de 100.000 mensajes nuevos al día.
- La Web superficial consta, al menos, de 4.000 millones de documentos[3], la Web oculta (Florescu, Levy y Mendelzon 1998), es decir, aquellas páginas accesibles sólo tras rellenar algún tipo de formulario, tendría según Isidro F. Aguillo (2002) entre 2 y 50 veces el tamaño de la primera y según Michael K. Bergman (2001) hasta 500.
- *Reuters* produce alrededor de 11.000 artículos de prensa diarios[4] (aproximadamente 2,5 millones de palabras) que ofrece de manera *online* mediante documentos *NewsML*[5].
- *Springer Verlag* editó 356 volúmenes de su serie *Lecture Notes in Computer Science* en 2003[6] (alrededor de 90 millones de palabras). Todos los artículos están disponibles como archivos *PDF*[7]).

La lectura de estos datos parece evocar una mezcla de imágenes: un ejército de tipógrafos escribiendo, no los volúmenes de la Biblioteca del Museo Británico, sino de la Biblioteca de Babel. Como se verá, la realidad no está muy lejos de la ficción.

Actualmente cualquier usuario puede publicar en *USENET* y, a poco que lo desee, en la Web lo cual supone alrededor de 400 millones[8] de potenciales "efectivos". Según un estudio realizado en 2003 por *Pew/Internet* en EE.UU. entre usuarios adultos (18 años o más) un 44% ha contribuido de una forma u otra a incrementar los contenidos disponibles[9], el 10% ha publicado en grupos de noticias (*USENET*), el 13% mantiene sitios web y el 2% bitácoras[10] (Lenhart, Horrigan y Fallows 2004).

---

[1] Una unidad de disco *IBM 1301* tenía una capacidad aproximada de 27MB y costaba, en 1961, 115.500 dólares (*IBM* 1994-2004). Un disco de 40GB de la misma marca costaba en 2004 170 dólares. Teniendo en cuenta la inflación (Sahr 2004), el precio del primer sistema equivaldría a 725.000 dólares actuales por lo que el coste por megabyte ha descendido desde 26.850 dólares a 0,4 centavos en 43 años.

[2] Fuente: *Google* (`http://www.google.com`)

[3] Fuente: *Google* (`http://www.google.com`)

[4] Fuente: *Reuters* (`http://www.reuters.com`)

[5] *NewsML* (*News Markup Language*, Lenguaje de Etiquetado de Noticias) es un vocabulario XML para "empaquetar" contenidos periodísticos así como para añadirles metainformación.

[6] Fuente: *Springer Verlag* (`http://www.springeronline.com`)

[7] *Portable Document Format* (Formato de Documento Transportable).

[8] En 2001 había alrededor de 119 millones de usuarios de Internet en la Unión Europea (David 2003), 143 en EE.UU. (*USDOC* 2002), 47 en Japón (*MPHPT* 2001) y, en 2003, 68 millones en China (*CNNIC* 2003).

[9] Compartiendo archivos, anotando comentarios en sitios web y bitácoras, respondiendo artículos en *USENET*, etc.

[10] Una bitácora, *weblog* o su contracción *blog* es un sitio web que contiene, en orden cronológicamente inverso, artículos publicados por una persona (en ocasiones por grupos) sobre los más diversos temas empleando un sistema de gestión de contenidos. Las bitácoras tienen dos características esenciales, suelen

De acuerdo con (*USDOC* 2002) de los 143 millones de usuarios en EE.UU. 95 millones son adultos. Combinando este dato con el informe anterior tendríamos que, sólo en dicho país, 10 millones de usuarios publican en *USENET*; 12 millones mantienen sitios web y casi 2 millones bitácoras. Si las mismas ratios demográficas y actitudes de uso fuesen extensibles a la "población total" de Internet tendríamos 27 millones de usuarios que habitualmente publican en *USENET*, 35 millones de *webmasters* y 5 millones y medio de *bloggers*.

Naturalmente, los datos no son directamente extrapolables aunque es razonable suponer que Europa y Japón exhiben comportamientos similares. Por ello, la cifra de 5 millones de creadores habituales de contenido textual parece bastante razonable como cota inferior y la imagen de un formidable ejército tecleando creíble.

Por otro lado, la Web[1] se asemeja bastante a la Biblioteca que describió Borges. Su tamaño real es desconocido. No todos los contenidos disponibles son realmente interesantes[2] y, lo que es más, a pesar de que muchas personas, en particular estudiantes (Grahan y Metaxas 2003), lo den por hecho tampoco son necesariamente veraces.

Sin embargo, ni el tamaño de la Web ni la veracidad de sus contenidos son problemas que se vayan a abordar en este trabajo sino uno más específico, a saber, la sobrecarga de información que sufren todos los usuarios que utilizan la Web como fuente de información. No obstante, para ahondar en la naturaleza del problema es necesario analizar por qué y cómo surgió la Web, cuál es el motivo por el que crece a un ritmo tan elevado y qué soluciones se han propuesto para localizar información en el aluvión disponible.

3 La Web como sistema de recuperación de información

Tim Berners-Lee (1989) inició el desarrollo de la Web en el *CERN*[3] como un medio para evitar la pérdida de información, inevitable en una organización de gran tamaño, y facilitar el acceso a la información disponible (bases de datos, directorios telefónicos, etc.) Dos características de la propuesta original permitieron transformar el proyecto original en la Web actual: su naturaleza distribuida (los documentos pueden residir en máquinas distintas) y la posibilidad de establecer vínculos (enlaces) entre documentos. Por otro lado, Berners-Lee insistía en la necesidad de construir un sistema que animase a los usuarios a incorporar nueva información, haciéndolo así aun más útil y atractivo, de tal forma que el conjunto de documentos creciese de forma continua.

Berners-Lee hace algunas reflexiones muy interesantes sobre posibles problemas y métodos para recuperar información en un sistema como el que proponía. Alerta, por ejemplo, sobre los inconvenientes de utilizar palabras clave para localizar documentos y sugiere la posibilidad de establecer enlaces, no sólo con documentos, sino también con

---

ser muy personales y los lectores pueden anotar los artículos con sus propios comentarios. La persona que mantiene una bitácora es un *blogger* y la totalidad de *blogs* es conocida como blogosfera.

[1] Entendiendo la Web como el conjunto formado por la Web superficial, la oculta y todos los servicios disponibles y/o accesibles vía web ahora y en el futuro (bitácoras, *USENET*, periódicos digitales, publicaciones científicas en formato electrónico, enciclopedias, foros, etc.)

[2] Como diría Borges *"por una línea razonable o una recta noticia hay leguas de insensatas cacofonías, de fárragos verbales y de incoherencias"*.

[3] *Centre Européen pour la Recherche Nucléaire*, Centro Europeo para la Investigación Nuclear.

conceptos facilitando la existencia de enlaces "indirectos" entre documentos de temática similar.

Por tanto, la propuesta original planteaba construir la Web sobre una base semántica más o menos sólida, partiendo de nodos conceptuales enlazados desde los distintos documentos. Por otra parte, para explotar las ventajas de los enlaces indirectos antes mencionados, los enlaces entre documentos y conceptos deberían ser bidireccionales. No obstante, Berners-Lee (1989) no hace ninguna mención explícita sobre enlaces bidireccionales, tan sólo en (Berners-Lee 1990) se plantea su utilidad aunque también señala que un programa que recorriese una Web monodireccional podría obtenerlos simplemente "invirtiendo" los enlaces que hubiese encontrado. Finalmente, las primeras versiones del lenguaje *HTML*[1] (*HyperText Mark-up Language* 1992) establecieron de manera permanente la unidireccionalidad de los enlaces, esto es, si un documento *A* enlaza a un nodo *B* es posible "llegar" a *B* partiendo de *A* e imposible[2] hacer el recorrido contrario comenzando en *B*.

Además, en ninguna de las versiones desarrolladas se incluyó nada similar a los "nodos conceptuales". Las primeras versiones del lenguaje *HTML* (*HyperText Mark-up Language* 1992) permitían dar título a los documentos, formatear texto (párrafos, listas, cabeceras, etc.), crear enlaces a otros recursos y, de un modo rudimentario mediante la etiqueta `ISINDEX`, crear interfaces con programas auxiliares para realizar búsquedas de documentos en cada servidor (mediante palabras clave). Como consecuencia, la Web se convirtió en un artefacto diseñado para crecer de un modo cada vez más acelerado sin proporcionar mecanismos adecuados para localizar información[3].

4 Los primeros directorios y motores de búsqueda[4]

El primer servidor web (`info.cern.ch`, antes `nxoc01.cern.ch`) entró en funcionamiento en 1990 y para finales de 1992 existían alrededor de veinte (Berners-Lee 1992c). En aquel momento resultaba relativamente sencillo mantener de manera manual un directorio de sitios web y, de hecho, organizaciones como el *CERN* o el *NCSA*[5]

---

[1] *HyperText Mark-up Language* (Lenguaje de Etiquetado Hipertextual) es un lenguaje de etiquetas basado en *SGML* que permite construir páginas web. El estándar para este lenguaje es mantenido por el *Consorcio W3* (`http://www.w3.org`).

[2] Es cierto que existen herramientas como *Google* (`http://www.google.com`) que permiten encontrar documentos que enlazan con uno específico (por ejemplo, la consulta `link:http://www.w3.org` proporciona como resultado documentos que enlazan con el sitio del *Consorcio W3*). Sin embargo, esto es posible tan sólo después de haber explorado una porción de la Web construyendo un grafo que la represente y permita un análisis posterior como sugería Berners-Lee (1990), no se trata de una característica intrínseca de la Web.

[3] Tim Berners-Lee (1992a) sugiere utilizar documentos índice (aquellos que emplean la etiqueta `ISINDEX`) que, a su vez, apuntasen a otros índices más específicos y en (Berners-Lee 1992b) plantea utilizar la estructura formada por los enlaces para encontrar información relevante. Tan sólo son unos breves apuntes y las ideas están esbozadas de un modo rudimentario pero señalan las dos líneas que posteriormente se seguirían para desarrollar mecanismos de búsqueda en la Web: índices (por ejemplo, *Yahoo!*) y buscadores basados en robots (por ejemplo, *Google*). A pesar de todo, los métodos para localizar información en la Web son externos a la misma.

[4] Un motor de búsqueda, o simplemente buscador, es un artefacto *software* que explora la Web almacenando en una base de datos parte o todo el texto de los documentos que analiza. Al ir procesando documentos se crea un índice que emplea las palabras que aparecen en cada página web. Cuando un buscador recibe una consulta toma las palabras utilizadas por el usuario y obtiene los documentos indexados por las mismas.

[5] *National Center for Supercomputing Applications*, Centro Nacional para Aplicaciones de Supercomputación.

gestionaban índices a los que iban añadiendo las notificaciones de nuevos sitios que recibían por correo electrónico.

Sin embargo, según un estudio realizado por Matthew Gray (1995), a finales de 1993 había 623 servidores web y su número se duplicaba cada 3 meses, existiendo en diciembre de 1994 más de 10.000. Esto hacía muy difícil, aunque no imposible, mantener manualmente un índice de sitios web, dejando a un lado el hecho de que muchos administradores no notificaban su existencia a los directorios existentes (Steinberg 1996, p. 2). Así, la carencia de un sistema adecuado para poder localizar los distintos servidores y documentos en la incipiente Web era ya un problema[1] y comenzaron a desarrollarse distintos sistemas en busca de soluciones.

Tan sólo en *WWW94* (*First International Conference on the World-Wide Web*, Primer Congreso Internacional sobre la Web) se presentaron nueve trabajos relativos al indexado automático de documentos y a la búsqueda de información. Entre ellos cabe destacar los sistemas creados por Martijn Koster (Koster 1994) y Oliver A. McBryan (McBryan 1994), *ALIWEB* y *WWW Worm*, respectivamente. Ambos desarrollaron programas para explorar la Web de manera automática[2], saltando de enlace en enlace y almacenando información sobre las páginas visitadas en una base de datos para su posterior consulta por parte de los usuarios.

ALIWEB (Koster 1994) comenzaba su exploración a partir de sitios web registrados manualmente por sus administradores almacenando una información relativamente escasa para cada documento indexado (título, descripción y algunas palabras clave) lo que limitaba las posibilidades de los usuarios al realizar sus consultas[3].

En el caso de *WWW Worm* (McBryan 1994) no queda muy claro cómo se construía la base inicial de sitios web para realizar el indexado, la información almacenada era aún más parca (título del documento y textos utilizados en los enlaces que apuntan al mismo) y se consultaba (internamente) mediante la orden *UNIX* `egrep`[4].

Otros sistemas destacables, similares a los anteriores y desarrollados en la misma época fueron *Jumpstation*, *Wanderer*, *WebCrawler* y *Lycos*.

Jumpstation, implementado por Jonathon Fletcher (1994), fue uno de los primeros motores de búsqueda. Entró en funcionamiento en diciembre de 1993 y lo hizo de manera errática hasta quedar desatendido en abril de 1994.

---

[1] A lo largo de 1993 tuvo lugar una interesante discusión en la lista de correo *WWW Talk* sobre distintas formas de enfocar la localización de recursos en la Web así como los retos que plantearía: recorrer todos los enlaces almacenando la información encontrada (Perry 1993), la necesidad de grandes recursos temporales y de almacenamiento así como la conveniencia de no visitar todos los enlaces posibles (Johnson 1993) o los problemas que surgirían con enlaces generados automáticamente (Putz 1993). Por otro lado, Thomas R. Bruce (1993) planteaba un enfoque no automático mediante el cual los distintos sitios web deberían "registrarse" para ser revisados por "editores" que los categorizarían, pudiendo dividirse las categorías en subgrupos que serían asignados a nuevos editores; el mismo planteamiento que, algunos años después, sigue *ODP* (*Open Directory Project*, Proyecto de Directorio Abierto, `http://www.dmoz.org`)

[2] *Spiders*, "arañas", también denominados robots.

[3] La probabilidad de coincidencia entre dos individuos (por ejemplo, entre el autor de un documento y un potencial lector) en el uso de la misma palabra para identificar un concepto está entre el 10 y el 20% (Furnas *et al.* 1987).

[4] `egrep` permite realizar búsquedas en un fichero de texto empleando expresiones regulares, como resultado muestra todas las líneas que contienen el patrón recibido como parámetro.

Wanderer fue inicialmente desarrollado para descubrir nuevos sitios web, posteriormente se utilizó para medir la expansión de la Web entre junio de 1993 y junio de 1995 (Gray 1995) y, finalmente, para construir el buscador *Wandex* (*Wanderer Index*).

WebCrawler (Pinkerton 1994) supuso una mejora respecto a *ALIWEB* o *WWW Worm* puesto que indexaba todo el texto de las páginas que exploraba. Esta estrategia permitía ofrecer más documentos para las consultas de los usuarios pero, al aumentar el número de páginas indexadas, reducía de manera drástica la **precisión**[1] de las respuestas.

Lycos (Mauldin y Leavitt 1994) constituyó una iniciativa intermedia entre *ALIWEB* y *WebCrawler* puesto que no indexaba el texto completo de los documentos ni únicamente su título y descripción. En su lugar generaba una versión "ligera" constituida por el título, las veinte primeras líneas y las cien palabras más relevantes[2].

Este tipo de sistemas automáticos podían enfrentarse al enorme crecimiento de la Web en mejores condiciones que los índices construidos de forma manual. Sin embargo, estos últimos ofrecían otras ventajas (organización en categorías jerárquicas, posible revisión por parte de "editores" especializados, referencias cruzadas, etc.) que también eran valoradas por los usuarios (Bruce 1993). Por ejemplo, *Galaxy, Yahoo!* y *ODP* se construyeron siguiendo esta línea.

Galaxy (Speyer y Allen 1994) empezó a dar servicio a comienzos de 1994. Según el anuncio original se habían empleado métodos semiautomáticos para construir la base de datos original (que incluía no sólo páginas web sino también servidores *Gopher*[3] y *WAIS*[4]). *Galaxy*, como sería común en todos los directorios posteriores, permitía que los administradores de sitios web notificasen la dirección de su sitio para su inclusión en una categoría previamente seleccionada entre las disponibles en la jerarquía, posteriormente, un editor revisaba el sitio web y decidía acerca de su inclusión en el directorio.

El sitio web precursor de *Yahoo!* (Filo y Yang 1994), *Jerry's Guide to the World Wide Web* – Guía de Jerry a la Web, fue creado a comienzos de 1994 como un proyecto personal y se transformó en un directorio comercial en 1995. Al igual que *Galaxy*, dispone de categorías predefinidas en las que los administradores de sitios web pueden solicitar la inclusión que, también, es revisada por empleados de la empresa.

ODP[5] (*Open Directory Project*, Proyecto de Directorio Abierto), antes *DMoz* (*Directory Mozilla*, Directorio Mozilla), antes *NewHoo* y, aún antes, *GnuHoo*[6], fue fundado en 1998. Su

---

[1] La precisión es la proporción entre el número de documentos relevantes retornados por un sistema para una consulta y el total de documentos retornados.

[2] Aplicando *tf*idf* como método de ponderación (véase página 4).

[3] *Gopher* es un protocolo para la búsqueda y recuperación de información textual en Internet, fue desarrollado en la Universidad de Minnesota (EE.UU.) y distribuido públicamente a partir de 1991. El objetivo de *Gopher* era muy similar al de la Web aunque resultaba menos flexible que HTML. Todavía existen algunos servidores *Gopher* activos (menos de 300 en Agosto de 2004, fuente: *Floodgap* <http://gopher.floodgap.com>) pero puede afirmarse que ha sido totalmente reemplazado por la Web.

[4] *WAIS* (*Wide Area Information Servers*, Servidores de Información a Nivel Global) era un sistema para búsqueda remota de texto en bases de datos distribuidas basado en el estándar *ANSI Z39.50*. Hasta donde sabe el autor no existe ningún servidor *WAIS* operativo en la actualidad.

[5] http://www.dmoz.org

[6] El cambio de *GnuHoo* a *NewHoo* estuvo motivado, aparentemente, por un artículo (Miller 1998) publicado en *Slashdot* que criticaba el uso del acrónimo *GNU* (vinculado a proyectos de *software* no

estructura y funcionamiento es similar a la de *Galaxy* o *Yahoo!* con la diferencia de que los editores no forman parte de la plantilla de la empresa sino que realizan su labor de manera desinteresada.

Así pues, en torno a 1998 ya existía toda una serie de recursos para la búsqueda de información en la Web que, sin embargo, seguían siendo insuficientes. Por un lado, no parecía que los directorios como *Yahoo!*, al utilizar editores humanos, pudiesen clasificar todos los sitios web existentes (mucho menos todas las páginas) al mismo ritmo que aparecían (Steinberg 1996). Por otro lado, aunque algunos expertos[1] afirmaban que era posible indexar la Web al mismo ritmo que crecía, otros, como Steve Lawrence y C. Lee Giles (1998), discrepaban. Éstos, tras analizar la "cobertura" de distintos buscadores[2] encontraron que ninguno cubría, individualmente, más de un tercio de la Web indexable[3] conocida y que la combinación de varios buscadores (en su caso seis) podía cubrir más del triple de páginas que un único sistema.

Lawrence y Giles (1998) concluyen que el uso de metabuscadores era una solución para localizar información en la Web puesto que garantizaba la cobertura del mayor número posible de páginas. Uno de los primeros metabuscadores, de hecho anterior a su estudio, fue *MetaCrawler* (Selberg y Etzioni 1995). Selberg y Etzioni señalan dos deficiencias fundamentales en los buscadores de aquel momento: (1) la porción de la Web sobre la que trabaja cada buscador es distinta del resto obligando a los usuarios a repetir la consulta en distintos buscadores y (2) gran parte de los resultados son irrelevantes o enlaces "muertos". *MetaCrawler* pretendía dar solución al primer problema ofreciendo una interfaz única para los distintos buscadores (esto es, distintas bases de datos) y el segundo filtrando los resultados recibidos.

Hay que señalar, sin embargo, que ni los argumentos que se presentan para señalar el primer problema ni los criterios que emplean para "evaluar" la relevancia de los documentos son totalmente acertados (Lawrence y Giles 1998, p. 98). Esto en absoluto invalida la propuesta de Selberg y Etzioni; de hecho, los metabuscadores siguen siendo comunes[4]. Sin embargo, no son la solución al problema de la relevancia ni tampoco al de la cobertura puesto que, en rigor, bastaría un solo buscador común que indexara un número suficiente de páginas web.

5 Motores de búsqueda modernos

En 1998 los buscadores existentes no cubrían, individualmente, toda la Web ofreciendo los metabuscadores una solución simple e inmediata para ese problema (Lawrence y Giles 1998). Sin embargo, el problema también podría solventarse si los

---

privativo) por parte de una empresa privada para denominar una alternativa a *Yahoo!* construida mediante el uso de voluntarios en lugar de personal propio.

[1] Eric A. Brewer, fundador y responsable técnico de *Inktomi* (actualmente parte de *Yahoo!*) afirmó en una entrevista (Steinberg 1996, p. 5) que era posible indexar la totalidad de la Web, al menos, hasta 2000.

[2] Los buscadores que analizaron fueron *AltaVista*, *Excite* (que había adquirido *WebCrawler*), *HotBot*, *Infoseek*, *Lycos* y *Northern Light*. Ninguno era un directorio, todos utilizaban robots.

[3] Aquella parte de la Web accesible para un robot, es decir, páginas web accesibles sin formularios, contraseñas o que no les están "vedadas" (existen una serie de estándares para evitar que los robots de búsqueda no accedan a ciertas partes de un sitio web).

[4] En Marzo de 2004 existían alrededor de 30 metabuscadores activos. Fuente: *SearchEngineWatch* <http://searchenginewatch.com>

buscadores cubriesen porciones de la Web mayores (idealmente su totalidad), algo teóricamente factible.

Por otro lado, no es de extrañar que búsquedas realizadas sobre una base documental con, al menos, 320 millones de *ítems* (Lawrence y Giles 1998) proporcionasen resultados de una relevancia mediocre al realizar búsquedas por palabras clave con esquemas de ponderación muy simples. Sin embargo, gracias a la naturaleza del hipertexto, podía tratar de encontrarse una solución empleando métodos análogos a los aplicados a un tema antiguo: la cita de trabajos científicos.

Una característica de los textos científicos es la referencia a trabajos de terceros. El estudio de los patrones subyacentes a las referencias en revistas científicas ha llegado a convertirse en una rama del conocimiento con, al menos, 75 años de antigüedad (Lotka 1926), (Gross y Gross 1927), (Brodman 1944) o (Fussler 1949) y hace algo más de treinta que se desarrolló el concepto de "índice de impacto" (Garfield 1972) para determinar el "prestigio" de las distintas publicaciones.

La idea tras dicho índice es muy simple: partiendo de una base de datos que almacene para cada trabajo científico su título, la revista en que fue publicado y la lista de obras que cita es posible obtener el número de referencias hechas a un trabajo, autor o revista. Obviamente, los trabajos más antiguos, los autores más prolíficos o las revistas con más artículos o números por año serán, probablemente, más citados a pesar de que otros trabajos, autores y publicaciones pueden ser tanto o más relevantes que los primeros. Así pues, es necesario algún tipo de "normalización" del número total de citas a fin de obtener una medida de índice de impacto "justa".

Sin embargo, no es necesario entrar en mayores detalles para encontrar paralelismos entre el problema de la relevancia de los trabajos científicos y la relevancia de los documentos en la Web. En el primer caso por medio de las referencias y en el segundo mediante los enlaces se establecen vínculos entre documentos que indican, implícitamente, que el autor que establece dicho vínculo considera al documento enlazado tanto o más relevante que el suyo propio[1] o una fuente autorizada sobre un tema en particular. Por otro lado, el contexto en que se hace la cita (o el texto que rodea y se utiliza en el enlace) aportan información muy valiosa sobre el contenido del documento referenciado. Ideas similares a estas podían ser aplicadas inmediatamente a las bases de datos construidas por los robots al explorar la Web aprovechando las características del hipertexto.

Jon Kleinberg (1998) sentó las bases sobre las que se apoyan los modernos buscadores al plantearse la viabilidad de un método algorítmico para estimar la relevancia de un documento, algo que según él era una característica subjetiva. Para ello definió los conceptos de **"autoridad"** y ***hub*** (concentrador). Una autoridad es un documento al que enlazan muchos otros puesto que, según Kleinberg, cada enlace recibido es un "voto"

[1] En algunas ocasiones se referencia un trabajo para criticarlo duramente o se establece un enlace a un sitio web por motivos maliciosos (véase más adelante). Puesto que el uso pretendido, y habitual, de referencias y enlaces tiene connotaciones positivas el abuso tiene efectos perturbadores. En el primer caso (referencias científicas) un trabajo polémico puede ser muy citado y, consecuentemente, valorado como muy relevante. En el segundo se puede abusar de los mecanismos de un motor de búsqueda para construir complicadas "bromas". Por ejemplo, al formular la consulta `ladrones` en *Google* se obtiene como primer resultado el sitio web de la *Sociedad General de Autores y Editores* [17 agosto 2004]. Esto se conoce como *Google Bomb* – Bomba *Google* y requiere un esfuerzo coordinado de varios sitios web distintos para lograr su objetivo. Este tipo de hechos, a pesar de encerrar cierta malicia, son difícilmente clasificables como "ataques".

emitido por el individuo que estableció dicho enlace. Analizando el texto empleado en los enlaces puede determinarse el contexto en el cual el documento enlazado es una autoridad. Por su parte, un concentrador será un documento que contiene enlaces a muchas autoridades y es, por tanto, un recurso valioso para localizar información relevante en la Web.

Estos conceptos fueron probados por Chakrabarti *et al.* (1998a) y (1998b) mediante varios prototipos que tenían como objetivo localizar únicamente los documentos más relevantes para cada consulta, esto es, las autoridades. Para evaluar el rendimiento de estas técnicas se realizaron una serie de consultas genéricas empleando dichos prototipos, *Yahoo!* (un directorio) y *Altavista* (un buscador basado en robots) obteniendo, en cada caso, los diez documentos más relevantes. Posteriormente, un grupo de usuarios evaluó de manera "ciega" cada documento y valoró su relevancia en relación con la consulta planteada. La relevancia media de los resultados proporcionados empleando la técnica de Kleinberg superaba el 50% frente al 40% de *Yahoo!* y el 20% de *Altavista* abriendo la posibilidad de construir automáticamente taxonomías de documentos similares a las construidas por expertos humanos.

Existen, sin embargo, tres escenarios (Bharat y Henzinger 1998) en los que la técnica de Kleinberg puede ser objeto de abuso o simplemente falla por basarse en suposiciones no totalmente correctas. Se trata de relaciones entre servidores "mutuamente fortalecedoras", enlaces generados automáticamente y documentos irrelevantes enlazados desde autoridades o concentradores.

Relaciones entre servidores "mutuamente fortalecedoras". El algoritmo de Kleinberg cuenta cada enlace como un voto diferente; de este modo, si varios documentos alojados en un único servidor apuntan a un único documento externo éste recibe muchos "votos" que Bharat y Henzinger consideran "fraudulentos". Para solucionarlo plantean la necesidad de reducir el peso otorgado a los enlaces que parten desde un único servidor a un único documento. No obstante, puesto que un servidor puede alojar múltiples sitios web (páginas personales por ejemplo) su solución es simplista y devalúa "votos" independientes por cuestiones meramente topológicas. Brian D. Davison (2000a) realiza un estudio mucho más riguroso sobre el difícil problema de los enlaces "nepotistas".

Enlaces generados automáticamente. Según Kleinberg un enlace es un "voto" emitido por un individuo a favor de la relevancia de un documento. Sin embargo, existen enlaces que no son creados por seres humanos sino generados automáticamente con lo cual ya no son "votos" válidos. Por ejemplo, al crear una página personal o una bitácora en algún servidor gratuito todos los documentos tendrán enlaces a la página principal del servicio o de los patrocinadores. Tales enlaces no pueden diferenciarse de los creados por una persona y se valorarán de igual modo, afectando de un modo difícil de determinar a los resultados. Bharat y Henzinger no ofrecen solución a este problema.

Documentos irrelevantes enlazados desde autoridades o concentradores. Se supone que si una autoridad o un concentrador enlazan un documento éste debe ser necesariamente una autoridad sobre el tema tratado en los documentos de partida. Sin embargo, esta suposición no siempre es cierta; por ejemplo, las páginas personales de los autores de documentos muy referenciados no tienen por qué ser necesariamente relevantes. Bharat y Henzinger proponen analizar el contenido de las páginas enlazadas para comprobar si realmente tienen relación con el tema tratado en el documento del que parte el enlace. Para mejorar el rendimiento de su analizador no emplean palabras clave sino raíces obtenidas mediante el algoritmo de *stemming* de Porter (1980). Hay que decir que la idea de

analizar los contenidos es atractiva pero lo cierto es que este planteamiento es poco escalable en un entorno multilingüe como la Web (habría que desarrollar un *stemmer* para cada idioma).

En 1998 comenzó a operar el buscador, tal vez, más popular de la actualidad: *Google* (Brin y Page 1998). Éste, al igual que los motores de búsqueda "tradicionales", emplea robots para explorar la Web en búsqueda de documentos pero, al contrario que estos, utiliza una técnica mucho más sofisticada para organizar los resultados de las consultas de los usuarios: el algoritmo ***PageRank*** (Page *et al.* 1998), similar en ciertos aspectos al propuesto por Kleinberg. Al igual que Kleinberg, *PageRank* se basa en el uso de autoridades; sin embargo, no todos los enlaces son valorados del mismo modo sino en función de un valor numérico otorgado a cada documento, denominado también *PageRank*. Dicho valor indica el "prestigio" o la relevancia del documento y se propaga de unos documentos a otros: el *PageRank* de una página se divide por el número de enlaces de salida y se "transfiere" a los documentos enlazados. Así, documentos que reciben muchos enlaces aunque de poco valor serán muy relevantes y documentos que reciben pocos enlaces pero desde páginas con *PageRank* elevado serán igualmente importantes.

Además del valor *PageRank*, *Google* utiliza otros factores para ordenar los resultados de una consulta, por ejemplo, el texto de los enlaces que reciben los documentos, la posición de las palabras clave dentro del documento, etc. De este modo, los primeros documentos se corresponden, aproximadamente, con las autoridades que se obtendrían aplicando el algoritmo de Kleinberg pero sin eliminar la posibilidad de consultar otros documentos con menor *PageRank*. De este modo *Google* recupera muchos documentos para cada consulta (en ocasiones cientos o miles) pero ofreciendo siempre los documentos más relevantes[1] entre los primeros resultados.

Este sistema parece adecuado para la mayor parte de usuarios. Según Jansen *et al.* (1998) y Silverstein *et al.* (1998) los usuarios de motores de búsqueda resuelven una necesidad de información con menos de dos consultas (en un 67% de los casos de acuerdo con el primer estudio y en un 78% según el segundo), no suelen pasar de la primera página de resultados (en un 58% según los primeros y en un 85% según los segundos) y, de acuerdo con Jansen y Spink (2003), un 66% de los usuarios examinan entre los resultados menos de 5 documentos y un 30% un único documento. Jansen y Spink argumentan que esto se debe a tres razones: (1) las necesidades de información de la mayoría de internautas no son complejas, (2) los primeros documentos retornados son realmente "autoridades" para la consulta formulada y, (3) en promedio, alrededor del 50% de los documentos retornados son relevantes para una consulta específica desde la perspectiva del usuario (Jansen y Spink 2003, p. 68).

Esto suscita varias cuestiones. ¿Por qué retornar entonces miles de documentos para cada consulta? ¿No serían suficientes los *n* más relevantes a la manera de Kleinberg? ¿En cuántos casos no se alcanza una precisión del 50% en la primera página de resultados? ¿Existen usuarios que no encuentran lo que buscan entre los documentos con mayor puntuación pero que podrían hacerlo en alguno de los centenares apenas puntuados?

Tratando de dar respuestas a estas preguntas el autor de este trabajo realizó un sencillo experimento que se pasa a describir. En primer lugar, se recopilaron tres conjuntos de consultas. El primero estaba formado por las diez consultas más frecuentes realizadas en *Google* por usuarios españoles durante julio de 2004 (véase Tabla 1). El segundo por las 50

[1] Los más relevantes según el criterio de *Google* que no siempre coincidirá con el criterio del usuario.

consultas más frecuentes, según *Wordtracker*[1], realizadas el 17 de agosto de 2004 (véase Tabla 2). Y el tercero por 50 consultas extraídas en tiempo real de *Dogpile – Searchspy*[2], también el 17 de agosto (véase Tabla 3).

Posteriormente, cada consulta era enviada a *Google* y se obtenían los siguientes datos:

- Número de resultados totales.
- Aparición de un sitio web "oficial" como primer resultado (podría indicar que se trata de una consulta navegacional).
- En caso de no existir sitio web "oficial", número de temas distintos tratados por los documentos resultantes (podría dar pistas sobre la ambigüedad de la consulta).
- Porcentaje de documentos relevantes[3] para la consulta entre los diez primeros resultados. En caso de que apareciese un sitio web "oficial" como primer resultado se asignaba una relevancia del 100% al considerar que la consulta era "navegacional", esto es, tan sólo pretendía alcanzar un sitio web cuya existencia era conocida por el usuario.

Consulta	Sitio web oficial	Temas distintos	Resultados	Relevancia
shrek 2	✓	-	2.260.000	100%
paginas amarillas	✓	-	383.000	100%
renfe	✓	-	320.000	100%
sport	✓	-	171.000.000	100%
iberia	✓	-	952.000	100%
el corte ingles	✓	-	387.000	100%
harry potter	✓	-	6.280.000	100%
hola	✓	-	1.870.000	100%
chistes		**1**	**2.460.000**	**100%**
postales		1	4.350.000	100%

Tabla 1. Diez consultas más frecuentes realizadas por internautas españoles en Julio de 2004 (Fuente: *Google*).

A partir de los datos recogidos en Tabla 1 y Tabla 2 puede concluirse que gran parte de las consultas más frecuentes se corresponden con uno de los tres tipos siguientes[4]: (1) entretenimiento (por ejemplo, `shrek 2`, `outback jack` o `chistes`), (2) servicios disponibles

[1] `http://www.wordtracker.com`

[2] `http://www.dogpile.com/info.dogpl/searchspy`

[3] Cada una de las páginas obtenidas como resultado fue visitada por el autor para determinar si era o no relevante (según criterios humanos) para la consulta formulada.

[4] Las categorías propuestas por el autor de este trabajo no se corresponden con las de la taxonomía habitualmente utilizada (Broder 2002). De hecho, las consultas de "entretenimiento" y "servicios" pueden ser navegacionales (por ejemplo, `shrek 2` o `hotmail`), transaccionales (por ejemplo, `juegos` o `weather`) e incluso informativas (por ejemplo, `chistes` o `jokes`) mientras que las de "celebridades" podrían ser navegacionales (por ejemplo, `avril lavigne` o `pamela anderson`) o informativas (por ejemplo, `paris hilton` o `dan castellaneta`). Dos motivos impulsaron al autor a proponer una taxonomía diferente. Por un lado, sólo se querían clasificar las consultas más frecuentes, no cualquier consulta posible. Por otro, la taxonomía de Broder se basa en las intenciones del usuario que subyacen a la consulta. Por ejemplo, una consulta "navegacional" indica que el usuario conoce la existencia de un sitio web y emplea el buscador para alcanzarlo. Así, desde ese punto de vista la consulta `avril lavigne` es navegacional desde mediados de 2002; antes, al no existir sitio web oficial, la consulta era "informativa", se deseaba encontrar información sobre la artista en una o más páginas. Dada la naturaleza de las consultas más frecuentes parecía más adecuada una taxonomía centrada en el objeto de la consulta que en las intenciones del usuario que la realiza.

vía Web (por ejemplo, `el corte ingles`, `ebay` o `weather`) y (3) celebridades (por ejemplo, `paris hilton` o `ian thorpe`).

En el primer caso las consultas son simples; si se trata de una película, libro o programa de televisión los usuarios emplean el título como consulta (por ejemplo, `harry potter`), si no, se limitan a indicar qué quieren (por ejemplo, `chistes`). Para el segundo tipo o bien se utiliza como consulta el propio nombre del servicio (por ejemplo, `ebay`, `yahoo` o `hotmail`) o palabras clave bien conocidas (`postales`, `jokes`, `weather`, `jobs`, etc.) para describir servicios "genéricos" (el usuario no tiene preferencia por ninguno en particular). Por último, en el tercer caso la consulta se reduce al nombre y/o apellido de la celebridad.

Este tipo de consultas frecuentes parecen obtener resultados satisfactorios de manera regular por tres motivos: existe un sitio web oficial que cubre las necesidades de información, o bien uno o más sitios web que ofrecen servicios análogos (por ejemplo, postales, juegos o información meteorológica) o se trata de información disponible en periódicos digitales[1].

Así, el experimento realizado por el autor parece llegar a las mismas conclusiones que Jansen *et al.* (1998), Silverstein *et al.* (1998) y Silverstein y Spink (2003): las consultas no suelen tener más de dos términos, no es necesario pasar de la primera página de resultados y, en promedio, más del 50% de los primeros resultados son relevantes para la consulta formulada (de hecho, el porcentaje es ligeramente superior al 90% en el caso de las consultas más frecuentes).

De manera análoga se procedió a estudiar las 50 consultas obtenidas en tiempo real que, en teoría, constituyen una muestra razonable de consultas "típicas" pero no frecuentes. Se eliminaron las que obtenían como primer resultado un sitio web "oficial" y aquellas para las cuales fue imposible determinar la relevancia de los resultados[2]. De este modo quedaron 35 consultas que obtuvieron unos resultados con una precisión promedio del 55%.

Ese dato también es coherente con los resultados de Silverstein y Spink (2003). Sin embargo, estos investigadores no indican la dispersión que muestra la relevancia de los resultados en su experimento. Por su parte, el autor ha detectado en esta muestra que el 20% de las consultas obtiene una precisión media de tan sólo el 21% y el 23% de las consultas no obtienen ningún documento relevante entre los 10 primeros.

---

[1] Aparentemente, algunas de las consultas del tercer tipo buscan información sobre "escándalos" recientes en los que estaría envuelta la celebridad en cuestión. Este sería el caso de las consultas `kobe bryant` o `mike wallace`, el primero, jugador de baloncesto, por estar sometido a juicio por una agresión sexual y el segundo, presentador de televisión, por desorden público.

[2] Fue imposible determinar la relevancia de los resultados obtenidos para consultas como `e-find`, `people` o `itt-tech and homework` puesto que, a la vista de la consulta y los resultados, era muy difícil deducir qué buscaba realmente el usuario con esa consulta.

Consulta		Sitio web oficial	Temas distintos	Resultados	Relevancia
olympics	ⓘ	✓	-	8.250.000	100%
hurricane charley			**1**	**598.000**	**100%**
avril lavigne	🗣	✓	-	1.590.000	100%
google	🛠	✓	-	58.400.000	100%
yahoo	🛠	✓	-	123.000.000	100%
ebay	🛠	✓	-	68.400.000	100%
paris hilton	🗣		3	3.420.000	40%
outback jack	❑	✓	-	156.000	100%
mapquest	🛠	✓	-	1.920.000	100%
yahoo.com	🛠	✓	-	123.000.000	100%
james mcgreevey	🗣	✓	-	129.000	100%
kobe bryant	🗣		2	931.000	80%
dan castellaneta	🗣		1	26.500	100%
lindsay lohan	🗣		1	241.000	100%
mike wallace	🗣		4	1.700.000	50%
dawn staley	🗣		1	32.900	100%
britney spears	🗣	✓	-	4.930.000	100%
michael phelps	🗣		3	447.000	80%
nudist+image			**4**	**709.000**	**¿?**
weather	🛠		1	76.800.000	100%
maps	🛠		1	133.000.000	100%
jokes			**1**	**19.800.000**	**100%**
amber frey	🗣		1	77.700	100%
hotmail	🛠	✓	-	21.700.000	100%
thong	**SEX**		**6**	**7.080.000**	**40%**
games			**3**	**274.000.000**	**70%**
jobs	🛠		1	160.000.000	100%
search engines			**1**	**9.380.000**	**100%**
john heffron	🗣	✓	-	36.200	100%
carmen electra	🗣		1	891.000	100%
pamela anderson	🗣	✓	-	2.710.000	100%
camel+toes	**SEX**		**3**	**206.000**	**80%**
hilary duff	🗣	✓	-	811.000	100%
path client			**¿?**	**3.500.000**	**0%**
tattoos			**2**	**6.020.000**	**90%**
nicky hilton	🗣		1	220.000	100%
ashlee simpson	🗣	✓	-	233.000	100%
thongs	**SEX**		**3**	**1.820.000**	**80%**
dictionary			**1**	**38.600.000**	**100%**
home depot	🛠	✓	-	2.630.000	100%
hotmail.com	🛠	✓	-	21.700.000	100%
www.thehun.com	🛠	✓	-	518.000	100%
share jackson			**¿?**	**2.640.000**	**10%**
inuyasha			**1**	**1.270.000**	**100%**
anime			**1**	**42.800.000**	**100%**
ian thorpe	🗣	✓	-	225.000	100%
travel			**1**	**174.000.000**	**100%**
ask jeeves	🛠	✓	-	1.100.000	100%
jessica simpson	🗣	✓	-	1.620.000	100%
ebay.com	🛠	✓	-	68.400.000	100%

Tabla 2. Cincuenta consultas más frecuentes realizadas por internautas de todo el mundo durante el 17 de Agosto de 2004 (Fuente: *Wordtracker*).

Consulta	Sitio web oficial	Temas distintos	Resultados	Relevancia
rapid serial visual presentation		**1**	**34.700**	**100%**
sesshomaru		**1**	**47.900**	**100%**
ffdo		**3**	**2.810**	**60%**
cape plumbago photo		**1**	**450**	**70%**
rosettanet	✓	-	97.900	100%
nike air force ones		2	43.200	¿?
"recetas de ensaladas"		**1**	**6.210**	**70%**
rental car coupons		1	1.560.000	¿?
hemmoroid *(sic)* treatment		**1**	**686**	**100%**
jets pizza		1	42.800	¿?
dee zee running boards	✓	1	8.730	100%
www.adopt a pokemon		**¿?**	**32**	**0%**
how much paper is in one tree?		**1**	**2.200.000**	**70%**
wedding tents		1	181.000	¿?
itt-tech and homework		¿?	1.010	¿?
captree school+west islip	✓	1	14.100	100%
prpcmonitor		**1**	**958**	**100%**
daleville indiana + images		**4**	**3.500**	**20%**
arlington washington public library		**1**	**277.000**	**0%**
fundraising software comparison blackbaud oficial		**1**	**505**	**100%**
handwoven yoga mats		**1**	**899**	**100%**
flood film project		**8**	**283.000**	**30%**
car insurace *(sic)* quotes		**1**	**15.100**	**50%**
"salas surgical group california"		**0**	**0**	**0%**
ravrn simone		**1**	**2**	**0%**
where can i upload roms		**¿?**	**106.000**	**0%**
pcgs + indian cent cameo		**1**	**2.270**	**100%**
what do black baby snakes eat		**1**	**63.300**	**20%**
compile time object		**¿?**	**720.000**	**0%**
ymca swimming lessons- kalamazoo, mi		**1**	**201**	**10%**
combat support flight		2	618.000	¿?
elaine beno		**3**	**836**	**30%**
ancient olympics		**1**	**511.000**	**100%**
methow.com	✓	-	54.100	100%
glenn county, ca		**1**	**891.000**	**100%**
people		¿?	316.000.000	¿?
win ace *(sic)*	✓	2	1.080.000	100%
photgraphed *(sic)* by george smith		**¿?**	**95**	**0%**
"motorcycle tent trailer"		**1**	**2.910**	**0%**
wimbledon	✓	1	2.000.000	100%
required octane		**1**	**90.900**	**20%**
e-find		¿?	53.200	¿?
autos		**1**	**15.500.000**	**100%**
escorps		**2**	**129**	**80%**
nra	✓	7	1.280.000	100%
origin of christian traditions		**1**	**342.000**	**100%**
galaxie 1963 ford power steering valve		**1**	**650**	**20%**
broadband internet tucson		**2**	**36.200**	**80%**
armitage funeral home kearny		**1**	**68**	**100%**
instagate pro vpn		**1**	**1.680**	**100%**

Tabla 3. Cincuenta consultas capturadas en tiempo real el 17 de Agosto de 2004 (Fuente: *Dogpile – Searchspy*).

Puesto que los resultados obtenidos con este pequeño experimento han coincidido con los obtenidos en otros realizados a mayor escala parece razonable no extrapolar sin más los datos obtenidos pero sí argumentar que un porcentaje relativamente elevado de las consultas "típicas" no obtiene resultados relevantes en su primera formulación. Al examinar

una a una las consultas de la muestra parece que algunas podrían reformularse y otras presentan errores ortográficos. No obstante, dada la reticencia de los usuarios a realizar más de una consulta para un mismo problema no parece adecuado exigirles lo primero y, por otro lado, un sistema debería ser robusto frente a errores tipográficos y ortográficos[1] sin recurrir a artefactos complejos.

En resumen, por un motivo u otro, un porcentaje desconocido pero elevado[2] (tal vez entre el 15 y el 20%) de las consultas que se envían a un buscador moderno como *Google* no obtienen ningún resultado positivo entre los diez primeros. Podría aducirse que si los diez documentos más relevantes no satisfacen la consulta entonces ninguno de los restantes lo hará; sin embargo, este argumento es poco sólido.

Por un lado, es necesario recordar que la "relevancia" de los documentos es, en realidad, una medida de su "prestigio" que se obtiene de manera algorítmica basándose, fundamentalmente, en los enlaces que apuntan hacia cada página. Ya se mencionaron algunos problemas de esta técnica señalados por Bharat y Henzinger (1998) –véase página 12. Tales problemas hacen necesario aceptar con reservas el grado de "relevancia" o "irrelevancia" otorgado por un buscador a los distintos documentos de la Web.

Por otro lado, aun en el caso de que todos los enlaces se estableciesen de un modo ideal para los algoritmos de Kleinberg y *PageRank*, habría que seguir desconfiando por una razón muy simple: aunque una página web muy enlazada sea relevante lo contrario, una página poco enlazada es irrelevante, no es necesariamente cierto. Un argumento a favor de esto puede encontrarse en el trabajo de Estelle Broadman (1944) que demostró que el valor de una publicación para un profesional no es directamente proporcional al número de veces en que es citada en otras obras. Recordemos que, de hecho, el índice de impacto de una publicación requiere una normalización del número de citas totales recibidas (Garfield 1972)[3].

Así pues, *Google* y casi todos los buscadores modernos resuelven muy bien aquellas consultas para las que existen una o más páginas "autorizadas" y no obtienen tan buenos resultados cuando no existen tales autoridades. En este último caso, el usuario simplemente recibe una avalancha de información. Para tratar de aliviar esta situación *Google* dispone del

---

[1] Uno de los puntos de interés de la técnica descrita por el autor en esta disertación es, precisamente, su tolerancia al "ruido". No debe confundirse, sin embargo, esta capacidad con la corrección de errores que ofrecen muchos buscadores en la Web. Un sistema "tolerante" aceptaría una consulta como `guttemberg` (*sic*) y ofrecería, de manera transparente, resultados con ese término y otros como `gutemberg`, `guttenberg` e, incluso, `gutenberg`. El hecho de que un documento presente un error tipográfico (o una grafía menos popular) no lo invalida como fuente de información; por ejemplo, alguien que busque información sobre `mao zedong` estará probablemente satisfecho con documentos que mencionen a `mao tse tung`.

[2] Dijimos que no se podían extrapolar los datos puesto que la muestra es muy reducida; no obstante, es posible emitir una suposición razonada. Según Silverstein *et al.* (1998) el 86,4% de las consultas que recibe un buscador (en su caso *Altavista*) se repiten un máximo de tres veces y el 63,7% aparecen una única vez (en ambos casos durante un período de 43 días). Si tomamos estos datos como unas cotas razonables para definir el porcentaje de "consultas típicas" en oposición al de "consultas frecuentes" y damos por bueno el 23% de consultas que no obtienen resultados tendríamos que entre el 15% y el 20% de las consultas que recibe un buscador son resueltas sin dar ningún documento relevante entre los diez primeros.

[3] Podría resultar un problema interesante estudiar posibles modificaciones del algoritmo de Kleinberg o *PageRank* mediante la aplicación de algunas de las técnicas que se emplean para calcular índices de impacto; sin embargo, tampoco es ese el tema de esta disertación.

servicio *Google Answers*[1] ("*Google* responde" o "Respuestas *Google*") que permite a los usuarios hacer preguntas que serán respondidas por otros usuarios expertos tras un pago en metálico. Es decir, una solución "manual" al problema en cuestión.

Sirva esto como un tercer argumento en apoyo de la existencia tangible de una sobrecarga de información en la Web a la espera de una solución automatizada que, como se dijo antes, es el problema que se afronta en este trabajo.

6 Distintas propuestas para luchar contra la sobrecarga de información

La cantidad de información disponible en Internet es enorme (véase página 4) y existe un porcentaje no despreciable (tal vez entre un 15 y un 20%) de consultas que los buscadores no son capaces de resolver satisfactoriamente. Tales consultas obtienen como resultado cientos o miles de documentos sin existir ninguno adecuado entre los primeros aunque es razonable suponer que alguno de los restantes sí es relevante para las necesidades del usuario. El problema radica en encontrar en un conjunto de documentos muy grande unos pocos que sean de interés para el usuario.

A lo largo de los años noventa se realizaron toda una serie de investigaciones sobre este asunto no sólo en la Web sino también en otros servicios como correo electrónico o grupos de *USENET*. Estos trabajos emplearon, de forma independiente o combinada, tres técnicas básicas: agentes *software*, filtrado colaborativo y recomendación por contenidos.

Un **agente** es un elemento *software* capaz de interactuar con su entorno (incluidos otros agentes) para realizar una tarea en representación de un usuario o de otro agente. Los agentes implementan algún tipo de inteligencia artificial que les permite actuar de manera autónoma y determinar las acciones apropiadas para responder a los eventos del entorno.

El **filtrado colaborativo** (Goldberg *et al* 1992) proporciona a un usuario lo que otros individuos similares encontraron de utilidad antes que él. Un ejemplo típico es el servicio de *Amazon*[2] *"Customers who bought this book also bought…"* ("Los clientes que compraron este libro también compraron…")

Por su parte, la **recomendación por contenidos** tiene como objetivo proporcionar documentos similares a un documento de partida y precisa, por tanto, de algún tipo de análisis del texto de los documentos.

A continuación se describirán muy brevemente algunas de las iniciativas más interesantes señalando aspectos innovadores potencialmente aplicables al problema de la sobrecarga de información en la Web así como aspectos débiles de cara a una solución totalmente automática.

Paul E. Baclace (1991 y 1992) utiliza agentes para filtrar la información que recibe un usuario. Dichos agentes evalúan, de manera individual, el interés de los documentos en función del autor y algunas palabras clave. Una vez hecho esto se obtiene una puntuación

---

[1] El servicio *Google* Answers (`http://answers.google.com`) está definido en los términos siguientes: *"El motor de búsqueda de Google es una gran manera de encontrar información en línea. Pero a veces incluso los usuarios experimentados necesitan ayuda para encontrar la respuesta exacta a una pregunta. Google Answers es una forma de conseguir ayuda de expertos en la búsqueda en línea. Al proponer una pregunta usted especifica la cantidad que está dispuesto a pagar por la respuesta y la diligencia con que necesita esa información. Un experto buscará la respuesta y le enviará la información que está buscando, así como enlaces útiles a páginas web sobre el tema. Si usted está satisfecho con la respuesta pagará la cantidad previamente estipulada."*

[2] `http://www.amazon.com`

media para cada documento y aquellos de interés son enviados al usuario que debe evaluarlos. Dicha evaluación permite recompensar a los agentes que evaluaron correctamente el documento y penalizar a los que lo hicieron incorrectamente. Tras una serie de iteraciones el usuario dispone de una población de agentes adaptada a sus intereses. Esta propuesta, aunque interesante, es difícilmente aplicable al problema que nos ocupa por varias razones. En primer lugar, requiere una evaluación explícita por parte del usuario de los documentos calificados como relevantes por los agentes, algo que puede ser inabordable en muchos casos. En segundo lugar, al requerir una serie de iteraciones para obtener un conjunto apto de agentes la técnica es útil para filtrar información acerca de intereses relativamente estables en el tiempo pero no para resolver consultas específicas.

Masahiro Morita y Yoichi Shinoda (1994) describen un experimento que trata el problema de proporcionar artículos interesantes de *USENET* a un grupo de usuarios en función de sus preferencias. El sistema presentado obtiene las valoraciones de manera implícita (a partir de los tiempos de lectura, de las acciones realizadas en el entorno y de las acciones realizadas sobre el texto del artículo) demostrando así que es posible extraer información relevante para el usuario sin necesidad de exigirle ningún esfuerzo consciente. Por otro lado, Morita y Shinoda no utilizan palabras clave (en realidad ideogramas clave) para seleccionar los documentos sino bigramas[1] de palabras.

Pattie Maes (1994) describe una serie de agentes con cometidos similares a los de Baclace (1991 y 1992): filtrar correo y artículos *USENET*, además de recomendar libros o música. Al igual que este último, Maes pretende que el usuario evalúe la calidad de la información que se le ofrece de una manera que se podría calificar de "abiertamente intrusiva". Por ejemplo, el sistema de recomendación musical, *Ringo*, requiere que un usuario evalúe en el momento del registro una lista de 125 artistas para indicar sus preferencias (Shardanand y Maes 1995, p. 211) algo que no parece demasiado razonable.

Menczer, Belew y Willuhn (1995) y Menczer y Belew (1998) describen una técnica similar a la de Baclace (1991 y 1992) aunque con diferencias importantes. En primer lugar, el sistema se emplea para realizar consultas en la Web y no para filtrar información. En segundo, los ecosistemas de agentes se crean para cada consulta individual por lo que no existen ni evolucionan de forma indefinida. Los agentes disponen de una cierta energía que consumen al explorar la Web y pueden recuperar parte de la energía consumida presentando algún documento al usuario que debe valorarlo de manera explícita. Como ya se dijo con anterioridad obligar al usuario a evaluar los resultados no es adecuado, especialmente existiendo formas de obtener una evaluación implícita (Morita y Shinoda 1994). Por otro lado, la evaluación de los prototipos se hace en subgrafos de la Web muy limitados: 116 documentos en el caso de (Menczer, Belew y Willuhn 1995) y 11.000 documentos del sitio web de la *Encyclopaedia Britannica* en (Menczer y Bellew 1998). Además, los resultados de estos experimentos se comparan con las técnicas empleadas por los buscadores de la época que no implementan algoritmos como *PageRank* o similares algo común en los buscadores actuales. Así pues, se trata de una técnica interesante aunque es difícil determinar en qué medida mejoraría los resultados de un buscador moderno.

Henry Lieberman (1995) desarrolló *Letizia*, un agente que asiste al usuario mientras éste navega por la Web. *Letizia* analiza las acciones del usuario sobre los documentos

[1] Un bigrama es una subcadena que contiene dos elementos (palabras o caracteres) y que se obtiene desplazando, elemento a elemento, una "ventana" sobre el texto. La oración anterior, por ejemplo, contendría los siguientes bigramas de palabras: `<Un bigrama>`, `<bigrama es>`, `<es una>`, `<una subcadena>`, etc.

(activar un enlace, grabar o imprimir el documento, etc.) para establecer su interés, determina de forma aproximada el contenido de los documentos extrayendo una serie de palabras clave y, además, explora la Web en segundo plano en búsqueda de documentos similares a los que el usuario considera interesantes. Los documentos valorados como potencialmente interesantes se almacenan en una lista que evoluciona a medida que avanza la exploración del usuario; de tal forma que puede, en cualquier momento, solicitar al agente una recomendación que éste extrae de la lista anterior.

Son varios los aspectos a destacar en esta propuesta: no requiere valoración explícita del usuario, determina un perfil aproximado para el mismo y explora la Web en su representación. Sin embargo, también presenta algunos inconvenientes: el análisis del contenido de los documentos es muy simple y puede conducir a recomendar documentos irrelevantes que coinciden en algunas palabras clave. *Letizia* sólo explora documentos próximos a aquel en que se encuentra el usuario y, además, la experiencia pasada del usuario o de otros usuarios con intereses similares no es tenida en cuenta.

LIRA (Balabanovic, Shoham y Yun 1995) es un agente que permite recomendar diariamente a un usuario un pequeño conjunto de páginas web potencialmente interesantes. Según sus autores, a lo largo del experimento el sistema ofreció en un 50% de los casos mejores resultados que los ofrecidos por un experto humano. Sin embargo, el experimento se hizo con un máximo de 6 usuarios simultáneamente y durante apenas 3 semanas por lo que no pueden considerarse unos datos excesivamente concluyentes. Por otro lado, son varias las críticas que se pueden hacer a *LIRA*: requiere de los usuarios una valoración explícita de los documentos, sólo funciona adecuadamente si el usuario manifiesta un único interés bien definido y, además, emplea extracción de palabras clave como herramienta de análisis de contenidos. Esto último se manifiesta en una limitación reconocida por los propios investigadores: "*Las páginas retornadas por el sistema son a menudo muy similares entre sí, tal y como han señalado muchos de los usuarios.* (Balabanovic, Shoham y Yun 1995, p. 8)"

MUSAG (Goldman, Langer y Rosenschein 1996) es uno de los primeros intentos de abandonar la técnica de coincidencia de palabras clave para la búsqueda de información en la Web. El prototipo utiliza dos agentes, *MUSAG* y *SAg*. El primero tiene como finalidad generar diccionarios "conceptuales" que agrupan las palabras que emplea el usuario en sus consultas con palabras que aparecen en los documentos resultantes. El segundo, *SAg*, emplea estos diccionarios para expandir las consultas. Esta técnica es similar a una de las utilizadas por Salton (1968) en el sistema *SMART* (véase página 3). No obstante, los diccionarios son simples tablas de expresiones asociadas a una palabra y, además, el único criterio de relevancia es la presencia de palabras del diccionario en el documento sin tener en cuenta los posibles intereses o necesidades del usuario.

Fab (Balabanovic y Shoham, 1997) es un sistema de agentes que recomienda páginas web mediante un sistema híbrido que combina colaboración entre usuarios y análisis automático de contenidos. Esta primera versión de *Fab* requería una evaluación explícita de los documentos, posteriormente sería modificado en la línea de (Morita y Shinoda 1994) para obtener valoraciones implícitas:

> *En escenarios típicos, los usuarios proporcionan feedback explícito sólo a regañadientes [...] por tanto, no es razonable imponer una carga extra a usuarios que ya intentan reducir su sobrecarga de información. Por tanto, el primer objetivo es aprender a recomendar documentos apropiados utilizando solamente feedback implícito.* (Balabanovic 1998, p. 6)

Es necesario indicar que en las pruebas que Balabanovic realizó de su sistema se emplearon 1.600 artículos de prensa cubriendo un período de dos semanas, documentos, quizás, demasiado homogéneos en cuanto a su estructura y muy diferentes de la mayor parte de páginas web existentes.

GroupLens (Konstan *et al.* 1997) describe un sistema que demuestra que la utilización del tiempo de lectura de un documento como sistema de evaluación implícita permite obtener recomendaciones similares a las producidas empleando valoración explícita.

Siteseer (Rucker y Marcos 1997) es un proyecto sencillo pero que señala un par de puntos interesantes. El sistema tomaba los *bookmarks* (páginas favoritas) de un usuario y su estructuración como un indicativo de sus intereses y las relaciones semánticas que establecía entre los mismos. Para realizar recomendaciones, se comparaban los intereses de cada usuario con los del resto y se le aconsejaba visitar documentos "favoritos" de otros usuarios que no estuviesen en su lista[1].

AntWorld (Kantor *et al.* 2000) es un proyecto que trata de ayudar a los usuarios a encontrar la información que buscan aprovechando la experiencia y valoraciones de anteriores usuarios del sistema. Como la mayoría de las propuestas revisadas, los desarrolladores de *AntWorld* piensan que la valoración de los documentos debe ser activa por parte del usuario y se muestran escépticos sobre la obtención pasiva de dicha valoración.

En resumen, para poder ofrecer a un usuario unos pocos documentos seleccionados de un conjunto muy grande es inevitable que el propio usuario u otros usuarios con intereses similares los hayan "evaluado". No obstante, no es necesario que la evaluación de los documentos sea explícita (por ejemplo, otorgando una calificación) puesto que el comportamiento del usuario al actuar sobre el documento proporciona indicios sobre el grado de interés del mismo (Morita y Shinoda 1994), (Balabanovic 1998) o (Jansen y Spink 2003).

Por otro lado, también es imprescindible realizar un análisis de los contenidos a fin de determinar qué documentos y/o perfiles de usuarios (páginas favoritas, historial de visitas, consultas realizadas, combinaciones de lo anterior, etc.) son similares y en qué grado. Morita y Shinoda (1994) señalaron que es posible emplear técnicas sencillas (en su caso bigramas de ideogramas) para comparar documentos con mejores resultados que empleando únicamente palabras clave.

7 La Web Semántica

Paralelamente al desarrollo de técnicas como las de Kleinberg o *Google* para localizar documentos en la Web y al mismo tiempo en que se buscaban soluciones al problema de la sobrecarga de información en Internet, Tim Berners-Lee (1998) esbozaba el concepto de **Web Semántica** que, junto con James Hendler y Ora Lassila, refinaría posteriormente (Berners-Lee, Hendler y Lassila 2001).

Simplificando enormemente puede decirse que el objetivo básico de la Web Semántica es permitir que agentes *software* sean capaces de "consumir" documentos disponibles en la Web para inferir nuevo conocimiento. Para ello los documentos deberían construirse empleando lenguajes "semánticos" que permitirían no sólo anotar

[1] Una iniciativa similar a la que ahora desarrolla *del.icio.us* que se define como *"un gestor social de enlaces favoritos"* (`http://del.icio.us`).

metainformación sino también especificar las relaciones existentes entre los metadatos. El *quid* de la cuestión radica en la forma de construir las etiquetas semánticas de los nuevos lenguajes e indicar las relaciones entre las mismas (véase Fig. 1 y Fig. 2). Para realizar esta labor se ha optado por la utilización de **ontologías**.

```
<INSTANCE KEY="http://www.cs.umd.edu/users/hendler/">
  <USE-ONTOLOGY ID="cs-dept-ontology" VERSION="1.0" PREFIX="cs"
  URL="http://www.cs.umd.edu/projects/plus/SHOE/cs.html" />

  <CATEGORY NAME="cs.Professor" FOR="http://www.cs.umd.edu/users/hendler/"/>

  <RELATION NAME="cs.member">
    <ARG POS=1 VALUE="http://www.cs.umd.edu/projects/plus/">
    <ARG POS=2 VALUE="http://www.cs.umd.edu/users/hendler/">
  </RELATION>

  <RELATION NAME="cs.name">
    <ARG POS=2 VALUE="Dr. James Hendler">
  </RELATION>

  <RELATION NAME="cs.doctoralDegreeFrom">
    <ARG POS=1 VALUE="http://www.cs.umd.edu/users/hendler/">
    <ARG POS=2 VALUE="http://www.brown.edu">
  </RELATION>

  <RELATION NAME="cs.emailAddress">
    <ARG POS=2 VALUE="hendler@cs.umd.edu">
  </RELATION>

  <RELATION NAME="cs.head">
   <ARG POS=1 VALUE="http://www.cs.umd.edu/projects/plus/">
   <ARG POS=2 VALUE="http://www.cs.umd.edu/users/hendler/">
  </RELATION>
</INSTANCE>
```

Fig. 1 Código *SHOE* (véase página 25) **utilizado para etiquetar una página *HTML*.**

Este código parte de una ontología (véase Fig. 2) que describe departamentos universitarios de informática. Indica que el documento (la página *HTML*) hace referencia al profesor James Hendler que es doctor por la Universidad de Brown y director de una organización cuya información está disponible en `http://www.cs.umd.edu/projects/plus/`.

El uso del término "ontología" no está exento de polémica (Soergel 1999) o (Bates 2002) debida, en parte, al origen filosófico[1] del mismo. Sin embargo, la definición de ontología aplicable al campo de la Web Semántica tiene poco que ver con la filosofía:

> *Una ontología es la especificación de una conceptualización. Esto es, una descripción de los conceptos y relaciones que pueden existir para un agente o una comunidad de agentes* (Gruber 1993).

Según Berners-Lee, Hendler y Lassila (2001, p. 4) una ontología es:

> *Un documento o fichero que define formalmente las relaciones entre términos. Una ontología típica para la Web consta de una taxonomía y de un conjunto de reglas de inferencia.*

Con anterioridad o simultáneamente a la propuesta de Berners-Lee (1998) para la Web Semántica se estaban realizando una serie de trabajos que tenían como objetivo desarrollar lenguajes que permitiesen definir tales ontologías y utilizarlas para etiquetar documentos en la Web, lo que podríamos denominar "pre-Web-Semántica". Cabe destacar los proyectos *SHOE* (Luke, Spector y Rager 1996), *WebKB* (Craven *et al.* 1998) y *Ontobroker/On2broker* (Fensel *et al.* 1998) y (Fensel *et al.* 1999), respectivamente.

[1] Según el Diccionario de la Lengua Española (*RAE* 2001) la ontología es la "parte de la metafísica que trata del ser en general y de sus propiedades trascendentales."

```
<!-- The ontology declarations start here. We begin by creating the ontology -->

<ONTOLOGY ID="cs-dept-ontology" VERSION="1.0" DESCRIPTION="An example ontology for
computer science academic department">

<!-- Now we declare that the ontology will be borrowing elements from
     the base-ontology.  -->

<USE-ONTOLOGY ID="base-ontology" VERSION="1.0" PREFIX="base"
URL="http://www.cs.umd.edu/projects/plus/SHOE/onts/base1.0.html">

<!-- Here we declare the categories in this ontology -->

<DEF-CATEGORY NAME="Person" ISA="base.SHOEEntity" SHORT="person">
<DEF-CATEGORY NAME="Worker" ISA="Person" SHORT="worker">
<DEF-CATEGORY NAME="Faculty" ISA="Worker" SHORT="faculty member">
<DEF-CATEGORY NAME="Professor" ISA="Faculty" SHORT="professor">
…
<DEF-CATEGORY NAME="Organization" ISA="base.SHOEEntity" SHORT="organization">
…
<DEF-CATEGORY NAME="ResearchGroup" ISA="Organization" SHORT="research group">
…

<!-- Here we declare the relations in the ontology -->

<DEF-RELATION NAME="emailAddress" SHORT="can be reached at">
  <DEF-ARG POS=1 TYPE="Person">
  <DEF-ARG POS=2 TYPE=".STRING" SHORT="email address">
</DEF-RELATION>

<DEF-RELATION NAME="head" SHORT="is headed by">
  <DEF-ARG POS=1 TYPE="Organization">
  <DEF-ARG POS=2 TYPE="Person">
</DEF-RELATION>
…
<DEF-RELATION NAME="doctoralDegreeFrom" SHORT="has a doctoral degree from">
  <DEF-ARG POS=1 TYPE="Person">
  <DEF-ARG POS=2 TYPE="University">
</DEF-RELATION>
…
<DEF-RELATION NAME="member" SHORT="has as a member">
  <DEF-ARG POS=1 TYPE="Organization">
  <DEF-ARG POS=2 TYPE="Person" SHORT="member">
</DEF-RELATION>

<!-- Here we declare some example inferences which might be useful to agents.  -->

<DEF-INFERENCE DESCRIPTION="Transitivity of Suborganizations.  If subOrganization(x,y)
and subOrganization(y,z) then subOrganization(x,z)">
  <INF-IF>
    <RELATION NAME="subOrganization">
      <ARG POS=FROM VALUE="x" VAR>
      <ARG POS=TO VALUE="y" VAR>
    </RELATION>
    <RELATION NAME="subOrganization">
      <ARG POS=FROM VALUE="y" VAR>
      <ARG POS=TO VALUE="z" VAR>
    </RELATION>
  </INF-IF>
  <INF-THEN>
    <RELATION NAME="subOrganization">
      <ARG POS=FROM VALUE="x" VAR>
      <ARG POS=TO VALUE="z" VAR>
    </RELATION>
  </INF-THEN>
</DEF-INFERENCE>
…

<!-- The end of the ontology marked here -->

</ONTOLOGY>
```

Fig. 2 Ontología *SHOE* (véase página 25) **que describe un departamento de informática.**

Obsérvese cómo la ontología describe categorías (p.ej. persona, trabajador, profesor, organización, grupo de investigación, etc.), relaciones (p.ej. dirección de correo, doctorado por, miembro de, etc.) y ofrece ejemplos de inferencia de conocimiento.

SHOE (Luke, Spector y Rager 1996) es una de las primeras iniciativas destinadas a proporcionar un lenguaje de marcado semántico. Se trata de una extensión del lenguaje *HTML* que permite desarrollar ontologías (véase Fig. 2) y utilizar las clases y relaciones definidas en una o más de esas ontologías para marcar zonas específicas de un documento *HTML* (véase Fig. 1). Luke *et al.* describen asimismo una herramienta, *Exposé*, que explora la Web en busca de páginas anotadas con *SHOE* y almacena los asertos que encuentra en una base de conocimiento que puede utilizarse posteriormente para realizar consultas.

WebKB (Craven *et al.* 1998) tenía como objetivo construir, de forma automática, una base de conocimiento que reflejase el contenido de la Web de una forma inteligible para una máquina. Para lograr esto el sistema debía recibir una ontología que describiese las clases y relaciones, así como un conjunto de documentos, etiquetados sobre la base de dicha ontología, que servirían como conjunto de entrenamiento. Así, tras un período de entrenamiento adecuado, el sistema sería capaz de procesar documentos *HTML* y producir documentos marcados semánticamente de acuerdo a la ontología de partida.

Ontobroker (Fensel *et al.* 1998) fue una iniciativa muy similar a *SHOE* puesto que proponía una serie de herramientas para definir ontologías, etiquetar documentos basándose en dichas ontologías y realizar consultas e inferencia sobre una base de conocimiento. Posteriormente evolucionaría hacia *On2broker* (Fensel *et al.* 1999) cuya principal novedad fue la utilización de tecnologías como *XML*[1] o *RDF*[2].

XML y *RDF* constituyen las bases sobre las que comenzar a construir la Web Semántica (Berners-Lee, Hendler y Lassila 2001, p. 3) puesto que el primero posibilita la construcción de nuevos lenguajes de etiquetas, por ejemplo *RDF* que, a su vez, permite expresar asertos. Sin embargo, son necesarias toda una serie de capas encima de *RDF* para desarrollar finalmente la Web Semántica.

Por ejemplo, aunque *RDF* permite dar valores a las distintas propiedades de diferentes recursos no dispone de mecanismos para describir esas propiedades ni para describir las relaciones entre las propiedades y otros recursos. Para ello es necesario un lenguaje que permita definir vocabularios *RDF*. Dicho lenguaje, construido mediante *RDF*, es *RDF Schema* o *RDF(S)* (Brickley y Guha 2004). Este lenguaje define clases y propiedades que permiten, a su vez, describir nuevas clases, propiedades y recursos.

Sin embargo, tampoco *RDF* ni *RDF Schema* son capaces por sí solos de modelar ontologías, razón por lo que comienzan a desarrollarse lenguajes para este fin análogos a los definidos durante la fase pre-Web-Semántica con la diferencia de que los nuevos lenguajes se construyen sobre el estándar *RDF(S)*. Ejemplos de estas extensiones ontológicas para

---

[1] *XML – eXtensible Markup Language* (Lenguaje de Etiquetado Extensible) es una recomendación del Consorcio W3 que permite crear lenguajes de etiquetado de propósito específico (vocabularios *XML*) `<http://www.w3.org/XML/>`.

[2] *RDF – Resource Description Framework* (Marco para la Descripción de Recursos) es la especificación de un modelo de metadatos que realiza descripciones de recursos mediante sentencias que combinan un objeto, una propiedad y un valor para dicha propiedad, todo ello serializado mediante *XML* `<http://www.w3.org/RDF/>`. A continuación se muestra la sentencia "*Daniel Gayo Avello es el autor de* `http://www.di.uniovi.es/~dani/`":

```
<rdf:Description about='http://www.di.uniovi.es/~dani/'>
<Author>Daniel Gayo Avello</Author>
</rdf:Description>
```

RDF Schema son las desarrolladas por Staab *et al.* (2000) y Horrocks *et al.* (2000) que definen *OIL*[1] o por McGuinness *et al* (2000) con *DAML-ONT*[2].

Posteriormente *DAML-ONT* y *OIL* convergieron en el lenguaje *DAML+OIL* (van Harmelen, Patel-Schneider y Horrocks 2001) que terminaría evolucionando hacia *OWL*[3] (Bechhofer *et al.* 2004) una recomendación del Consorcio W3 y, por tanto, el estándar para la construcción de ontologías.

8 Consultas en la Web Semántica

En estos momentos existen toda una serie de tecnologías estandarizadas por el Consorcio W3 que permiten construir parte de la "pila" de la Web Semántica (véase Fig. 3). Así, se utiliza *XML* para desarrollar vocabularios como *RDF*, el cual permite expresar asertos acerca de recursos disponibles en la Web. Éste, a su vez, es la base para construir *RDF(S)* que posibilita la creación de nuevos vocabularios *RDF* y, por tanto, la creación de un lenguaje como *OWL* para definir ontologías. Ontologías que, a su vez, permiten etiquetar los documentos de la Web Semántica.

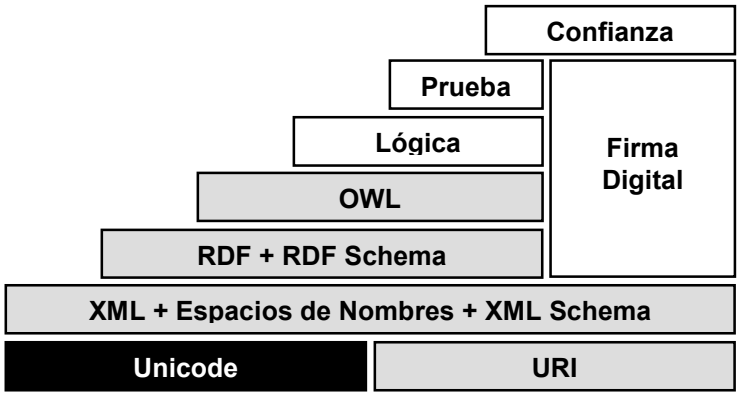


Fig. 3 Pila de la Web Semántica.

Se muestran sombreados aquellas capas de la Web Semántica para las que ya existe un estándar. En gris claro se representan los estándares propuestos por el Consorcio W3.

Sin embargo, aún quedan por desarrollar una serie de capas de esa "pila" y, quizás, una de las más urgentes sea la capa de consulta[4]. No obstante, ya se han investigado toda una serie de lenguajes entre los que cabe destacar *Metalog*, *SquishQL*/*RDQL* o *RQL*/*SeRQL*.

Metalog (Marchiori y Saarela 1998, 1999a y 1999b) fue el primer sistema en añadir una capa de lógica y consulta sobre *RDF* permitiendo inferir conocimiento nuevo a partir de los metadatos disponibles. Para facilitar su uso emplea un lenguaje "pseudo-natural" (*LPN*), es decir, una versión simplificada y controlada del inglés que permite expresar hechos, reglas y consultas que se aplicarán sobre metainformación *RDF*. Estas acciones pueden expresarse no sólo en *LPN* sino también en *RDF* o mediante algún otro lenguaje de programación lógica.

---

[1] *Ontology Inference Layer*, Capa de Inferencia Ontológica.

[2] Lenguaje ontológico del programa *DARPA Agent Markup Language* (Lenguaje de Etiquetado para Agentes DARPA).

[3] *Web Ontology Language* (Lenguaje Ontológico para la Web).

[4] Más bien inferencia (Guha *et al.* 1998).

SquishQL (Brickley y Miller 2000) (Miller, Seaborne y Reggiori 2002) es un lenguaje similar a *SQL* para realizar consultas sobre *RDF*. *RDQL* (Seaborne 2004) puede considerarse una evolución de *SquishQL* y ha sido propuesto por *Hewlett-Packard* al Consorcio W3 como un posible lenguaje de consulta para *RDF*.

RQL (Karvounarakis et al. 2001) es un lenguaje funcional, integrado dentro de *Sesame*[1], que permite consultar datos *RDF* y *RDF(S)*. *Sesame* tiene como objetivo ofrecer una plataforma estable para almacenar, consultar, manipular y administrar ontologías y metadatos expresados no sólo en *RDF* o *RDF(S)* sino también en *OWL*. En la actualidad, el lenguaje *SeRQL* (*Aduna B.V.* y *Sirma AI Ltd.* 2002-2004) ha sustituido a *RQL*.

Estos lenguajes probablemente tendrán una gran influencia en un futuro estándar W3C para un lenguaje de consulta sobre *RDF*, actualmente[2] en estudio por parte del *RDF Data Access Working Group* (Grupo de Trabajo para Acceso a Datos RDF). Después de todo Janne Saarela (*Metalog*), Alberto Reggiori (*SquishQL*), Andy Seaborne (*SquishQL* y *RDQL*) o James Hendler (*SHOE* y *DAML-ONT*) son algunos de los miembros de este grupo. Hasta el momento se han especificado casos de uso, requisitos y objetivos para el lenguaje de consulta (Clark 2004) y se ha empezado a esbozar el lenguaje *BRQL*[3] (Prud'hommeaux y Seaborne 2004) que permitirá seleccionar información, extraer subgrafos *RDF* y construir nuevos grafos *RDF* a partir del resultado de una consulta. Por el momento no se contempla utilizar el lenguaje de consulta sobre *RDF(S)* ni *OWL*.

```
SELECT ?title
PREFIX dc: <http://purl.org/dc/elements/1.1/>
WHERE { <http://example.org/book/book1> dc:title ?title . }
```

Fig. 4 Consulta BRQL para determinar el título de un libro.

Si *BRQL*, o un lenguaje similar, se convierte finalmente en un primer estándar para consultas en la Web Semántica (algo muy probable) se abrirán toda una serie de posibilidades en campos como agregación de contenidos, transporte, sistemas de producción, turismo, gestión de información personal, pruebas de *software*, comercio electrónico, etc. Los casos de uso planteados (Clark 2004) auguran un lenguaje expresivo y potente, aunque para resolver consultas fundamentalmente "metasemánticas", por ejemplo[4]:

- Encontrar la dirección de correo de Jonhny Lee Outlaw.
- Encontrar en la web de un proveedor información sobre el repuesto de una pieza así como la lista de piezas que deben ser sustituidas junto con la defectuosa.
- Recibir puntualmente información sobre libros, películas y música que cumplan unos criterios de título, precio y autor.
- Grabar todos los programas de televisión sobre el jugador de béisbol Ichiro.

La utilidad de una tecnología que permita resolver necesidades como las anteriores está fuera de toda duda; sin embargo, el problema que se afronta en este trabajo no es ese sino la forma de resolver en la Web de manera adecuada y automática consultas

[1] `http://www.openrdf.org`

[2] Agosto de 2004.

[3] *Bristol RDF Query Language.*

[4] Se han indicado las "necesidades de información" que se podrían resolver con un lenguaje como *BRQL* no las consultas expresadas en dicho lenguaje.

informativas[1] mucho más abiertas y ambiguas, formuladas en cualquier lenguaje natural[2] (tal vez con errores tipográficos, ortográficos o gramaticales) y susceptibles de sobrecargar de información al usuario. En definitiva, consultas como las siguientes[3]:

- *history and cultural Bengal* (historia y Bengal cultural).
- *acute predictors of aspiration pneumonia: how important is dysphagia?* (predictores adecuados para la neumonía por aspiración: ¿cuán importante es la disfagia?)
- *degenerative disk disease* (enfermedad degenerativa de disco). En el contexto médico es más común la forma *"disc"* que *"disk"*.
- *muscel* (sic) *aches during pregnancy* (dolores *muscualres* durante el embarazo). Consulta con error tipográfico.

9 La Web Cooperativa

Más de cuatro décadas transcurrieron entre la descripción de Vannevar Bush del dispositivo "memex" (Bush 1945) y el desarrollo de la Web, el sistema que tal vez se haya aproximado más a sus ideas. Durante ese tiempo se fueron solventando las "dificultades técnicas de todo tipo" que Bush auguraba y se ha llegado a la situación actual en la que centenares de miles de millones de documentos[4] están, en principio, a un *clic* de distancia de millones de usuarios.

No obstante, la realidad es muy distinta. A no ser que los usuarios conozcan la dirección de los documentos estos son inalcanzables puesto que la Web no dispone, por sí misma, de ningún mecanismo de recuperación de documentos. Por esa razón se han desarrollado sistemas de búsqueda capaces de proporcionar a los usuarios direcciones de páginas web en respuesta a sus consultas. Sin embargo, debido al tamaño de la Web las respuestas son, en general, demasiado numerosas y surge un problema de "sobrecarga de información".

A lo largo de los apartados anteriores se ha expuesto la existencia de dicho problema no sólo en la Web sino también en otros servicios de Internet. Se han estudiado con cierto detalle técnicas interesantes para solucionarlo como la recuperación y filtrado de información, la evaluación explícita o implícita de la relevancia de un documento por parte de los usuarios, las técnicas empleadas para explorar la Web a fin de localizar documentos desconocidos, los métodos para determinar el "prestigio" de un sitio web de modo análogo a como se calcula el "impacto" de una publicación científica así como la futura evolución de la Web hacia la Web Semántica.

---

[1] *"El propósito de las consultas informativas es encontrar información que se supone está disponible en la Web de forma estática. No se prevé más interacción que la lectura. Por forma estática se entiende que el documento no se crea en respuesta a la consulta."* (Broder 2002, p. 5)

[2] Aunque las técnicas que se describirán más adelante podrían ser aplicadas, en teoría, a idiomas ideográficos con resultados análogos a los obtenidos con idiomas alfabéticos, el autor se ha centrado en los segundos.

[3] Fuente: *Dogpile – Searchspy* <http://www.dogpile.com/info.dogpl/searchspy>

[4] Si se acepta que la Web Oculta tiene un tamaño 50 veces superior al de la Web superficial (Aguillo 2002), (Bergman 2001) y se toma el número de páginas indexadas por *Google* como cota inferior del tamaño de la segunda.

De este modo se ha podido delimitar mejor el problema que se afronta en este trabajo:

> *"La sobrecarga de información que experimentan los usuarios al tratar de resolver en la Web consultas informativas formuladas en lenguaje natural de manera tal vez ambigua y, en ocasiones, con errores tipográficos, ortográficos o gramaticales."*

La **Web Cooperativa**[1] (Gayo Avello y Álvarez Gutiérrez 2002) es una propuesta del autor para solucionar ese problema que se sustenta en los siguientes puntos:

- La utilización de conceptos, generados automáticamente, como una alternativa intermedia entre las ontologías y las palabras clave.
- La clasificación de documentos en una taxonomía a partir de tales conceptos.
- La cooperación entre usuarios, en realidad, entre agentes que actúan en representación de los usuarios y que no requieren su participación explícita.

9.1 Conceptos frente a palabras clave

La recuperación de información mediante palabras clave utilizada por los actuales motores de búsqueda plantea dos graves problemas: una tasa de recuperación excesiva y una precisión relativamente baja. La utilización de ontologías puede mejorar la precisión en algunos casos. Sin embargo, desarrollar ontologías que den soporte a cualquier consulta concebible en la Web supondría un esfuerzo inabordable[2].

Existe, sin embargo, una posibilidad intermedia: la utilización de conceptos. Un concepto sería una entidad más abstracta y, por tanto, con mayor carga semántica que una palabra clave. No obstante, no requeriría "artefactos" complejos como lenguajes ontológicos o sistemas de inferencia. Un concepto podría ser considerado como un grupo de palabras con un significado similar, o relacionado, dentro de un ámbito determinado ignorando tiempo, género y número[3]. Por ejemplo, en un área del conocimiento podría existir el concepto (`ordenador`, `máquina`, `servidor`) mientras que en otro existiría (`actor`, `actriz`, `artista`, `celebridad`, `estrella`).

Los conceptos, así entendidos, serán útiles si permiten proporcionar semántica de forma análoga a las ontologías y, simultáneamente, son generados y procesados automáticamente como las palabras clave. El autor tiene puestas grandes esperanzas en las

---

[1] El resto de este apartado y el siguiente se han elaborado a partir de los capítulos tercero y cuarto del trabajo de investigación defendido por el autor en 2002. El primero de ellos se corresponde con un artículo presentado en *COMPSAC* (Gayo Avello y Álvarez Gutiérrez 2002) y en el segundo se aclaran algunos puntos en respuesta a comentarios de los revisores.

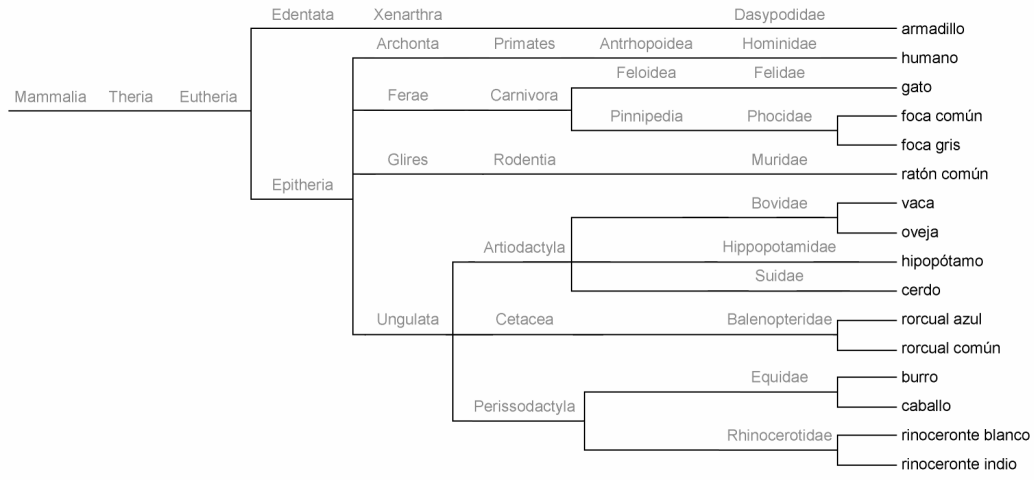
[2] El autor no es el único en sugerir la necesidad de un complemento para la Web Semántica que opere sobre la Web actual. Flake, Pennock y Fain (2003) afirman: *"Muchos han defendido la Web Semántica como un medio para mejorar la recuperación de información en la Web [argumentando que], en su forma actual, no resulta adecuada para el procesamiento automático puesto que la información no está estructurada. […] En la Web Semántica los autores utilizarán un lenguaje para anotar con etiquetas semánticas los datos. […] Resulta sencillo prever el etiquetado [semántico] implícito de catálogos de productos pero podría ser desalentador anotar semánticamente largos pasajes de texto –por ejemplo, artículos de revista. […] Un escenario complementario prevé algoritmos suficientemente inteligentes como para inferir semántica de la Web actual, no estructurada pero auto-organizada, sin ayuda de etiquetas semánticas. […] Los usuarios se beneficiarán más si el trabajo para la creación de la Web Semántica se realiza en paralelo al desarrollo de herramientas para el análisis de datos en la Web auto-organizada."*

[3] Estos conceptos serían similares a los *synsets* (conjuntos de sinónimos) empleados por *WordNet* <`http://wordnet.princeton.edu`>. Los *synsets* se definen como conjuntos de palabras intercambiables en algún contexto.

técnicas de Semántica Latente[1] (Foltz 1990) o de indexación de conceptos (Karypis y Han 2000). En la siguiente sección se examinará la forma en que es posible obtener semántica a partir de conceptos sin emplear ningún soporte ontológico.

9.2 Taxonomías de documentos

Para dotar de significado a un documento, la Web Semántica precisa una ontología que defina una serie de términos y relaciones entre los mismos. Dichos términos son utilizados para etiquetar diferentes partes del documento proporcionando así un "marcado semántico". La Web Cooperativa, por su parte, pretende utilizar el texto completo del documento, sin ningún tipo de etiquetado, como fuente de semántica. ¿Es esto posible sin "comprender" el significado del texto? A lo largo de esta sección se presentará una forma de procesar lenguaje natural para obtener, de manera totalmente automática, una clasificación conceptual de documentos.



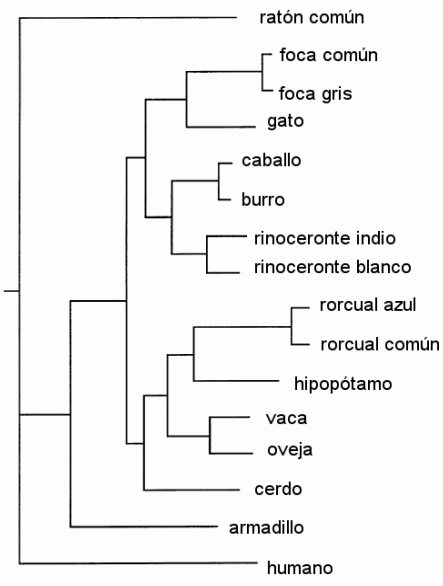


Fig. 5 Clasificación tradicional y clasificación automática basándose en el ADN mitocondrial (Ursing y Arnason 1998).

El método automático ha agrupado correctamente especies muy similares (foca común y gris, rorcual común y azul, vaca y oveja) y ha establecido categorías de un nivel superior (las dos especies de foca junto con el gato, o el caballo y el burro junto a las dos especies de rinoceronte). El método automático ha situado una especie, hipopótamo, en una categoría más próxima a las dos especies de rorcual de lo que aparece en la clasificación tradicional (fenotípica). Por otro lado, especies muy distintas como el ratón, el armadillo o los humanos no han sido agrupados en ninguna categoría próxima al resto de especies, lo cual es correcto. El método genético, sin embargo, no ha podido determinar que el armadillo pertenece a una categoría propia. Este hecho no es importante de cara a la utilización de estos algoritmos como sistema de clasificación semántica de documentos puesto que lo fundamental es establecer relaciones entre documentos similares y separar los documentos diferentes más que determinar el grado exacto de estas diferencias.

[1] Foltz y Dumais (1992) describen una experiencia en la que se combinan dos técnicas diferentes para describir los intereses de un grupo de usuarios (palabras clave y valoración de documentos) y dos técnicas de recuperación de información (búsqueda por palabras clave y semántica latente); la combinación que mejores predicciones produjo fue la semántica latente combinada con valoración de documentos.

Un documento puede considerarse como un individuo de una población. Entre los seres vivos un individuo está definido por su genoma, el cual se compone de cromosomas que se dividen en genes construidos a partir de bases genéticas. De forma similar, los documentos están compuestos por pasajes que se dividen en sentencias construidas mediante conceptos. Siguiendo esta analogía se puede conjeturar que dos documentos estarán semánticamente relacionados si sus respectivos "genomas" son similares y que grandes diferencias entre dichos "genomas" implicarán una relación semántica baja.

El autor considera que esta analogía puede ser puesta en práctica y que es posible adaptar algoritmos empleados en biología computacional al campo de la clasificación de documentos. Simplificando mucho, estos algoritmos se limitan a trabajar con largas cadenas de caracteres que representan fragmentos del genoma de individuos de la misma o de distintas especies. Individuos o especies similares muestran similitudes en sus códigos genéticos de tal forma que es posible mostrar la relación existente entre individuos y especies en taxonomías o dendrogramas[1] sin la necesidad de conocer, o lo que es lo mismo, comprender, la función de cada gen.

Estos dendrogramas permiten, en cierto modo, agrupar a las distintas especies en "categorías"; dichas "categorías" aportan información muy útil para comprender la evolución de las especies y, en muchas ocasiones, confirman (y en otras refutan) el sistema de clasificación de las especies clásico, basado en el sistema linneano.

Este sistema establece los grupos taxonómicos sobre la base de características observables en los seres vivos, es decir, su fenotipo. Los dendrogramas, sin embargo, establecen los distintos grupos basándose en el genotipo de las especies. El fenotipo depende del genotipo pero también está influenciado por el ambiente y por la interacción entre éste y el genotipo. Por esta razón, categorías obtenidas de forma automática a partir de la bioquímica de las especies pueden parecerse extraordinariamente a aquellas otras establecidas mediante un criterio de clasificación consciente e inteligente[2].

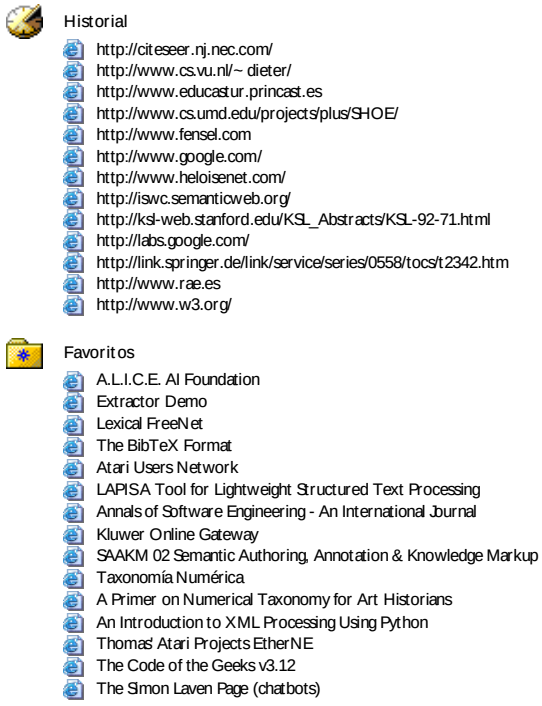


Fig. 6 Historial y lista de favoritos de un usuario.

[1] Un dendrograma es una representación gráfica de un proceso de agrupamiento que muestra las relaciones entre una serie de grupos. Puede verse un dendrograma como un árbol jerárquico, donde los grupos de la misma rama están más relacionados entre sí que con grupos de otras ramas (véase Fig. 5).

[2] Chakrabarti *et al.* (1998b) también plantearon la posibilidad de construir taxonomías de páginas relevantes de forma automática, los resultados que obtuvieron con su sistema *CLEVER* mostraban que las técnicas que empleaban proporcionaban mejores resultados que un directorio generado de forma semi-automática como *Yahoo!*. Sin embargo, el autor cree que es posible generar taxonomías para cualquier documento (no sólo los más relevantes) además de poder emplearse mejores indicadores de la relevancia que los empleados en *CLEVER*.

De forma análoga, los documentos podrían ser clasificados automáticamente en dendrogramas en función de las similitudes encontradas en sus respectivos "genomas conceptuales". La importancia de semejante sistema de clasificación radica en el hecho de que proporcionaría información semántica (similitudes a un nivel conceptual entre distintos documentos o entre documentos y consultas de usuario) sin utilizar ningún tipo de información semántica durante el proceso de clasificación. De hecho, debería ser capaz de agrupar documentos en categorías análogas a las que establecería un ser humano independientemente de la naturaleza del documento y del idioma en que el documento estuviera escrito.

9.3 Colaboración entre usuarios

Ya se ha dicho con anterioridad que la Web actual no permite sacar el máximo provecho al conocimiento experimental que obtienen los usuarios al explorarla. También se han estudiado algunas iniciativas de filtrado y recomendación de información que permitían la participación de los usuarios pero obligaban a estos a valorar documentos o proporcionar información de forma explícita. La Web Cooperativa pretende utilizar estas experiencias para extraer semántica de las mismas de forma no intrusiva y transparente para el usuario. Para ello cada usuario de la Web Cooperativa dispondría de un agente con dos objetivos: aprender de su "maestro" y recuperar información para él.

Fig. 7 Perfil de usuario extraído de los documentos anteriores (véase Fig. 6).

A partir de los documentos presentes en el historial de navegación y la lista de favoritos de un usuario será posible determinar sus principales temas de interés. Estos temas de interés configurarán un perfil que, con fines únicamente ilustrativos, se representa aquí como una "bolsa" de conceptos asociada a documentos representativos. Los distintos temas supondrán un porcentaje determinado del perfil del usuario y cada tema, a su vez, podrá matizar los conceptos que lo constituyen (representado aquí mediante una escala de gris donde los tonos oscuros indican conceptos importantes para el usuario y los claros señalan los menos relevantes).

9.3.1 Aprendizaje de los intereses del maestro

Para alcanzar este objetivo el agente debe desarrollar un perfil que describa de forma precisa los intereses del usuario. Esta descripción se haría mediante los conceptos anteriormente descritos y se construiría a partir de los documentos que el usuario almacena en su equipo, visita con frecuencia, añade a su lista de favoritos, etc. Todo ello sin intervención explícita del usuario.

Una vez un usuario es vinculado a un perfil es posible utilizar esta información para dar una semántica a los documentos de la Web que no es implícita a los mismos sino que depende de los usuarios. Ni la Web actual ni la Web Semántica tienen en cuenta la "utilidad" de los documentos. Los documentos son buscados y procesados por la utilidad que los usuarios esperan obtener de ellos. La utilidad de un documento no reside en sus contenidos sino que es un "juicio de valor" emitido por un usuario particular para un documento específico.

La Web Cooperativa, al tener asociado cada usuario a un perfil, puede asignar a cada par `(perfil, documento)` un nivel de utilidad. El agente asignado a cada usuario sería el responsable de determinar dicho nivel de utilidad. Este proceso de evaluación, para ser verdaderamente práctico, debería determinarse de una forma implícita (únicamente "observando" el comportamiento del usuario, sin necesidad de interrogarle). Por otro lado, el nivel de utilidad no sería asignado al documento como un todo sino a pasajes individuales dentro de un mismo documento[1].

Ya se ha visto que la mayor parte de iniciativas relacionadas con la valoración de recursos por parte de los usuarios requieren una participación voluntaria con los problemas que esto conlleva. Sin embargo, también se han presentado algunas experiencias interesantes en el campo de la valoración implícita que han mostrado que es factible. La segunda opción es preferible de cara a una implementación práctica.

9.3.2 Recuperación de información para el maestro

Un agente de la Web Cooperativa tendría dos formas de obtener información para su maestro:

- Buscar información para satisfacer una consulta.
- Explorar en representación del usuario para recomendarle documentos desconocidos.

Para poder llevar a cabo ambas tareas se pretende emplear dos técnicas bien conocidas: Filtrado Colaborativo y Recomendación por Contenidos. En la Web Cooperativa, si el agente empleara filtrado colaborativo recomendaría al usuario documentos a los que usuarios de su mismo perfil han otorgado un elevado nivel de utilidad.

Por otro lado, si emplease recomendación por contenidos proporcionaría documentos relacionados conceptualmente con el perfil del usuario, con una consulta o con un documento de partida, independientemente del nivel de utilidad que pudieran tener asociado.

---

[1] J. Allan realizó un estudio que *"apoya claramente la hipótesis de que los documentos largos contienen información que diluye el feedback [la valoración del usuario]. Recortar estos documentos seleccionando un pasaje adecuado tiene un acentuado impacto en la eficiencia. (Allan 1995)."* En este caso no se reduciría un documento a un único pasaje sino que se extraería y trataría individualmente cada pasaje del texto.

Los agentes de la Web Cooperativa utilizarían un híbrido de ambas técnicas ya que esta forma de actuar facilita la localización de nuevos recursos en una comunidad incipiente (Burke 1999), aquella en la que aún no se han evaluado muchos documentos. En el siguiente punto se presentan ejemplos ilustrativos de ambos modos de funcionamiento del sistema.

9.4 Aplicaciones y limitaciones de la Web Cooperativa

En esta propuesta existen dos mecanismos de recuperación de información; el primero es comparable a los actuales motores de búsqueda mientras que el otro exploraría la Web en búsqueda de información que pudiera recomendar a los usuarios.

El primer sistema permitiría "consultas" similares a las descritas a continuación:

- "Encuentra documentos con el término `estrella`". Al tratarse de un término muy genérico el sistema no debería proporcionar ningún resultado sino indicar al usuario términos relacionados con el original en función del contexto. Así, podría ofrecer contextos que contuvieran, cada uno, conceptos como `Star Wars`, `astronomía`, `cine`, `música pop`, etc. Obviamente, un aspecto muy importante sería la interfaz que permitiría visualizar tales opciones.
- Encontrar documentos relacionados con una sentencia, párrafo o documento seleccionado por el usuario. El usuario introduciría un fragmento de texto o un *URI* y el sistema procedería a clasificar dicha información en una rama del árbol taxonómico retornando documentos de esa rama (o de ramas vecinas). De nuevo, en caso de que el texto de partida fuera excesivamente genérico no se proporcionarían resultados sino sugerencias para refinar la búsqueda.

Por supuesto, esto es sólo un primer esbozo del sistema de búsqueda; aspectos fundamentales para el mismo serían las técnicas de visualización de datos, así como aquellas para explorar los árboles taxonómicos u ordenar los resultados en función del usuario.

El sistema de recomendación funcionaría de forma ligeramente distinta; se trataría, básicamente, de un asistente personal que ayudaría al usuario realizando tareas como las siguientes:

- Buscar información en representación del usuario. El usuario proporcionaría al agente algunas consultas como las presentadas antes para que las procesara y extrajera un conjunto reducido de resultados.
- Recomendar documentos no solicitados pero interesantes. Para llevar a cabo esta tarea el asistente debería buscar documentos similares a otros procesados recientemente por el usuario así como intercambiar información con agentes similares; de esta forma sería posible satisfacer demandas latentes de información.

Si se compara la Web Cooperativa con la Web actual y con la Web Semántica está claro que esta propuesta proporciona menos resultados que los motores de búsqueda tradicionales aunque mucho más relevantes puesto que se están empleando taxonomías conceptuales. Por otro lado, al obtener semántica a partir del texto completo de los documentos la Web Cooperativa permite consultas difíciles para la Web Semántica a menos que se proporcione una ontología capaz de describir todos los conceptos y relaciones existentes, algo imposible en la mayor parte de los casos (p. ej., ¿Sería posible desarrollar una ontología lo suficientemente sutil como para describir la Informática y permitir cualquier consulta concebible?)

Por supuesto, consultas admisibles en la Web Semántica como "*Encontrar el artículo más reciente sobre SHOE en el que James Hendler figure como coautor* (Denker *et al* 2001)" no podrían ser resueltas satisfactoriamente en la Web Cooperativa. Por esa razón la Web Cooperativa se propone como complemento de la Web Semántica y no como sustituto.

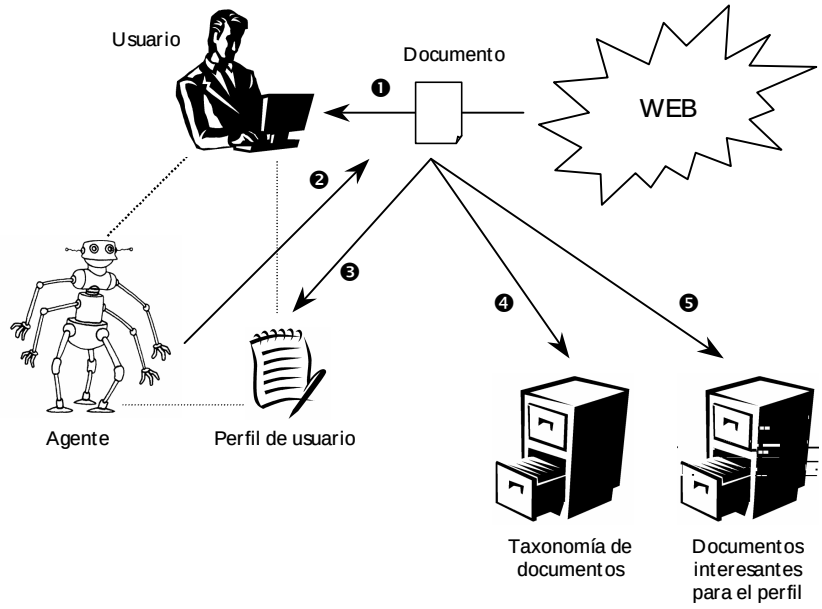


Fig. 8 Funcionamiento básico de la Web Cooperativa.

El usuario navega por la Web de la manera usual y descarga un documento ❶, su agente observa todas sus acciones y en función de éstas valora el interés del documento para el usuario ❷. Una vez evaluada la relevancia del documento, el agente actualiza el perfil del usuario sobre la base de la nueva información ❸, clasifica el documento en caso necesario dentro de la taxonomía de documentos ❹ (que estaría alojada en un servidor central) y agrega el documento, en caso de que la valoración sea positiva, a un "repositorio" de documentos de interés para el perfil del usuario que representa ❺ (también alojado en un servidor central).

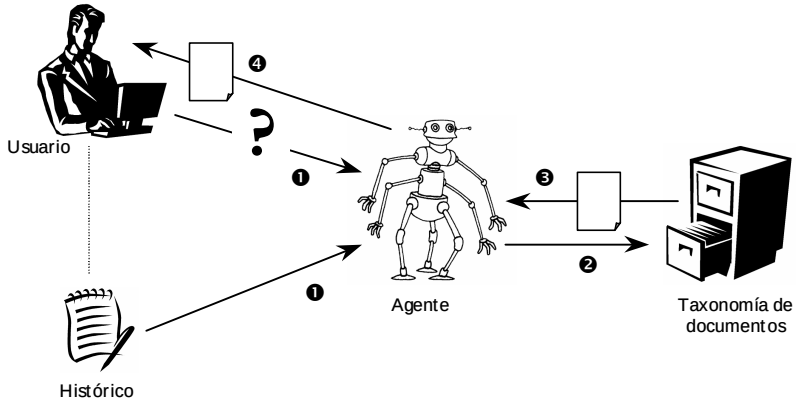


Fig. 9 Resolución de consultas y recomendación por contenidos en la Web Cooperativa.

Los agentes de la Web Cooperativa pueden resolver consultas de los usuarios además de explorar en representación de los mismos (recomendar documentos cuyos contenidos pueden ser interesantes). El agente puede examinar el histórico de navegación del usuario o recibir una consulta ❶. Con esta información el agente lleva a cabo una exploración taxonómica ❷, es decir, clasifica dentro de la taxonomía conceptual los datos de partida y obtiene como resultados documentos próximos en el dendrograma ❸. Estos documentos son proporcionados al usuario como recomendaciones en caso de que el agente haya actuado *motu proprio* o como resultados de una consulta ❹.

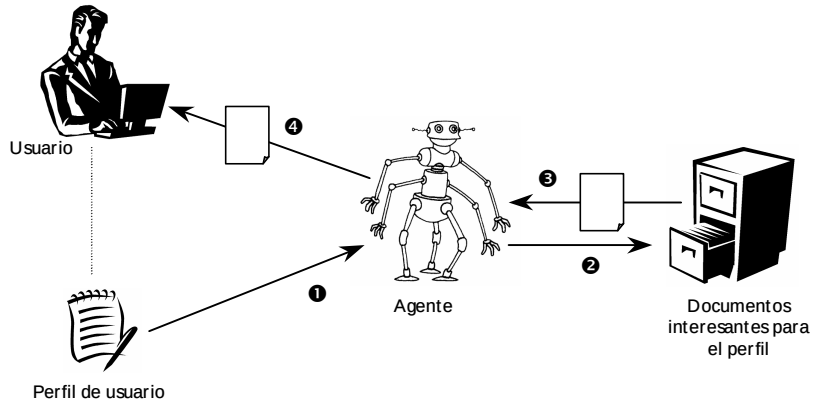


Fig. 10 Recomendación por filtrado colaborativo en la Web Cooperativa.

Los agentes de la Web Cooperativa pueden recomendar documentos de interés para el usuario basándose en las preferencias de usuarios similares. Periódicamente, cada agente accedería a distintos repositorios ❷ en función del perfil de su maestro ❶. De cada repositorio se obtendrían una serie de documentos potencialmente interesantes ❸ que serían presentados al usuario como recomendaciones ❹.

10 ¿Qué NO es la Web Cooperativa?

A la luz de lo visto hasta ahora es posible proporcionar una definición para la Web Cooperativa:

> *"La Web Cooperativa es una capa situada directamente sobre la Web actual con el fin de dotarla de semántica de manera global, automática, transparente e independiente del idioma. Requiere la participación de los usuarios pero no de forma consciente y directa sino indirectamente a través de agentes autónomos y cooperantes. La Web Cooperativa se apoya sobre el uso de conceptos y taxonomías documentales; unos y otras pueden obtenerse, sin intervención humana, a partir del texto libre de los documentos."*

En los apartados anteriores se ha planteado el problema, se ha situado en un contexto más amplio y se han mostrado iniciativas que han tratado de resolverlo parcialmente y la forma en que éstas han inspirado y motivado al autor en la propuesta de la Web Cooperativa dentro de la cual, como se verá más adelante, se enmarca su tesis. Existen, sin embargo, distintos proyectos que no estando relacionados con esta propuesta podrían parecer, engañosamente, similares; la finalidad de este apartado es diferenciar la propuesta de Web Cooperativa de estas otras.

10.1 La Web Cooperativa NO es la Web Semántica

La Web Cooperativa pretende extraer semántica de los documentos existentes en la Web, "clasificar" los documentos en una taxonomía o dendrograma y utilizar agentes. A la vista de esto es posible intentar compararla con la Web Semántica; sin embargo, eso sería un error puesto que las diferencias entre ambas iniciativas son enormes.

La Web Semántica requiere ontologías, sean estas construidas automáticamente o desarrolladas por un ser humano; dichas ontologías definen clases y relaciones que permiten etiquetar documentos para, así, facilitar un proceso de inferencia a los agentes de la Web Semántica.

De este modo, en la Web Semántica hasta que un concepto no está recogido en una ontología no existe pues no puede ser nombrado de ningún modo. Por otro lado, ya se ha

comentado anteriormente que la Web Semántica, a pesar de su nombre, ofrece a la Web más metasemántica que semántica.

La Web Cooperativa, por otro lado, no emplea ontologías, sólo conceptos. Este enfoque es mucho más simple puesto que no interesa explicitar en modo alguno las relaciones entre los conceptos. Esto no quiere decir que la Web Cooperativa ignore las relaciones entre conceptos sino que son manipuladas implícitamente.

Como ya se dijo, cada pasaje de cada documento es una secuencia de conceptos y el autor cree que dichas secuencias conceptuales pueden ser procesadas de modo similar a como el ADN es procesado para establecer clasificaciones de seres vivos. Esta clasificación conceptual automática, en caso de ser posible, sería capaz de separar documentos de un modo similar a como haría un ser humano dejando patentes las relaciones implícitas entre conceptos.

Por otro lado, los agentes de la Web Semántica y la Web Cooperativa tendrían misiones muy distintas. Los primeros tendrían como finalidad procesar documentos etiquetados "semánticamente" y realizar inferencias. Los segundos procesarían documentos no etiquetados, aprenderían de sus maestros e intercambiarían información entre ellos con el objetivo de recomendar información interesante.

Por todo ello, aun cuando tanto la Web Semántica como la Web Cooperativa emplean agentes, elementos semánticos y establecen algún tipo de catalogación de documentos, se trata de propuestas totalmente distintas (aunque como se ha señalado anteriormente complementarias).

10.2 La Web Cooperativa NO son las categorías *dmoz* o *Yahoo!*

Un aspecto vital de la Web Cooperativa es la clasificación de documentos en taxonomías o dendrogramas. Tales dendrogramas permitirían mostrar las relaciones conceptuales existentes entre los documentos de forma análoga a como se visualizan las relaciones que hay entre distintas especies biológicas y deberían obtener, de forma automática, "categorías" de documentos muy similares a las que podría establecer un ser humano.

Estas categorías pueden recordar a las disponibles en directorios como *dmoz*[1], *looksmart*[2] o *Yahoo!*[3]; no obstante, aun cuando es posible un parecido superficial, las diferencias de fondo entre las taxonomías documentales de la Web Cooperativa y las de estos directorios son notables.

Recuérdese que la propuesta que se plantea en este trabajo pretende generar de forma totalmente automática, no supervisada e independiente del idioma una o más taxonomías para los documentos disponibles en la Web; la estrategia seguida por los directorios es, sin embargo, muy distinta.

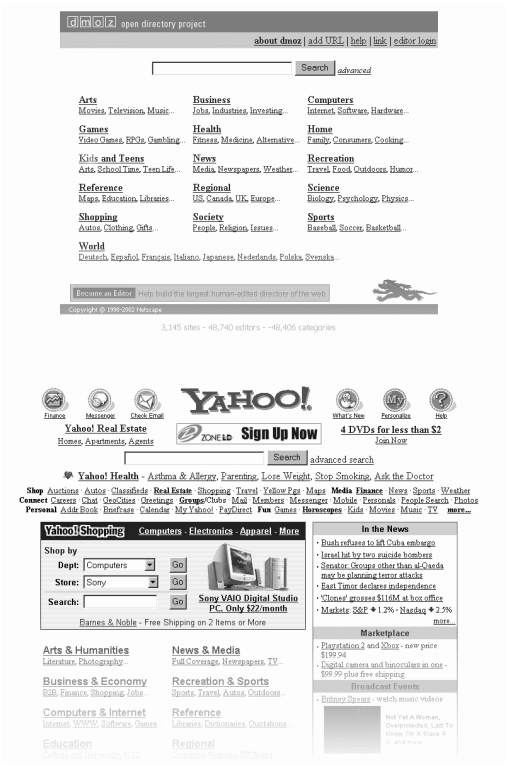
Todos los directorios requieren supervisión humana tanto para la creación de las categorías como para la asignación de documentos a las mismas. La forma en que se llevan a cabo estas tareas varía de un directorio a otro pero en ningún caso pueden realizarse de forma totalmente automática.

---

[1] `http://www.dmoz.org`

[2] `http://www.looksmart.com`

[3] `http://www.yahoo.com`



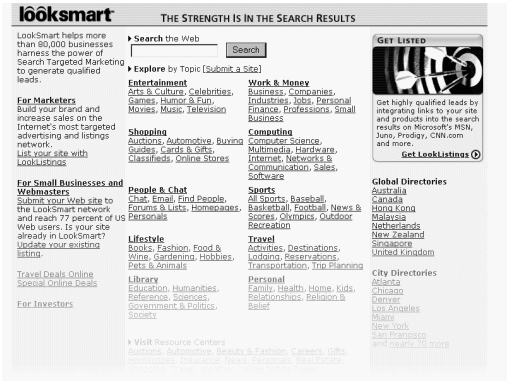


Fig. 11 Directorios *dmoz*, *looksmart* y *Yahoo!*

Página principal de los tres principales directorios de la Web a fecha de 20 de mayo 2002. Los tres comparten básicamente la misma estructura de categorías y proporcionan un motor de búsqueda que puede operar, al menos, sobre los documentos indexados en el directorio.

Yahoo! y *looksmart* cuentan con una plantilla dedicada a construir sus respectivos directorios. Estos empleados reciben sugerencias de los usuarios de la Web para añadir sitios web al directorio (no así para crear nuevas categorías). Esta estrategia está muy limitada y difícilmente puede desarrollar un directorio que abarque la totalidad, o cuando menos una parte importante, de la Web y que sea, al mismo tiempo, de calidad.

Además, esta forma de construir directorios tiende a "prostituirse" al promocionar determinadas categorías o documentos dentro de una categoría a cambio de una suma de dinero. Este hecho, aunque comercialmente justificable, sin duda degrada la utilidad que el directorio pudiera tener para los usuarios lo cual hace este método desaconsejable.

En el otro extremo se sitúan iniciativas como la de *dmoz*, también conocido como *Open Directory Project* u *ODP*. Se trata de un directorio desarrollado por una comunidad de usuarios que actúan desinteresadamente, de un modo similar a como se desarrolla el Software Libre. Cada categoría está gestionada por uno o más editores que revisan las sugerencias enviadas por los usuarios (documentos o propuestas de nuevas categorías), proporcionan una descripción para las mismas y organizan los documentos que aparecen en la categoría creando, en caso necesario, subcategorías.

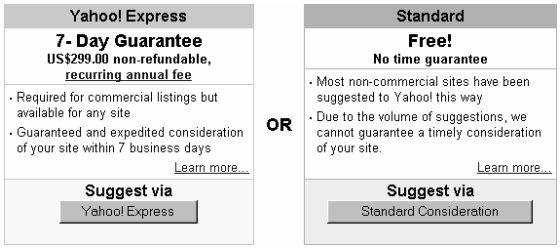


Fig. 12 Sugerencia de un nuevo sitio web para el directorio Yahoo!

Yahoo! permite a los usuarios sugerir nuevas entradas al directorio. A cambio de 299 dólares se "tomará en consideración" la "sugerencia" en sólo 7 días. La segunda opción (gratuita) no garantiza la inclusión del enlace en el directorio en ningún momento.

Esta estrategia colaborativa y altruista es superior a la empleada por directorios comerciales como los anteriormente mencionados puesto que es más fácilmente escalable y menos susceptible a la "corrupción". Sin embargo, a pesar de su mayor escalabilidad sigue sin poder abarcar una parte importante de la Web[1].

Por tanto, aun cuando las taxonomías que se proponen para la Web Cooperativa podrían coincidir en muchas ocasiones con las categorías disponibles en directorios como *dmoz* o *Yahoo!*, existen diferencias muy claras entre ambas iniciativas: los directorios son supervisados por humanos mientras que las taxonomías de la Web Cooperativa serían obtenidas de forma totalmente automática.

10.3 La Web Cooperativa NO es la Web Colaborativa

El término elegido para la propuesta, Web Cooperativa, tal vez no haya sido excesivamente afortunado puesto que puede llevar a confusión con algunas iniciativas calificadas, en ocasiones, como Web Colaborativa.

A diferencia de la Web Semántica que da nombre a una serie de líneas de investigación bien delimitadas, se ha empleado el término "Web Colaborativa" en varios proyectos, académicos y comerciales, que tienen poco o nada que ver entre sí ni con la propuesta de Web Cooperativa. A continuación se citan algunas de las aplicaciones calificadas en una u otra ocasión como Web Colaborativa.

GroupWeb (Greenberg y Roseman 1996) introduce el concepto de navegación colaborativa (*collaborative web browsing*) al presentar un sistema que permite a varios usuarios navegar de forma conjunta (recomendarse enlaces, seguir la ruta de navegación de otro usuario, explorar de forma combinada, etc.) Posteriormente, surgieron otra serie de iniciativas muy similares. Todos estos proyectos son, sin embargo, aplicaciones para trabajo en grupo y no sistemas de recuperación de información.

Sparrow Web (Chang 1998) fue un proyecto desarrollado en *Xerox PARC* que permitía a varios usuarios modificar una página web directamente mediante su navegador. Esta iniciativa se parece bastante a la idea de los *Wikis*[2] y, como se puede ver, no tiene ningún punto en común con la propuesta aquí descrita.

Kovács y Micsik (2000) emplean el término "Web Colaborativa" para hacer referencia a aplicaciones web que permiten el trabajo simultáneo de varios individuos. Sin embargo, describen aplicaciones relativamente tradicionales de filtrado colaborativo en *USENET*, Web y en una biblioteca digital (empleando en todos los casos valoración explícita) así como un sistema de encuestas y votaciones.

En resumen, la Web Colaborativa permite la colaboración de individuos de manera transparente ya sea para modificar documentos, explorar la Web o intercambiar información. Sin embargo, en el caso de la Web Cooperativa las entidades que cooperan son agentes que actúan en representación de los usuarios. Esta cooperación resulta transparente para el usuario que sigue empleando la Web de la manera usual.

---

[1] El directorio *dmoz* tenía indexados 3.429.012 ($3{,}4 \cdot 10^6$) sitios web a fecha de 28 de mayo de 2002, contando para ello con 49.030 editores; *Google* tenía indexadas 2.073.418.204 páginas ($2{,}1 \cdot 10^9$). Teniendo en cuenta que un directorio almacena una única página por sitio, *dmoz* está aún a tres órdenes de magnitud del volumen de documentos procesados por un sistema automático como *Google*.

[2] Un *Wiki* es un sitio web donde las páginas pueden ser editadas por cualquier visitante. Cualquier usuario puede ayudar a mejorar el sitio o plantear sus dudas, editando la página web, esperando que otro usuario las resuelva.

11 Formulación definitiva del problema y de la tesis

Parece innecesario decir que aún no existe ninguna implementación de la Web Cooperativa; sin embargo, utilizándola como una "vista desde 20.000 pies" resulta muy útil al plantear toda una serie de problemas interesantes:

- ¿Qué acciones del usuario sobre un documento son altamente discriminantes para determinar implícitamente su relevancia?
- ¿Es posible utilizar tales reglas de un modo eficiente dentro de un navegador web para determinar la relevancia del documento visualizado en un momento dado?
- ¿Es posible clasificar textos libres empleando métodos tomados de la biología computacional?
- ¿Es posible obtener un "pseudo-ADN" a partir de texto escrito en un lenguaje natural?
- Si existiera ese pseudo-ADN, ¿sería posible combinarlo, mutarlo o construir "algo" a partir del mismo como sucede con el ADN real?
- ¿Se debe suponer que idiomas distintos constituyen "bioquímicas" diferentes?
- ¿Cómo se daría el salto desde ese pseudo-ADN a los conceptos?
- ¿En qué forma podría un agente realizar búsquedas eficientes sobre una taxonomía de documentos construida a partir de ese pseudo-ADN?
- ¿Cómo obtendrían, representarían y almacenarían los agentes el perfil de sus maestros?
- ¿Cómo y dónde se comunicarían los agentes entre sí?
- ¿Qué información sobre el perfil de los respectivos maestros podría ser intercambiada?

Así pues, la Web Cooperativa involucra muy diversos aspectos: tratamiento de lenguaje natural, evaluación implícita de documentos, agentes *software*, interacción persona-ordenador, usabilidad o privacidad. En este trabajo se prescinde de lo que serían las "capas superiores" de la Web Cooperativa y se delimita aún más el problema encuadrándolo dentro del campo del procesamiento de lenguaje natural por medios estadísticos.

De este modo, a partir del problema original se formula otro más concreto sobre el que finalmente versa el presente trabajo:

> *La cantidad de texto no estructurado disponible en la Web seguirá aumentando y, a pesar de sus inconvenientes, el método preferido por la mayor parte de usuarios para recuperar información continuarán siendo las consultas formuladas en lenguajes naturales. En ambos casos (publicación y consulta) será inevitable un uso generalmente ambiguo de los distintos idiomas y la presencia de errores tipográficos, ortográficos o gramaticales.*

En relación con dicho problema, el autor sostiene la siguiente tesis:

> *Se puede obtener para los distintos n-gramas, g_i, de un texto escrito en cualquier idioma una medida de su significatividad, s_i, distinta de la frecuencia relativa de aparición de los mismos en el texto, f_i, pero calculable a partir de la misma. Esta métrica de la significatividad intradocumental de los n-gramas permite asociar a cada documento, d_i, un único vector, v_i, susceptible de comparación con cualquier otro vector obtenido del mismo modo aun cuando sus respectivas longitudes puedan diferir. Puesto que tales vectores almacenan ciertos aspectos de la semántica subyacente a los textos originales, el mayor o*

menor grado de similitud entre los mismos constituye un indicador de su nivel de relación conceptual, facilitando la clasificación y categorización de documentos, así como la recuperación de información. Asimismo, cada vector individual es capaz de transformar el texto original a partir del cual fue obtenido dando lugar a secuencias de palabras clave y resúmenes automáticos.

Siendo ésta su versión resumida:

Una única técnica sencilla, basada en el uso de vectores de n-gramas de longitud variable, independiente del idioma y aplicable a diversas tareas de tratamiento de lenguaje natural con resultados similares a los de otros métodos 'ad hoc' es viable.

A lo largo de los siguientes capítulos se procederá a describir los fundamentos de la nueva técnica propuesta por el autor y su relación con otras existentes. Se demostrará que, efectivamente, es posible obtener de forma sencilla y automática representaciones de documentos que conservan aspectos semánticos de los mismos con independencia del idioma. Se describirán las aplicaciones de la técnica a diversas tareas de procesamiento de lenguaje natural. Y, para finalizar, se expondrán las conclusiones a las que ha llegado el autor del trabajo y se esbozarán posibles líneas de trabajo futuro.

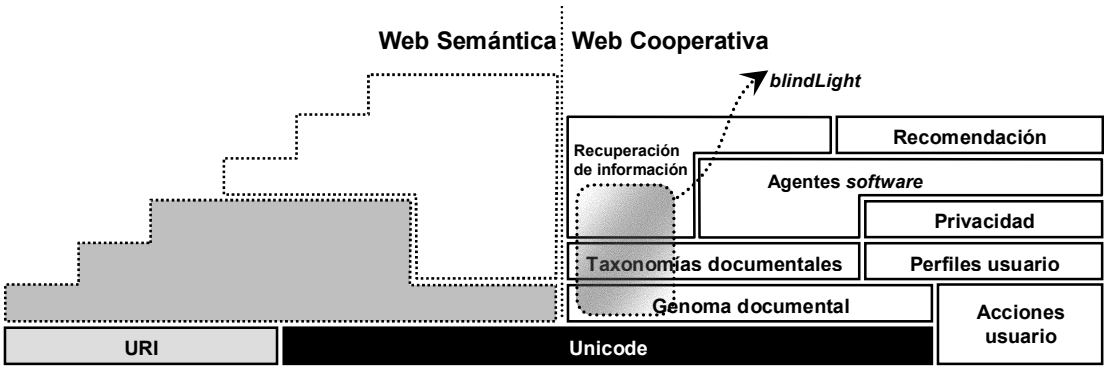


Fig. 13 Web Semántica vs. Web Cooperativa y relación de esta última con *blindLight*.