RECUPERACIÓN DE INFORMACIÓN.

Francisco Javier Martínez Méndez – José Vicente Rodríguez Muñoz

RECUPERACIÓN DE INFORMACIÓN.

ISBN:
Depósito legal:
Edita, Distribuye y Vende:
Imprime:

Índice General.

1 La Recuperación y los Sistemas de Recuperación de Información.

.

Este capítulo constituye una presentación del concepto de recuperación de información, y del conjunto de diferencias que esta idea posee con otras aplicaciones de la Informática en todo lo relacionado con la gestión y recuperación de datos. Al mismo tiempo se exponen los distintos modelos sobre los que se basan los sistemas que permiten llevar a cabo estas operaciones de recuperación de información.

1.1 Hacia una definición de la Recuperación de Información.

Si bien no es algo infrecuente, puede resulta cuando menos curioso que un concepto tan empleado como el de *recuperación de información* presente una cierta problemática a la hora de establecer una definición que lo sitúe adecuadamente dentro del campo de las *Ciencias de la Información*, a causa de las distintas y divergentes perspectivas con las que este concepto ha sido analizado a lo largo del tiempo.

Es 'Keith' Rijsbergen[1] el autor que mejor introduce este problema al considerar que "se trata de un término que suele ser definido en un sentido muy amplio" (Rijsbergen, 1999). Ha sido, en realidad, el profuso uso de este término, al igual que ocurre en otras disciplinas con otros vocablos que también pueden parecer básicos, lo que ha propiciado que el mismo no se encuentre bien utilizado en muchas ocasiones, siendo el fallo más común presentarlo como sinónimo de la recuperación de datos llevada a cabo desde la perspectiva de las base de datos. En otros casos, encontramos autores que expresan las diferencias que, a su juicio, presentan ambos conceptos, con lo cual la definición de recuperación de información queda, en cierto modo, supeditada a la de recuperación de datos (se define a partir de diferencias más que aportar algo nuevo y característico). También disponemos de definiciones muy genéricas, elaboradas de manera superficial y sin entrar en mayores consideraciones sobre estas diferencias.

[1] El profesor C. J. 'Keith' Rijsbergen es miembro del Departamento de Informática de la Universidad de Glasgow y es director del Grupo de Recuperación de Información (http://ir.dcs.gla.ac.uk/). Es el autor de la obra más afamada dentro de este campo, el libro 'Information Retrieval', publicado inicialmente en el año 1979 y del cual afortunadamente disponemos de una edición en línea publicado en 1999 y que está accesible en la URL http://www.dcs.gla.ac.uk/Keith/Preface.html [Consulta: 10 octubre 2006].

Finalmente, llama la atención que algunos autores pasa de largo sobre este problema, obviando este debate y profundizando más en la explicación de los *sistemas de recuperación de información*[2] (SRI en adelante).

El primer grupo de definiciones se encuentra claramente influenciado por la tecnología informática, cuya evolución ha inducido a considerar sinónimos ambos conceptos, llegándose a olvidar que se puede recuperar información sin recurrir a procedimientos informáticos (aunque evidentemente no es lo más común hoy en día). El frecuente y necesario empleo de una tecnología no debe sustituir el adecuado uso de los conceptos terminológicos. Un claro ejemplo de este desacierto es el *Glosario de la Asociación de Bibliotecarios Americanos* que define eltérmino "information retrieval" como *recuperación de la información* en primera acepción y como *recuperación de datos* en una segunda acepción (ALA, 1983), considerando ambos términos sinónimos en lengua inglesa[3]. Igualmente, el *Diccionario Mac Millan de Tecnología de la Información* presenta la recuperación de información como "el conjunto de técnicas empleadas para almacenar y buscar grandes cantidades de datos y ponerlosa disposición de los usuarios" (LON, 1989).

Un segundo grupo de autores establecen diferencias entre ambos conceptos. Meadow afirma que la recuperación de la información es "una disciplina que involucra la localización de una determinada información dentro de un almacén de información o base de datos" (Meadow, 1992), estableciendo implícitamente una asociación entre la recuperación de información y el concepto de *selectividad* en el cual la información específica ha de extraerse siguiendo algún tipo de criterio discriminatorio (selectivo por tanto). Pérez-Carballo y Strzalkowski redundan en esta tesis: "una típica tarea de la recuperación de información es *traer* documentos relevantes desde una gran archivo en respuesta a una pregunta formulada y ordenarlos de acuerdo con su *relevancia*" (Pérez-Carballo, 2000). Del mismo modo, Grossman y Frieder indican que recuperar información es "encontrar documentos relevantes, no encontrar simples correspondencias a unos patrones de bits" (Grossman, 1998). Meadow considera que no es lo mismo la recuperación de información entendida como traducción del término inglés *information recovery*[4] que cuando se traduce el término

[2] Quizás el esfuerzo realizado por los autores en definir los sistemas ha favorecido que el concepto de recuperación haya quedado relegado a un segundo plano.

[3] Este Glosario afirma que "document retrieval" es un término sinónimo de "information retrieval".

[4] De hecho, con el paso del tiempo, este término se viene utilizando más para referirse a los procesos y a las aplicaciones que permiten recuperar datos perdidos o dañados en un soporte de almacenamiento.

information retrieval, ya que "en el primer caso no es necesario proceso de selección alguno" (Meadow, 1992).

El autor que más extensamente se dedica a presentar estas diferencias es Blair[5], utilizando como criterios distintivos, entre otros (Blair, 1990):

1. En recuperación de datos se emplean preguntas altamente formalizadas, cuya respuesta es directamente la información deseada. En cambio, en recuperación de información las preguntas resultan difíciles de trasladar a un lenguaje normalizado (aunque existen lenguajes para la recuperación de información, son de naturaleza mucho menos formal que los empleados en los sistemas de bases de datos relacionales, por ejemplo) y la respuesta será un conjunto de documentos que probablemente contendrá lo deseado, con un evidente factor de indeterminación.
2. Según la relación entre el requerimiento al sistema y la satisfacción de usuario, la recuperación de datos es *determinista* y la recuperación de información es *posibilista*, debido al nivel de incertidumbre presente en la respuesta.
3. Éxito de la búsqueda. En recuperación de datos el criterio a emplear es la exactitud de lo encontrado, mientras que en recuperación de información, el criterio de valor es el grado en el que la respuesta satisface las necesidades de información del usuario, es decir, su percepción personal de utilidad.

Tramullas Saz resalta especialmente un aspecto de las reflexiones de Blair: la importancia (en muchas ocasiones ignorada) que tiene el factor de predicción por parte del usuario. No debemos olvidar que el usuario ha de intuir, en numerosas ocasiones, los términos utilizados para representar el contenido de los documentos, independientemente de la presencia de mecanismos de control terminológico. Este criterio "es otro de los elementos que desempeñan un papel fundamental en el complejo proceso de la recuperación de información" (Tramullas Saz, 1997) y además no se presenta en el campo de la recuperación de datos. Rijsbergen compendia en la siguiente tabla las diferencias fundamentales existentes entre *recuperación de datos* y *recuperación de información:*

	Recuperación de datos	**Recuperación de información**
Acierto	Exacto	Parcial, el mejor

[5] Blair dedica gran parte de la presentación de su libro '*Language and representation in information retrieval*' a asentar las diferencias entre recuperación de datos y recuperación de información.

Inferencia	Algebraica	Inductiva
Modelo	Determinístico	Posibilístico
Lenguaje de consulta	Fuertemente Estructurado	Estructurado o Natural
Especificación consulta	Precisa	Imprecisa
Error en la respuesta	Sensible	Insensible

Tabla 1.1 Recuperación de datos vs Recuperación de Información. Fuente: Rijsbergen, C.J. *Information Retrieval*. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 21 de octubre de 2001]

Baeza-Yates plantea las diferencias entre ambos tipos de recuperación con argumentos quizá algo menos abstractos que losempleados por otros autores, incidiendo en que "los datos se pueden estructurar en tablas, árboles, etc. para recuperar exactamente lo que se quiere, el texto no posee una estructura clara y no resulta fácil crearla" (Baeza-Yates, 1999). Para este autor, el problema de la recuperación de información se define de la siguiente manera: "dada una necesidad deinformación (consulta + perfil del usuario + ...) y un conjunto de documentos, ordenar los documentos de más a menos relevantes para esa necesidad y presentar un subconjunto de aquellos de mayor relevancia".

En la solución de este problema se identifican dos grandes etapas:

1. Elección de un modelo que permita calcular la *relevancia* de un documento frente a una consulta.
2. Diseño de algoritmos y estructuras de datos que implementen este modelo de forma eficiente.

Baeza-Yates se preocupa especialmente de las estructuras de datos y métodos de acceso a los mismos siendo este autor una verdadera referencia en esta materia[6]. Curiosamente, a la hora de definir la recuperación de información, en lugar de proponer una definición propia, emplea la elaborada por Salton: "la recuperación de la información tiene que ver con la representación, almacenamiento, organización y acceso a los ítem de información" (Salton & McGill, 1983). En principio, no deben existir limitaciones a la naturaleza del objeto informativo. Baeza-Yates incorpora la reflexión siguiente: "la representación y organización debería proveer al usuario un fácil acceso a la información en la que se encuentre interesado.

[6] Para todos los estudiosos de la recuperación de información resultan de obligada lectura los libros *Information retrieval: data structures & algorithms* (Baeza-Yates, 1992) y *Modern information retrieval* (Baeza-Yates, 1999). El texto completo de la *Introducción* y del capítulo 10 (*Interfaces de usuario y Visualización*) están disponibles para su consulta en la URL: http://www.ischool.berkeley.edu/~hearst/irbook/ junto con tablas de contenidos, ejercicios y recursos de otros capítulos.

Desafortunadamente, la caracterización de la necesidad informativa de un usuario no es un problema sencillo de resolver" (Baeza-Yates, 1999).

El tercer grupo de autores emplea la definición de Salton (base de la mayoría de definiciones de a bibliografía especializada), añadiendo como rasgo diferenciador común que estos autores no profundizan en escrutar las diferencias entre "recuperación de datos" y "recuperación de información", bien por no ser objeto de sus trabajos o bien por considerarlas suficientemente establecidas en trabajos previos. Feather y Storges ven a la recuperación de información como "el conjunto de actividades necesarias para hacer disponible la información a una comunidad de usuarios" (IEI, 1997). Croft concibe la recuperación de información como "el conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental, etc." (Croft, 1987). Tramullas Saz impregna su definición del aspecto selectivo de Blair comentado anteriormente, afirmando que "elplanteamiento de la recuperación de información en su moderno concepto ydiscusión, hay que buscarlo en la realización de los tests de *Cranfield* y enla bibliografía generada desde ese momento y referida a los mecanismos más adecuados para extraer, de un conjunto de documentos, aquellos que fuesen pertinentes a una necesidad informativa dada" (Tramullas Saz, 1997).

El cuarto y último grupo de autores se distinguen básicamente porque eluden definir la recuperación de la información. Tienen como máximo exponente a Chowdhury, quien simplemente dedica el primer párrafo de su libro *'Introduction to modern information retrieval'* a señalar que "el término recuperación de la información fue acuñado en 1952 y fue ganando popularidad en la comunidad científica de 1961 en adelante[7]", mostrando después los propósitos, funciones y componentes de los SRI (Chowdhury, 1999). Otro autor de esta corriente es Korfhage[8], quien se centra en el almacenamiento y recuperación de la información, considerando a estos procesos como las dos caras de una moneda. Para este autor, "un usuario de un sistema de información lo utiliza de dos formas posibles: para almacenar información en anticipación de una futura necesidad, y para encontrar información en respuesta una necesidad" (Korfhage, 1997).

[7] Chowdhury introduce una cita de Rijsbergen y Agosti, correspondiente al artículo 'The Context of Information', *The Computer Journal*, vol 35 (2), 1992.

[8] Resulta especialmente curiosa la actitud de Chowdhury y de Korfhage quienes son autores de dos excelentes manuales sobre la recuperación de información.

1.2 Sistemas de Recuperación de Información.

Tomando como base de partida la definición propuesta por Salton y uniéndole las aportaciones de Rijsbergen, sería el momento (siguiendo la opinión de Baeza-Yates), de elegir el mejor modelo para el diseño de un SRI, aunque para ello creemos totalmente necesario definir de forma previa y adecuada qué se entiende por "sistema de recuperación de información".

1.2.1 Vista funcional de un SRI.

Las notorias similitudes existentes entre la recuperación de información y otras áreas vinculadas al procesamiento y manejo de la información, se repiten en el campo de los sistemas encargados de llevar a cabo esta tarea. Para Salton "la recuperación de información se entiende mejor cuando uno recuerda que la información procesada son documentos", con el fin de diferenciar a los sistemas encargados de su gestión de otro tipo de sistemas, como los gestores de bases de datos relacionales. Salton piensa que "cualquier SRI puede ser descrito como un conjunto de ítems de información (DOCS), un conjunto de peticiones (REQS) y algún mecanismo (SIMILAR) que determine qué ítem satisfacen las necesidades de información expresadas por el usuario en la petición" (Salton & McGill, 1983)

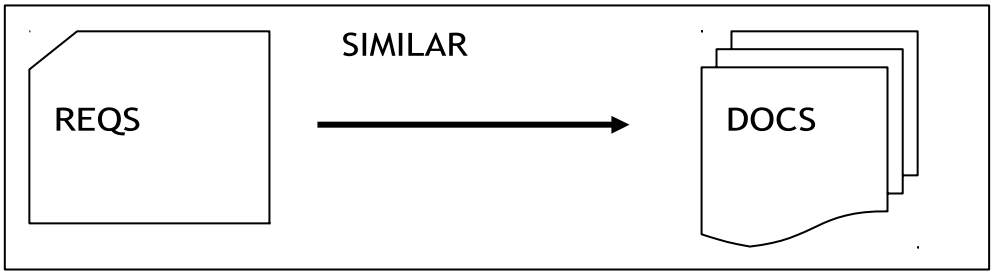

Ilustración 1.1 Esquema simple de un SRI. Fuente Salton , G. and Mc Gill, M.J. *Introduction to Modern Information Retrieval*. New York: Mc Graw-Hill Computer Series, 1983.

Es el mismo Salton quien reconoce que, una vez llevado a la práctica cotidiana, este esquema resulta muy simple y precisa deampliación, "porque los documentos suelen convertirse inicialmente a un formato especial, por medio del uso de una clasificación o de un sistema de indización, que denominaremos LANG" (Salton & McGill, 1983)

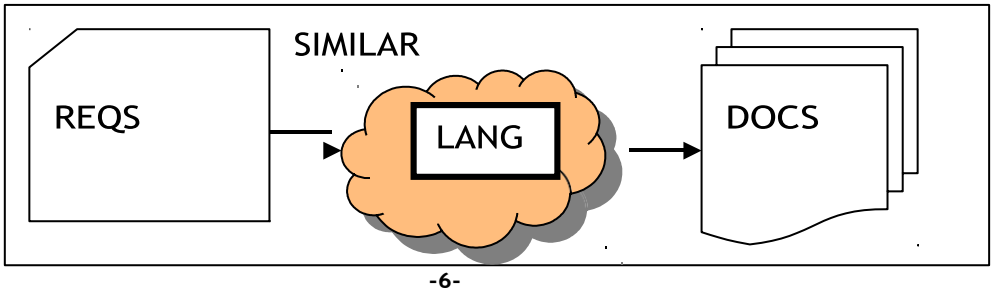

Ilustración 1.2 Esquema avanzado de un SRI. Fuente Salton , G. and Mc Gill, M.J. *Introduction to Modern Information Retrieval*. New York: Mc Graw-Hill Computer Series, 1983.

En la anterior ilustración se observa que el proceso establecido entre la entrada REQS y SIMILAR es la formulación de la búsqueda y el establecido entre SIMILAR y el conjunto de documentos DOCS es la recuperación. SIMILAR es el proceso de determinación de la similitud existente entre la representación de la pregunta y la representación de los ítems de información. Independientemente de la posible complejidad que pueda llegar a tener este proceso, la definición de Salton no puede ser más sencilla e intuitiva, pudiendo considerarse sublime[9].

Chowdhury identifica el siguiente conjunto de funciones principales en un SRI [CHO, 1999]:

1. Identificar las fuentes de información relevantes a las áreas de interés de las solicitudes de los usuarios.
2. Analizar los contenidos de los documentos.
3. Representar los contenidos de las fuentes analizadas de una manera que sea adecuada para compararlas con las preguntas de los usuarios.
4. Analizar las preguntas de los usuarios y representarlas de una forma que sea adecuada para compararlas con las representaciones de los documentos de la base de datos.
5. Realizar la correspondencia entre la representación de la búsqueda y los documentos almacenados en la base de datos.
6. Recuperar la información relevante
7. Realizar los ajustes necesarios en el sistema basados en la retroalimentación con los usuarios

Evolución de los SRI.

Son varios los autores que presentan la evolución de estos sistemas, pero quien mejor simplifica este progreso es Baeza-Yates, destacando tres fases fundamentales:

[9] Sublime según el Diccionario de la RAE es lo que "se dice especialmente de las concepciones mentales y de las producciones literarias y artísticas o de lo que en ellas tiene por caracteres distintivos grandeza y sencillez admirables". Es perfectamente válido este adjetivo para calificar la definición de Salton.

1. *Desarrollos iniciales*. Nunca debe olvidarse que ya se disponía de métodos de recuperación de información en las antiguas colecciones de papiros. Otro ejemplo clásico, que se ha venido utilizando hasta nuestros días, es la *tabla de contenidos* de un libro, sustituida por otras estructuras más complejas a medida que ha crecido el volumen de información. La evolución de la tabla de contenidos es el *índice*, estructura que aún constituye el núcleo de los SRI actuales.

2. *Recuperación de información en las bibliotecas*. Fueron lasprimeras instituciones en adoptar estos sistemas. Originalmente fueron desarrollados por ellas mismas y posteriormente se ha creado un mercado tecnológico altamente especializado en el que participan múltiples empresas e instituciones.

3. *La World Wide Web*. La dirección más natural de la evolución de los SRI ha sido hacia la web, entorno donde han encontrado una alta aplicación práctica y un aumento del número de usuarios, especialmente en el campo de los directorios y motores de búsqueda[10]. El alto grado de consolidación de la web está siendo favorecido por el vertiginoso abaratamiento de la tecnología informática, por el espectacular o desarrollo de las telecomunicaciones y por la facilidad de publicación de cualquier documento que un autor considere interesante, sin tener que pasar por el filtro de los tradicionales círculos editoriales (Baeza-Yates, 1999).

Lesk lleva a cabo una curiosa presentación de la evolución de la recuperación de la información considerándola como un ser humano que va atravesando diversos períodos de su existencia[11]:

1. *El niño de guardería* (1945-1955): el niño nace con los planteamientos teóricos de Vannevar Bush (1945) -muchos de los cuales se han alcanzado posteriormente con la evolución de la tecnología- y los índices KWIC propuestos por Luhn, el precursor de la Indización Automática (Rijsbergen, 1999).

2. *El niño del colegio* (1960s): es la época donde algunos de los hoy principales sistemas de recuperación de información (*Dialog* y *BRS* por ejemplo) son proyectos embrionarios. Al

[10] Estos sistemas, conocidos de forma más común como buscadores, se presentan posteriormente en el apartado dedicado a la *recuperación de la información en la web*, dentro de este mismo capítulo.

[11] El autor toma como referencia un texto de la obra 'As you like it', Acto 2º, Escena 7ª, escrita por el dramaturgo inglés **William Shakespeare** hacia 1599. El melancólico Jaques, asistente del exiliado Duque Mayor, protagonista de la obra, compara la vida con una obra de teatro, coloca en el catálogo las siete etapas del crecimiento: *infante*, *escolar*, *amante*, *soldado*, *justiciero*, *pantalonero* y *segunda niñez*.

mismo tiempo comienzan a desarrollarse las primeras bases de datos de repertorios bibliográficos (*Chemical Abstracts* o *ERIC* por ejemplo). También en esa época tienen lugar los experimentos de Cleverdon en el *Instituto Cranfield* (Sparck Jones, 1981), (Cleverdon, 1991).

3. *El adulto* (1970s): cuando comienzan a tomar forma los primeros proyectos gracias al desarrollo de los periféricos de entrada de textos (teclados) que permitían construir grandes colecciones de textos mejor que los lectores de tarjetas perforadas. El otro factor clave son los sistemas de proceso de datos en tiempo real. En esta época surge –de forma incipiente- OCLC el mayor recolector de información bibliográfica a nivel mundial y se desarrolla el formato MARC para la catalogación automatizada de las colecciones de fondos de las bibliotecas. También es cuando se funda NSF ('National Science Fundation') institución que tanta importancia va a tener en las décadas siguientes. En esta época la comunidad de investigadores en recuperación de información coincide mucho con los dedicados a la Inteligencia Artificial. A nivel conceptual, el mayor avance lo constituye el *modelo probabilístico de recuperación de información,* introducido por 'Keith' Rijsbergen (1999).

4. *El maduro* (1980s): al mismo tiempo que aumentan las facilidades para la entrada de la información disminuye el coste de los dispositivos de almacenamiento, aunque este proceso tiene su culminación en la década siguiente. Especialmente destaca en estos años la expansión del CD-ROM, hecho que revolucionó la entonces incipiente *Industria de la Información*. Paralela la distribución de información en este formato asistimos al desarrollo de los primeros módulos de consulta de catálogos en línea (OPAC), que también alcanzan su plena expansión en la década siguiente gracias a la World Wide Web. En esta época se afianza la investigación en la producción de algoritmos eficientes para la recuperación de la información, correspondiendo a Croft y Fox los más conocidos. Pero si algo merece ser recordado de forma prioritaria en esta época,no es otra cosa que la propuesta de aplicación del *modelo del espacio vectorial* de Gerard Salton en la recuperación de la información.

5. *La crisis de los cuarenta* (1990s): para quien no haya llegado aún a esa edad sólo decirle que no es un mito pero que se

supera. En esta década el niño tiene ya 45 años aproximadamente y durante los primeros años parece funcionar con el piloto automático contentándose con los desarrollos del período anterior. Pero todo comienza a cambiar radicalmente en cuanto Tim Berners-Lee y Paul F. Kunz crean la *World Wide Web*, entorno que para muchos representa la plasmación de los postulados de Vannevar Bush y es cuando cabe preguntarse si el niño ha alcanzadosu meta. Desde este momento todo cambia, tanto a nivel del sector industrial (surgen Netscape, Yahoo!, Lycos, Altavista y finalmente, en 1999, nace Google) como en el número de usuarios de los sistemas de recuperación de información (su volumen crece de forma exponencial). WAIS se conviertenen el estándar de los sistemas de búsqueda enbases de datos en la web, surgen los primeros índices y motores de búsqueda de recursos en los servidores web y el protocolo Z39.50 se consolida como la base para el desarrollo de las OPAC. En esta época nace, se consolida y finalmente desaparece *Hytelnet*, plataforma integradora para la consulta de catálogos en línea de multitud de bibliotecas de todo el mundo. La crisis de los cuarenta se supera, fijo.

6. *El deber cumplido* (2000s): habiendo llegado a satisfacer (y a mejorar) gran parte de las ideas originarias de Bush, bien podría considerarse que la recuperación de información ha desarrollado con éxito su tarea y puede aspirar a un merecido descanso. Aunque esta idea pueda resultar apetecible, quizá no sea aún ese momento. Si bien se han producido avances en el proceso y la recuperación de la información textual, aún se presentan muchos problemas en la Indización Automática y no digamos ya en el campo de la recuperación de información multimedia. También está por ver si los algoritmos y sistemas desarrollados en los años 80sy 90s pueden hacer frente a las grandes colecciones de documentos que se están construyendo. Finalmente, el sectorindustrial está haciendo grandes esfuerzos por trasladar el escritorio de trabajo de cada usuarios desde el ordenador personal hacia la red.

7. *El retiro:* si bien Lesk lo fija de manera muy optimista para el año 2010 no creo que todos tengamos tanta suerte. El mismo podría resultar válido cuando todos los países del mundo posean un similar nivel en cuanto al desarrollo de los

servicios y productos de la Sociedad de la Información. Aún queda mucho para ello (Lesk, 1995).

Los sistemas de recuperación de la información han evolucionado con el fin de adaptarse a este nuevo entorno, habiéndose llegado a desarrollar algunos de los sistemas más innovadores, al mismo tiempo que extensos, por no hablar de su popularidad, si bien aún no disponemos de metodologías suficientemente consolidadas que evalúen su efectividad. Esta evolución no es un proceso finalizado, sino más bien un proceso en realización, que lleva al establecimiento de nuevos términos, tales como WIS ('web information systems' o "sistemas de información basados en la tecnología web) destinados a integrarse plenamente con otros sistemas convencionales, llegando a ser más extendidos y de mayor influencia tanto en negocios como en la vida familiar" (Wan, 2001).

Modelos para la recuperación de información.

El diseño de un SRI se realiza bajo un modelo, donde queda definido "cómo se obtienen las representaciones de los documentos y de la consulta, la estrategia para evaluar la *relevancia* de un documento respecto a una consulta y los métodos para establecer la importancia (orden) de los documentos de salida" (Villena Román, 1997). Existen varias propuestas de clasificación de modelos, una de las síntesis más completas la realiza Dominich en cinco grupos (Dominich, 2000):

Modelo	**Descripción**
Modelos clásicos	Incluye los tres más comúnmente citados: *booleano*, *espacio vectorial* y *probabilístico*.
Modelos alternativos	Están basados en la Lógica Fuzzy
Modelos lógicos	Basados en la Lógica Formal. La recuperación de información es un proceso inferencial.
Modelos basados en la interactividad	Incluyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la *relevancia* de los documentos recuperados.
Modelos basados en la Inteligencia Artificial[12]	Bases de conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural.

Tabla 1.2 Clasificación de los Modelos de Recuperación de Información según Dominich. Fuente: Dominich, S. 'A unified mathematical definition of classical

[12] Respetando los grupos de Dominich, discrepamos con su opinión de no considerar los Modelos Lógicos parte de los Modelos basados en la Inteligencia Artificial. Realmente podrían englobarse en el mismo grupo de modelos.

information retrieval'. Journal of the American Society for Information Science, 51 (7), 2000. p. 614-624.

Baeza-Yates clasifica los modelos de recuperación de información con base en la tarea inicial que realiza el usuario en el sistema: (1) recuperar información por medio de una ecuación de búsqueda (*retrieval*) que se inserta en un formulario destinado a ello, o (2) dedicar un tiempo a consultar (*browse*) los documentos en la búsqueda de referencias (Baeza- Yates, 1999), dando entrada en su clasificación al *hipertexto* [CON, 1988] [NIE, 1990], modelo en el cual se basa la web [BER, 1992].

Este mismo autor divide a los modelos basados en la recuperación en dos grupos: clásicos y estructurados. En el primero de ellos incluye a los modelos booleano, espacio vectorial y probabilístico. Posteriormente, presenta una serie de paradigmas alternativos a cada modelo: teoría de conjuntos (conjuntos difusos y booleano extendido), algebraicos (vector generalizado, indización por semántica latente y redes neuronales), y por último, probabilísticos (redes de inferencia y redes de conocimiento); los modelos estructurados corresponden a listas de términos sin solapamiento y a nodos próximos (son modelos escasamente difundidos). Los modelos basados en la navegación entre páginas web son de tres tipos: estructura plana, estructura guiada e hipertexto.

El primero es una simple lectura de un documento aislado del contexto, el segundo incorpora la posibilidad de facilitar la exploración organizando los documentos en una estructura tipo directorio con jerarquía de clases y subclases y el tercero se basa en la idea de un sistema de información que de la posibilidad de adquirir información de forma no estrictamente secuencial sino a través de nodos y enlaces (Baeza-Yates, 1999). Es también Baeza-Yates quien proporciona una clasificación adicional de estos modelos de recuperación de información, realizada en función de la modalidad de consulta y de la vista lógica de los documentos:

Vista lógica de los documentos

	Términos Índice	**Texto Completo**	**Texto Completo y Estructura**

Recuperación	Clásicos Conjuntos teóricos Algebraicos Probabilísticos	Clásicos Conjuntos teóricos Algebraicos Probabilísticas	Estructurados
Navegación	Estructura plana	Estructura plana Hipertexto	Estructura guiada Hipertexto

Tabla 1.3 Clasificación de los Modelos de Recuperación de Información según Baeza-Yates. Fuente: Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York: ACM Press ; Harlow [etc.]: Addison-Wesley, 1999 XX, 513 p.

Tanto Baeza-Yates (Baeza-Yates, 1999) como Villena Román (Villena Román, 1997) llevan a cabo una presentación detallada de cada uno de los modelos, siendo también interesante la lectura de Grossman y Frieder[13] [GRO, 1998], para conocer las alternativas a los modelos clásicos.

Modelo del Espacio Vectorial.

Vamos a prestar un poco más de atención a este modelo, el más utilizado en la actualidad en los SRI (especialmente en la web).

Este modelo entiende que los documentos pueden expresarse en función de unos vectores que recogen la frecuencia de aparición de los términos en los documentos. Los términos que forman esa matriz serían términos no vacíos, es decir, dotados de algún significado a la hora de recuperar información y por otro lado, estarían almacenados en formato "stemmed" (reducidos los términos a una raíz común, tras un procedimiento de aislamiento de la base que agruparía en una misma entrada varios términos).

Si nuestro SRI contiene los siguientes cuatro documentos:

D1: el río Danubio pasa por Viena, su color es azul

D2: el caudal de un río asciende en Invierno

D3: el río Rhin y el río Danubio tienen mucho caudal D4:

si un río es navegable, es porque tiene mucho caudal

Su matriz correspondiente dentro del modelo del Espacio Vectorial podría ser la siguiente:

[13] Estos autores prefieren emplear el término "estrategias de recuperación de información" en lugar del término "modelo".

	Río	Danubio	Viena	color	azul	caudal	invierno	Rhin	navegable
D1	1	1	1	1	1	0	0	0	0
D2	1	0	0	0	0	1	1	0	0
D3	2	1	0	0	0	1	0	1	0
D4	1	0	0	0	0	1	0	0	1

Tabla 1.4 Ejemplo de Matriz de términos y documentos en el Espacio Vectorial. Fuente: elaboración propia.

Por medio de un proceso denominado *stemming*, quizá el SRI hubiera truncado algunas de las entradas para reducirlas a un formato de raíz común, pero para continuar con la explicación resulta más sencillo e ilustrativo dejar los términos en su formato normal. En cuanto a las palabrasvacías, hemos supuesto que el SRI elimina los determinantes, preposicionesy verbos ("el", "pasa", "por", etc.), presentes en los distintos documentos.

Para entregar la respuesta a una determinada pregunta se realizan una serie de operaciones. La primera es traducir la pregunta al formato de un vector de términos. Así, si la pregunta fuera **"¿cuál es el caudal del río Danubio?**", su vector de términos sería Q = (1,1,0,0,0,1,0,0,0). El siguiente paso es calcular la similitud existente entre el vector pregunta y los vectoresde los documentos (existen varias funciones matemáticas diseñadas para ello) y ordenar la respuesta en función de los resultados de similitud obtenidos.

Este procedimiento simple ha sido ligeramente modificado. Algunos autores comenzaron a considerar que la ***tf*** (la frecuencia absoluta de aparición de un término en un documento), es un factor que precisa de una corrección, porque la importancia de un término en función de su distribución puede llegar a ser desmesurada (por ejemplo, una frecuencia de **2** es 200% más importante que una frecuencia de 1, y la diferencia aritmética es sólo de una unidad).

Otros autores, como es el caso de Sparck-Jones, apreciaron la capacidad de discriminación de un término frente a otro. Esta importancia o generalidad de un término dentro de la colección ha de ser vista en su conjunto no en un único documento, y se pensó en incentivar la presencia de aquellos términos que aparecen en menos documentos frente a los que aparecen en todos o casi todos, ya que realmente los muy frecuentes discriminan poco o nada a la hora de la representación del contenido de un documento. Para medir este valor de discriminación se propone la medida ***idf*** (frecuencia inversa de documento). Así, para la construcción de la matriz de términos y documentos, se consideran las siguientes definiciones:

- **n** = número de términos distintos en la colección de documentos

- **tf_{ij}** = número de ocurrencias de término t_j en el documento D_i [frecuencia del término o ***tf***]
- **df_j** = número de documentos que contienen el término t_j
- **idf_j** = el log(d/df_j), donde d es el número total de documentos [frecuencia inversa del documento]

El vector para cada documento tiene n componentes y contiene una entrada para cada término distinto en la colección entera de documentos. Los componentes en el vector se fijan con los pesos calculados para cada término en la colección de documentos. A los términos en cada documento automáticamente se le asignan pesos basándose en la frecuencia con que ocurren en la colección entera de documentos y en la aparición de un término en un documento particular.

El peso de un término en un documento aumenta si este aparece más a menudo en un documento y disminuye si aparece más a menudo en todos los demás documentos. El peso para un término en un vector de documento es distinto de cero sólo si el término aparece en el documento. Para una colección de documentos grande que consiste en numerosos documentos pequeños, es probable que los vectores de los documentos contengan ceros principalmente. Por ejemplo, una colección de documentos con 10000 términos distintos genera un vector 10000-dimensional para cada documento. Un documento dado que tenga sólo 100 términos distintostendrá un vector de documento que contendrá 9900 ceros en sus componentes.

El cálculo del factor de peso (d) para un término en un documento se define **como combinación de la frecuencia de término (*tf*), y la frecuencia inversa del documento (*idf*)**. Para calcular el valor de la j- esima entrada del vector que corresponde al documento i, se emplea la ecuación siguiente: $\mathbf{d_{ij}=tf_{ij}\ x\ idf_j}$. El cálculo de las frecuencias inversas de los términos en los documentos y la posterior aplicación de esta fórmula sobre la matriz de nuestro ejemplo, proporcionaría la siguiente matriz de pesos (a la que añadimos una fila con el vector pregunta).

Cálculo de frecuencias inversas

Idf (río) = Log (4/4) = log (1) = 0

Idf (Danubio) = Log (4/2) = log 2 = 0.301

Idf (Viena) = Log (4/1) = log 4 = 0.602

Idf (color) = Log (4/1) = log 4 = 0.602

Idf (azul) = Log (4/1) = log 4 = 0.602

Idf (caudal) = Log (4/3) = log 1.33 = 0.124

Idf (invierno) = Log (4/1) = log 4 = 0.602

Idf (Rhin) = Log (4/1) = log 4 = 0.602

Idf (navegable) = Log (4/1) = log 4 = 0.602

Matriz ***tf-idf.***

	Río	**Danubio**	**Viena**	**color**	**azul**	**Caudal**	**invierno**	**Rhin**	**navegable**
D1	0	0.301	0.602	0.602	0.602	0	0	0	0
D2	0	0	0	0	0	0.124	0.602	0	0
D3	0	0.301	0	0	0	0.124	0	0.602	0
D4	0	0	0	0	0	0.124	0	0	0.602
Q	0	0.301	0	0	0	0.124	0	0	0

Tabla 1.5 Ejemplo de Matriz de términos y documentos en el Espacio Vectorial con los pesos calculados. Fuente: elaboración propia.

Ahora corresponde calcular las similitudes existentes entre los distintos documentos (D1, D2, D3 y D4) y el vector Q de la pregunta. Hay que multiplicar componente a componente de los vectores y sumar los resultados. El modo más sencillo de obtener la similitud es por medio del producto escalar de los vectores (es decir, multiplicando los componentes de cada vector y sumando los resultados).

Cálculo de similitudes

Sim (D1,Q) = 0*0 + 0.301*0.301 + 0.602*0 + 0.602*0 + 0.602*0 + 0*0.124+ 0*0 + 0*0 +0*0 = **0.09**

Sim (D2,Q) = 0*0 + 0*0.301 + 0*0 + 0*0 + 0*0 + 0.124*0.124 + 0.602*0 + 0*0 + 0*0 = **0.01**

Sim (D3,Q) = 0*0 + 0.301*0.301 + 0*0 + 0*0 + 0*0 + 0.124*0.124 + 0*0 + 0.602*0 + 0*0 = **0.10**

Sim (D4,Q) = 0*0 + 0*0.301 + 0*0 + 0*0 + 0*0 + 0.124*0.124 + 0*0 + 0*0 + 0.602*0 = **0.01**

Con estos valores de similitud, se obtiene la siguiente respuesta: {D3, D2, D1, D4}. Podemos observar en este ejercicio un ejemplo de acierto y un ejemplo de fallo de este modelo, ya que el primero de los documentos recuperados sí responde a la pregunta (D3) y al mismo tiempo los demás no responden adecuadamente (realmente la similitud es muy baja). Casos como el presente, justifican la presencia de documentos no

relevantes en la respuesta de los SRI y que este esquema básico de alineamiento haya sufrido muchos cambios.

2 La Recuperación de Información en la world wide web.

La consolidación de la web como plataforma de diseño de los sistemas de información en Internet, propició la creación de los SRI más voluminosos y avanzados que hasta ahora se han desarrollado. Este capítulo presenta con detalle la base que sustentaestos sistemas y la problemática que la misma ofrece.

Recuperar información en la web.

Hu recuerda que "el primer motor de búsqueda desarrollado en la red Internet fue ARCHIE[14], creado en 1990, aunque no fue hasta la creación del primer navegador, *Mosaic*[15], que propició el crecimiento de los documentos y la gestión de información multimedia cuando se expandió el uso de estos sistemas" [HU, 2001].

La web es un nuevo contexto, con una serie de particularidades muy definidas, que precisa de una adaptación del concepto de recuperación de información, bajo estas premisas Delgado Domínguez afirma que "se puede definir el objetivo de la recuperación como la identificación de una o más referencias de páginas web que resulten relevantes para satisfacer una necesidad de información" [DEL, 1998]. En este caso, los SRI que se emplean en la web nos van a devolver referencias a los documentos, en lugar de los propios documentos.

Breve perspectiva histórica de la web.

El nacimiento y crecimiento exponencial de la web es un hecho suficientemente conocido y cuyo alcance ha traspasado los límites de la comunidad científica hasta llegar a todo el entorno social. En agosto de 1991, Paul F. Kunz, físico de la *Universidad de Stanford* leyó una noticia enla que difundía la invención de la *World Wide Web* y contactó con Tim Berners-Lee, becario británico del CERN[16]. Berners-Lee estaba decidido a desarrollar un método eficiente y rápido para intercambiar datos científicos

[14] ARCHIE es una base de datos que contiene información sobre el contenido de servidores FTP Anónimo dispuestos en la red Internet. Permite así localizar en qué servidor se puede encontrar un determinado recurso.

[15] *Mosaic* es en la práctica el primer navegador gráfico, creado por Marc Andreessen en 1993, cuando era un estudiante de 22 años en la *Universidad de Urbana-Champaign* en Illinois.

combinando dos tecnologías ya existentes: el *hipertexto* y el *protocolo de comunicaciones TCP/IP*, implantando un nuevo modelo de acceso a la información en Internet intuitivo e igualitario: la *World Wide Web* (o WWW o *web*).

El objeto que movía a Berners-Lee en su iniciativa era disponer de un sistema de creación y distribución de documentos, que permitiera compartir información desarrollada en diferentes aplicaciones, de forma sencilla y eficiente, entre equipos de investigadores ubicados en distintos lugares geográficos y que cumpliera además los siguientes requisitos:

- Disponer de una interface sólida, es decir, el sistema debería permitir una conexión que al menos asegurara una transferencia de datos consistente.
- Integración de un amplio rango de tecnologías y distintos tipos de documentos.
- Proporcionar una herramienta para los documentos desde cualquier lugar de Internet y por cualquier individuo que este navegandodentro de este almacén, permitiendo accesos simultáneos.

Kunz almacenó la información de su departamento en la Universidad de Stanford en un servidor, con el fin de que otros científicos pudieran acceder a ella a través de Internet gracias al formulario web de la siguiente ilustración. Fue el propio Berners-Lee fue el primero en probarlo.

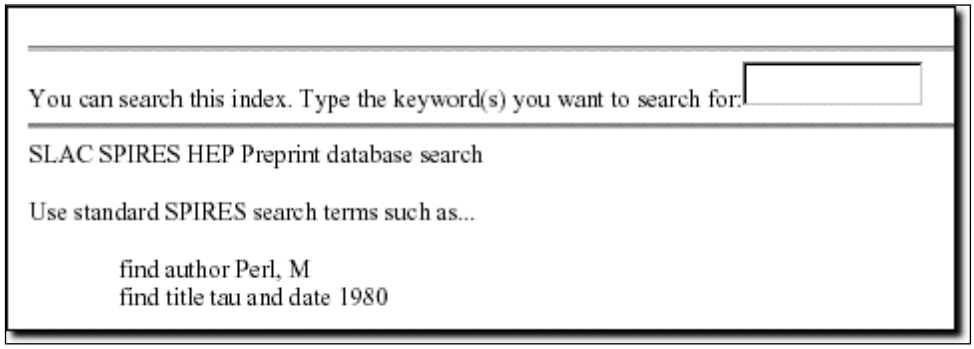

Ilustración 2.1 Sección de la primera página web diseñada por Kunz. Esta página sigue activa en la dirección <http://www.slac.stanford.edu/spires/hep/> de la Universidad de Stanford.

El posterior desarrollo de *Mosaic* aumentó el número de usuarios que accedía a este novedoso sistema, crecimiento que continuó con el desarrollo de nuevos navegadores: N*etscape* e *Internet Explorer*[17]. A

[16] C.E.R.N. son las siglas del "Centro Europeo de Investigación Nuclear" de Ginebra. Actualmente es denominado "European Organization for Nuclear Research" ("Organización Europea de Investigación Nuclear").

principios del año 1995 se formó la organización *Consorcio World Wide Web*[18] (o *W3C*), que está bajo la dirección del fundador de la Web

Métodos de recuperación de información en la web.

Sustancialmente, las técnicas de recuperación de información empleadas en Internet, proceden de las empleadas en los SRI tradicionales, y por ello surgen problemas cuando se realizan operaciones de recuperación de información, en tanto que el entorno de trabajo no es el mismo y las características intrínsecas de los datos almacenados difieren considerablemente.

Al mismo tiempo, en la web surgen nuevos problemas, tales como los populares fenómenos denominados *cloaking*[19], *links farms*[20] y *guest spamming*[21], o los vinculados con el enorme tamaño del índice de estos SRI, que poco a poco llega a alcanzar magnitudes impresionantes, muy difíciles de gestionar adecuadamente con los modelos tradicionales.

Baeza-Yates afirma que hay básicamente tres formas de buscar información en la web: "dos de ellas son bien conocidas y frecuentemente usadas. La primera es hacer uso de los *motores de búsqueda*[22], que indexan una porción de los documentos residentes en la globalidad de la web y que permiten localizar información a través de la formulación de una pregunta. La segunda es usar *directorios*[23], sistemas que clasifican documentos web

[17] Estas aplicaciones desencadenaron un litigio legal antimonopolio sólo comparable al entablado a principios del siglo XX por la extracción del petróleo, entonces en manos de Rockefeller y su compañía 'Standard Oil Company'.

[18] Traducción del término inglés "World Wide Web Consortium". La URL de esta organización es <http://www.w3c.org/>.

[19] Constructores de páginas web que insertan en su descripción términos que nada tienen que ver con el contenido de las mismas ("mp3", "sex", "pokemon", "Microsoft"), todos ellos de uso frecuente por los usuarios de los motores. El objetivo buscado es que su página sea recuperada también. También se incluyen bajo esta denominación las páginas optimizadas para un motor específico.

[20] Páginas donde puedes insertar un enlace a tu página web. Cuanto mayor sea el número de enlaces que consigas para tu página, mayor será tu *Pagerank* y por tanto más posibilidades de que *Google* te coloque en los primeros puestos de una respuesta. *E*ste aumento de la popularidad de tu página es algo artificioso.

[21] Consiste en firmar en multitud de *libros de visitas*; dado que muchos ofrecen la posibilidad de introducir una dirección web, es una manera rápida de conseguir enlaces sin el engorroso trámite de pedírselo al webmaster correspondiente.

[22] Empleamos el término "motor de búsqueda" como traducción de "web search engine". También se usa coloquialmente el término "buscador" como traducción del inglés, aunque su uso suele englobar en algunas ocasiones a los directorios.

[23] Empleamos el término "directorio" como traducción del término "web directory". En términos coloquiales se utiliza también la palabra "índice", aunque esta última palabra puede llevar a confusión en tanto que los motores de búsqueda, al igual que los directorios, hacen uso de índices para almacenar su información.

seleccionados por materia y que nos permiten navegar por sus secciones o buscar en sus índices. La tercera, que no está del todo disponible actualmente, es buscar en la web a través de la *explotación de su estructura hipertextual* (de los enlaces de las páginas web[24])" (Baeza-Yates, 1999).

Centrando el estudio en las primeras formas, resulta conveniente tener en cuenta el cierto grado de confusión existente entre los usuarios de estos sistemas, que a veces no tienen muy claro qué modalidad de sistema están empleando. Muchas veces, los usuarios no distinguen las diferencias que existen entre un directorio (*Yahoo!*, por ejemplo) y un motor de búsqueda (como pueden ser *Alta Vista* o *Lycos)*, ya que las interfaces de consulta de todos estos sistemas resultan muy similares y ninguno explica claramente en su página principal si se trata de un directorio o de un motor de búsqueda. Algunas veces aparece un directorio ofreciendo resultados procedentes de un motor de búsqueda (*Yahoo!* y *Google* tienen un acuerdo para ello[25]), o bien un motor también permite la búsqueda por categorías, como si fuera un directorio (*Microsoft Network*, por ejemplo). Estas situaciones no contribuyen a superar ese grado de confusión.

Los directorios son aplicaciones controladas por humanos que manejan grandes bases de datos con direcciones de páginas, títulos, descripciones, etc. Estas bases de datos son alimentadas cuando sus administradores revisan las direcciones que les son enviadas para luego ir clasificándolas en subdirectorios de forma temática. Los directorios más amplios cuentan con cientos de trabajadores y colaboradores revisando nuevas páginas para ir ingresándolas en sus bases de datos. Los directorios están "organizados en categorías temáticas, que se organizan jerárquicamente en un árbol de materias de información que permite el hojeo de los recursos descendiendo desde los temas más generales a losmás específicos. Las categorías presentan un listado de enlaces a las páginas referenciadas en el buscador. Cada enlace incluye una breve descripción sobre su contenido" [AGU, 2002].

La mayoría de los índices permiten el acceso a los recursos referenciados a través de dos sistemas: navegación a través de la estructura de las categorías temáticas y búsquedas por palabras claves sobre el conjunto de referencias contenidas en el índice. El directorio más grande y famoso es *Yahoo!*, aunque existen otros bastante conocidos: *Dmoz* (un directorio alimentado por miles de colaboradores), *Looksmart*, *Infospace* e *Hispavista*. Con mucha diferencia, el más utilizado es *Yahoo!*. Los directorios son más usados que los motores de búsqueda especialmente cuando "no se conoce exactamente el objetivo de la búsqueda" [MAN,

[24] En algunos textos se usa "hiperenlace" como traducción de "hyperlink".

[25] En la URL <http://es.docs.yahoo.com/info/faq.html#av> se amplía información sobre esta colaboración, que también mantienen otros sistemas.

2002], ya que resulta difícil acertar con los términos de búsqueda adecuados.

Los motores de búsqueda son robustas aplicaciones que manejan también grandes bases de datos de referencias a páginas web recopiladaspor medio de un proceso automático, sin intervención humana. Uno ovarios agentes de búsqueda recorren la web, a partir de una relación de direcciones inicial y recopilan nuevas direcciones generando una serie de etiquetas que permiten su indexación y almacenamiento en la base de datos.Un motor no cuenta con subcategorías como los directorios, sino con avanzados algoritmos de búsqueda que analizan las páginas que tienen ensu base de datos y proporcionan el resultado más adecuado a una búsqueda.También almacenan direcciones que les son remitidas por los usuarios[26]. Entre los motores más populares destacan *Altavista*, *Lycos*, *Alltheweb*, *Hotbot*, *Overture*, *Askjeeves*, *Direct Hit*, *Google*, *Microsoft Network*, *Terra* y *WISEnut*, entre otros.

	Descubrimiento de recursos	Representación del contenido	**Representación de la consulta**	**Presentación de los resultados**
Directorios	Lo realizan personas	Clasificación manual	Implícita (navegación por categorías)	Paginas creadas antes de la consulta. Poco exhaustivos, muy precisos
Motores de búsqueda	Principalmente de forma automática por medio de robots	Indización automática	Explícita (palabras clave, operadores, etc.)	Páginas creadas dinámicamente en cada consulta. Muy exhaustivos, poco precisos

Tabla 2.1 Características de directorios y motores de búsqueda. Fuente: Delgado Domínguez, A. Mecanismos de recuperación de Información en la WWW [En línea]. Palma de Mallorca, Universitat de les Illes Balears, 1998. <http://dmi.uib.es/people/adelaida/tice/modul6/memfin.pdf> [Consulta: 18 de septiembre de 2001]

Delgado Domínguez resume en la tabla 2.1 las características básicas de estos dos métodos de recuperación de información en la web. Es

[26] Aunque algunas fuentes cifran entre el 95% y 97% el número de URL que son rechazadas por estos motores por diversos motivos.

oportuno puntualizar que el razonamiento que lleva a la autora a considerar un directorio más preciso que un motor, se basa, sin duda alguna, en la fiabilidad de la descripción del registro, realizada manualmente de forma detallada y ajustada, entendiendo en este caso *precisión* como *ajuste* o *correspondencia* de la descripción realizada con el contenido de la página referenciada, en lugar de la acepción del mismo término empleada para medir el acierto de una operación de búsqueda[27]. Evidentemente, este nivel de ajuste varía sustancialmente cuando la descripción se ha realizado a través de un proceso automático, como suele ser el caso de los motores de búsqueda.

El tercer método de recuperación enunciado por Baeza-Yates es la *búsqueda por explotación de los enlaces* recogidos en las páginas web, incluyendo los *lenguajes de consulta a la web* y la *búsqueda dinámica*. Estas ideas no se encuentran todavía suficientemente implantadas debido a diversas razones, incluyéndose entre las mismas las limitaciones en la ejecución de las preguntas en estos sistemas y la ausencia de productos comerciales desarrollados (Baeza-Yates, 1999).

Los *lenguajes de consulta a la web*[28] pueden emplearse para localizar todas las páginas web que contengan al menos una imagen y que sean accesibles al menos desde otras tres páginas. Para ser capaz de dar respuesta a esta cuestión se han empleado varios modelos, siendo el más importante un modelo gráfico etiquetado que representa a las páginas web (los nodos) y a los enlaces entre las páginas y un modelo semiestructurado que representa el contenido de las páginas web con un esquema de datos generalmente desconocido y variable con el tiempo, tanto en extensión como en descripción (Baeza-Yates, 1999).

Chang profundiza más en este tipo de lenguajes de recuperación, "estos lenguajes no sólo proporcionan una manera estructural de acceder a los datos almacenados en la base de datos, sino que esconden detalles de la estructura de la base de datos al usuario para simplificar las operaciones de consulta" [CHA, 2001], este aspecto cobra especial importancia en un contexto tan heterogéneo como es la web, donde se pueden encontrar documentos de muy diversa estructuración. Es por ello que estos lenguajes simplifican enormemente la recuperación de información. Los más desarrollados pueden entenderse como extensiones del lenguaje SQL[29] para el contexto de la web empleado en los tradicionales gestores relacionales.

---

[27] Más vinculada a términos como *relevancia* o *pertinencia* del documento recuperado con respecto de la temática objeto de la pregunta.

[28] Traducción literal del término inglés "web query languages".

[29] SQL: Structured Query Language.

Todos estos modelos constituyen adaptaciones o propuestas de desarrollo de sistemas navegacionales para la consulta de hipertextos [CAN, 1990], [NIE, 1990], (Baeza-Yates, 1999), combinando la estructura de la red formada por los documentos y por sus contenidos. Al igual que sucedió en el entorno de los hipertextos, estos modelos resultan difíciles de implantar cuando se trata de gestionar inmensas cantidades de datos, como ocurre en la web.

La búsqueda dinámica es "equivalente a la búsqueda secuencial en textos" (Baeza-Yates, 1999). La idea es usar una búsqueda online para descubrir información relevante siguiendo los enlaces de las páginas recuperadas. La principal ventaja de este modelo es que se traslada la búsqueda a la propia estructura de la web, no teniendo que realizarse estas operaciones en los documentos que se encuentran almacenados en los índices de un motor de búsqueda. El problema de esta idea es su lentitud, loque propicia que se aplique sólo en pequeños y dinámicos subconjuntos de la web. Chang, dentro los SRI en web basados en la recuperación de información por medio de palabras clave, identifica cuatro tipos: motores de búsqueda, directorios, *metabuscadores*[30] y técnicas de *filtrado de información*[31] [CHA, 2001].

Los *metabuscadores* son sistemas desarrollados para mitigar el problema de tener que acceder a varios motores de búsqueda con el fin de recuperar una información más completa sobre un tema, siendo estos mismos sistemas los que se encargan de efectuarlos por el usuario. Un metabuscador colecciona las respuestas recibidas y las unifica, "la principalventaja de los metabuscadores es su capacidad de combinar los resultados de muchas fuentes y el hecho de que el usuario pueda acceder a varias fuentes de forma simultánea a través de una simple interfaz de usuario" (Baeza-Yates, 1999). Estos sistemas no almacenan direcciones y descripciones de páginas en su base de datos, "en lugar de eso contienen registros de motores de búsqueda e información sobre ellos. Envían la petición del usuario a todos los motores de búsqueda (basados en directorios y crawlers[32]) que tienen registrados y obtienen los resultados que les devuelven.

[30] "Metabuscador" es la traducción más aceptada del término "meta-search engine".

[31] Traducción literal del término inglés "information filtering".

[32] Este término se refiere al robot que recopila páginas web para el índice de los motores de búsqueda.

Algunos más sofisticados detectan las URL duplicadas provenientes de varios motores de búsqueda y eliminan la redundancia" [AGU, 2002]. , es decir solo presentan una al usuario.

Por muy grande y exhaustiva que pudiera llegar a ser la base de datos de un motor de búsqueda o de un directorio, nunca va a cubrir un porcentaje muy elevado del total de la web, "incluso si tienes un motor de búsqueda favorito, o incluso varios de ellos, para asegurarte de que tu búsqueda sobre una materia es suficientemente exhaustiva necesitarás haceruso de varios de ellos" [BRA, 2000].

Estos sistemas se diferencian en la manera en que llevan a cabo el alineamiento de los resultados[33] en el conjunto unificado[34], y cómo de bien traducen estos sistemas la pregunta formulada por el usuario a los lenguajes específicos de interrogación de sus sistemas fuentes, ya que el lenguaje común a todos será más o menos reducido. Algunos metabuscadores se instalan como cliente en entorno local (*Webcompass* o *Copernic*, por ejemplo), o bien se consultan en línea (*Buscopio*, por ejemplo). Otra diferencia sustancial existente entre estos sistemas es la presentación de los resultados, "los llegan a clasificar en dos tipos, los multi buscadores y los meta buscadores: los multi buscadores ejecutan la consulta contra varios motores de forma simultánea y presentan los resultados sin más organización que la derivada de la velocidad de respuesta de cada motor (unejemplo es *All4One* que busca en una gran cantidad de motores de búsqueda y directorios); los meta buscadores funcionan de manera similar alos multi buscadores pero, a diferencia de éstos, eliminan las referencias duplicadas, agrupan los resultados y generan nuevos valores de *pertinencia* para ordenarlos (algunos ejemplos son *MetaCrawler*, *Cyber411* y *digisearch)*" [AGU, 2002]. Otros metabuscadores muestran los resultados en diferentes ventanas, correspondiendo cada una de ellas a una fuente distinta (*Oneseek* o *Proteus*, por ejemplo).

Generalmente, el resultado obtenido con un metabuscador no es todo el conjunto de páginas sobre una materia que se almacenan en las fuentes del metabuscador, ya que suele limitarse el número de documentos recuperados de cada fuente. Aún así, "el resultado de un metabuscador suele ser más relevante en su conjunto" (Baeza-Yates, 1999). Puede sorprender esta limitación, pero no debemos olvidar uno de los elementos más considerados en las evaluaciones de los SRI, el *tiempo de respuesta delsistema* [LAN, 1993]. Si un metabuscador devolviera todas las referencias de todos los motores y directorios fuentes en relación con la materia objeto de una búsqueda, el tiempo de respuesta del sistema alcanzaría valores que

[33] Aunque el Diccionario de la R.A.E. admite el uso del vocablo "ranking", preferimos emplear "alineamiento", al tratarse el anterior de un anglicismo.

[34] En algunos casos este proceso no se realiza [BAE, 1999].

seguramente alejarían a los usuarios del metabuscador por excesivo. Por ello resulta necesario establecer un número límite de documentos recuperados por motor, con el fin de que el tiempo de respuesta, que de por sí, ya sería siempre mayor que el precisado por un único motor, no aumente excesivamente.

Actualmente, se encuentra en desarrollo una nueva generación de metabuscadores, destacando *Inquirus* como prototipo que emplea "búsqueda de términos en contexto y análisis de páginas para una más eficiente y mejor búsqueda en la web, también permite el uso de los operadores booleanos" [INQ, 2002]. Este sistema muestra los resultados progresivamente, es decir a medida que van llegando (tras analizarlos), con lo que el usuario apenas tiene tiempo de espera y las referencias que le entrega el metabuscador son siempre correctas (es decir no va a entregar una dirección de página inexistente o páginas que hubieran cambiado su contenido desde la indización) (Baeza-Yates, 1999).

Las técnicas de filtrado de la información son más un complemento de los motores que un modelo alternativo. El concepto de "filtrado" tiene que ver con la decisión de considerar (a priori) si un documento es relevante o no, eliminándolo del índice en caso contrario. Estas técnicas "sebasan en una combinación de sistemas de autoaprendizaje y SRI, y han sidoempleadas en la construcción de motores de búsqueda especializados" [CHA, 2001]. El filtrado de términos mejora la calidad del índice, rechazando términos de escaso o nulo valor de discriminación y aumenta la velocidad de la recuperación de información, aligerando las dimensiones del índice. Según el esquema de la ilustración 2.1, el agente (o los agentes) encargado/os de recopilar la información, someten las páginas recopiladasal filtrado. Si son aceptadas se almacenarán en el índice del motor, mientrasque las rechazadas son descartadas.

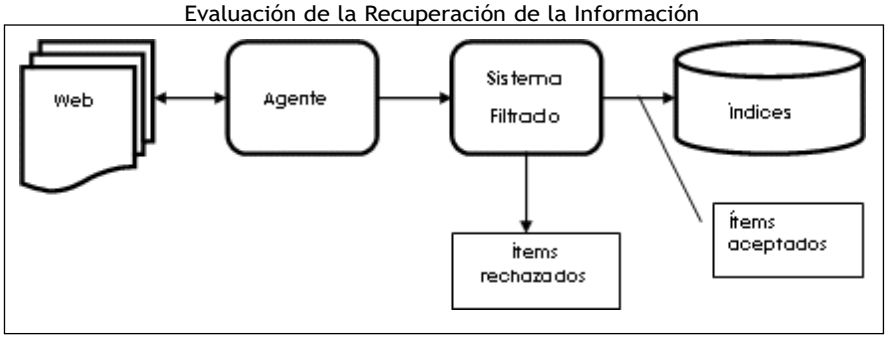

Ilustración 2.1 Proceso de construcción de un motor de búsqueda específico a partir de un filtrado de documentos. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001.

Hu habla de seis tipos de tecnologías diferentes empleadas en la búsqueda de documentos en la web [HU, 2001]:

1. exploración de la estructura hipertextual
2. recuperación de la información
3. metabuscadores
4. lenguajes de consulta basados en SQL
5. buscadores multimedia basados en el contexto
6. otros

Hu considera que "los enlaces establecidos entre las páginas web pueden resultar de tremenda utilidad como fuentes de información para los indicadores" [HU, 2001]. El hecho de que el autor de una página web introduzca un enlace en la misma, representa, implícitamente, un respaldo para la página enlazada[35]. La exploración de los enlaces[36] insertados en una página web y la exploración de los enlaces que apuntan hacia esa misma página, ha respaldado la creación de una nueva familia de motores de búsqueda, de la que *Google*[37] es su mejor exponente [BRI, 1998]. *Google*

[35] Igual que en las técnicas de análisis de citas, si un artículo es citado por los autores de otros trabajos, este primer artículo se dice que "aumenta su impacto".

[36] No confundir "exploración de los enlaces" con la "explotación de la estructura hipertextual" que mencionaba Baeza-Yates. El propio autor incluye a *Google* y a su algoritmo de alineamiento *Page Rank* dentro del campo de los motores de búsqueda y los SRI basados en esa explotación de enlaces son una alternativa a los motores. Otros autores emplean el término "hyperlink spreading" ("extensión de enlaces") para referirse al método que emplea el motor de búsqueda Google.

[37] *Google* es un juego de palabras con el término "googol", acuñado por Milton Sirotta, sobrino del matemático norteamericano Edward Kasner, para referirse al número representado por un 1 seguido de 100 ceros. El uso del término por parte de *Google* refleja la misión de la compañía de organizar la inmensa cantidad de información disponible en la web y en el mundo. Fuente: *Todo acerca de Google* [En línea] Mountain View, CA: Google, 2001. <http://www.google.com/intl/es/profile.html>

hace uso de la conectividad de la web para calcular un grado de calidad de cada página, esta graduación se denomina *PageRank* (coincide con el nombre del algoritmo de alineamiento empleado por este motor) y utiliza esta propia capacidad de conexión para mejorar los resultados de búsqueda.

Hu cita en segundo, tercer y cuarto lugar, a los directorios, motores de búsqueda, metabuscadores y lenguajes de consulta para la web, sistemas todos presentados anteriormente. En quinto lugar menciona la búsqueda de documentos multimedia, área en plena expansión, máxime cuando cada vez es mayor el número de documentos de esta naturaleza en la web y el número de usuarios que los demandan (gracias a las mejores conexiones a Internet). Para este autor, este desarrollo "está considerado uno de losmayores desafíos en el campo de la recuperación de información" [HU, 2001]. El último grupo citado por Hu engloba a los sistemas de recuperación con interface basada en procesamiento de lenguaje natural y los aún incipientes desarrollos de sistemas de recuperación de documentos en formato XML.

Los motores de búsqueda: paradigma de la recuperación de información en Internet.

De la totalidad de los SRI que se han desarrollado en la web, los motores de búsqueda son los que más se incardinan con su naturaleza dinámica, siendo sistemas de evolución paralela al crecimiento de la web y al aumento del número de usuarios. Constituyen además uno de los desarrollos más consolidados de las técnicas de *Indización Automática* (Salton & McGill, 1983) [GIL, 1999] y, al mismo tiempo, son los sistemas más sensibles a toda la amplia serie de situaciones peculiares que se presentan en la red y que no tenían lugar en los tradicionales SRI.

Independientemente de su método de rastreo y posteriores criterios y algoritmos usados en el alineamiento de los documentos, todos los motores parten de una situación inicial parecida: una lista de direcciones que sirve de punto de partida para el robot (o robots) del motor. Esta similitud inicial propicia, ineludiblemente, posteriores comparaciones del resultado obtenido, es decir, de la porción de web indexada y de la calidad de esta indexación. Otro factor favorecedor de estas comparaciones es el ocultismo de los métodos seguidos por cada motor, lo que nos lleva, al igual que ocurre en el caso anterior, a comparar el resultado obtenido con el fin de poder apreciar cuál de esos sistemas es de uso más recomendable.

Si se asume que de lo completa, representativa y actualizada que sea la colección de un motor de búsqueda, depende su calidad; en un directorio, en cambio, esta calidad depende de la capacidad de los indicadores y en su

[Consulta: 21 de enero de 2002].

número, motivos ambos más relacionados con capacidades presupuestarias que con prestaciones tecnológicas. En cambio, los motores representan un claro ejemplo de la aplicación de las técnicas de recuperación de información a la resolución de un reto, tan antiguo como moderno, en el campo de la Información y la Documentación: disponer en un índice las referencias a la mayor parte de los documentos existentes.

Funcionamiento de un motor de búsqueda.

El funcionamiento general de un motor de búsqueda se estudiadesde dos perspectivas complementarias: la *recopilación* y la *recuperación de información*. Un motor compila automáticamente las direcciones de las páginas formarán parte de su índice tras indizarlas. Una vez estén estos registros depositados en la base de datos del motor, los usuarios buscarán en su índice por medio de una interface de consulta (más o menos avanzada en función del grado de desarrollo del motor).

El módulo que realiza esta recopilación es conocido comúnmente

como *robot*[38], es un programa que "rastrea la estructura hipertextual de la web, recogiendo información sobre las páginas que encuentra. Esa información se indiza y se introduce en una base de datos que será explorada posteriormente utilizando un motor de búsqueda" [DEL, 1998]. Estos robots recopilan varios millones de páginas por día, y actualizan la información depositada en los índices en períodos de tiempo muy pequeños (si se considera la extensión del espacio al que nos estamos refiriendo). Parten generalmente de una lista inicial de direcciones de sitios web, que son visitados y a partir de los cuales cada robot rastrea a su manera la web, de ahí que la información almacenada en cada base de datos de cada motor sea diferente. A diferencia de Delgado Domínguez, Baeza-Yates distingue en un robot las funciones de análisis o rastreo ("crawling") de las deindexación ("indexing"), con lo cual habla de dos módulos independientes, el "crawler" o robot y el indexador (Baeza-Yates, 1999).

Arquitectura de un motor de búsqueda.

La mayoría de los motores usan una arquitectura de tipo robot-indexador centralizada (ver Ilustración 2.2). A pesar de lo que puede inducir

su nombre[39], el robot no se mueve por la red, ni se ejecuta sobre las máquinas remotas cuyas páginas consulta. El robot funciona sobre el sistema local del motor de búsqueda y envía una serie de peticiones a los servidores remotos que alojan las páginas a analizar. El índice también se gestiona localmente. Esta arquitectura clásica es la que implementa, entre otros, el motor *Alta Vista*, "precisando para ello, en 1998, de 20 ordenadores multiprocesadores, todos con más de 130 Gb de memoriaRAM y sobre 500 Gb de espacio en disco; sólo el módulo de interrogación del índice consume más del 75% de estos recursos" (Baeza-Yates, 1999).

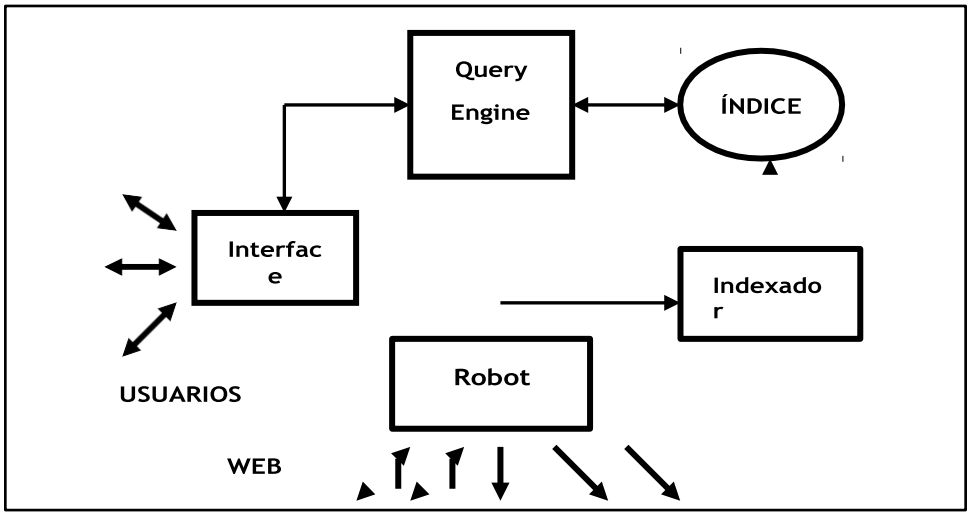

Ilustración 2.2 Arquitectura simple de un motor de búsqueda. Fuente: o a partir de un filtrado de documentos. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p.

Este modelo presenta algunos problemas para gestionar adecuadamente en el entorno local la ingente cantidad de datos:

- La actualización de los índices es complicada y lenta.
- No sigue el ritmo de crecimiento de la web, indexando nuevos documentos en un nivel menor.
- El trasiego de páginas por la red consume un gran ancho de banda y produce una sobrecarga de tráfico [DEL, 1998].
- Suelen ignorarse los contenidos dinámicos de la red, creación de páginas de consulta, ficheros en otros formatos, etc.

Estos problemas propician que uno de los campos de estudio más recientes sea el desarrollo de alternativas a este modelo de arquitectura simple. Baeza-Yates destaca la arquitectura del sistema *Harvest ("cosecha") co*mo la más importante de todas. Este sistema es un paquete integrado de herramientas gratuitas para recoger, extraer, organizar, buscar, y duplicar información relevante en Internet desarrollado en la *Universidad de Colorado* [BOW, 1994].

Harvest hace uso de una arquitectura distribuida para recopilar y distribuir los datos, que es más eficiente que la arquitectura centralizada. El principal inconveniente que presenta es la necesidad de contar con varios servidores para implementarla.

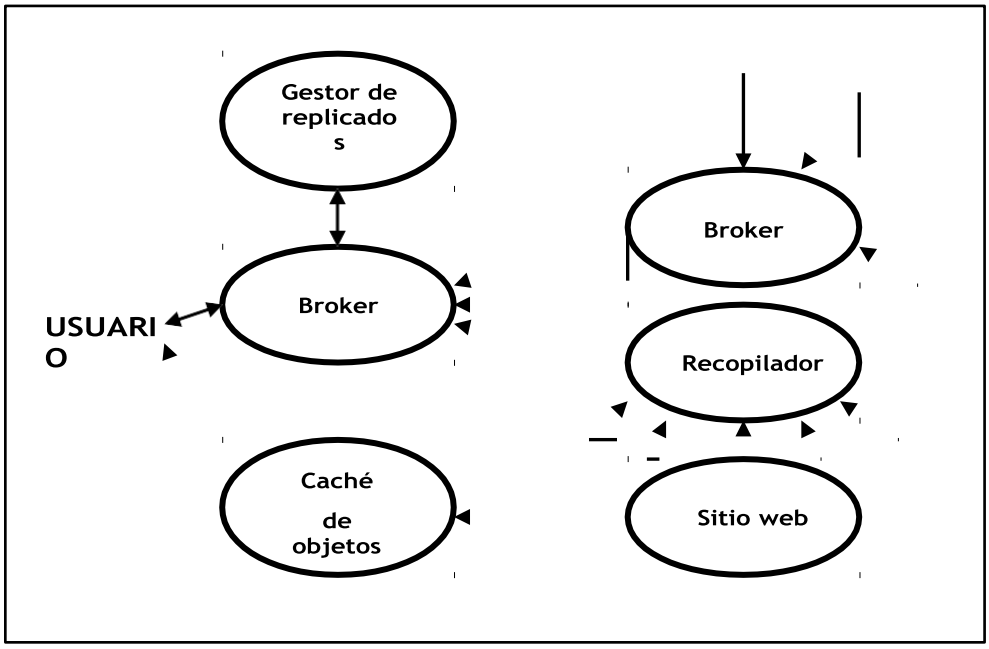

Ilustración 2.3 Arquitectura Harvest. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p.

En esta arquitectura distribuida, los servidores web reciben las peticiones de distintos robots (analizadores) de forma simultánea, aumentándose así la capacidad de carga de nuevas páginas del motor.

Esta arquitectura solventa el problema de la carga de tráfico en las conexiones con el motor, porque aumenta la velocidad de conexión con los robots en tanto que estos descartan gran cantidad de contenidos de las páginas que analizan y no las transfieren al entorno local, aliviando mucho la carga de tráfico. En último lugar, la información se recopila de forma independiente por cada robot, sin tener que realizar una gestión sincronizada.

Harvest tiene dos componentes principales: el *recopilador* y los *brokers*. El primero colecciona páginas y extrae de ellas a información necesaria para crear el índice del motor. El *broker* es el encargado de proporcionar el mecanismo de indexación de las páginas recopiladas y la interface de consulta. Al mismo tiempo, los *brokers* son los servidores de búsquedas, recuperan información desde uno o varios *recopiladores* o desde uno o varios *brokers*, actualizando constantemente sus índices.

Interface de usuario.

El estudio de la interface debe abordarse bajo dos perspectivas: la interface que el sistema dispone para que el usuario exprese sus necesidades de información (Chang la denomina "interface de consulta" [CHA, 2001]) y la interface de respuesta que dispone el sistema para mostrar al usuario el resultado de su operación de búsqueda (Baeza-Yates, 1999). No todos los sistemas poseen iguales prestaciones en la recuperación de información. La más simple interface de usuario es la clásica caja dentro de un formulario web (ver Ilustración 2.4). En esta caja, el usuario inserta su ecuación de búsqueda, pudiendo, a veces, introducir alguna restricción a la búsqueda, como por ejemplo: el idioma de las páginas a recuperar, el tipo de objeto, si se desea emplear la búsqueda por frase literal o "búsqueda exacta", etc.

Ilustración 2.4 Formulario de búsqueda simple del motor All the Web. Fuente: <http://www.alltheweb.com>

Aunque generalmente los motores de búsqueda suelen disponer de una interface de búsqueda avanzada (como la que se muestra en la Ilustración 2.5), en la cual el usuario puede incorporar a su ecuación de búsqueda una serie de parámetros adicionales, tales como: uso de operadores booleanos, búsqueda por frase literal, búsqueda aplicando operadores de adyacencia, búsqueda por términos opcionales (Baeza-Yates los denomina "invitados", son esos términos que podrían aparecer o no en un documento objeto de una consulta), restricciones geográficas, restricciones por tipo de dominio, restricciones por idioma, etc.

Algunos sistemas permiten refinar la búsqueda, es decir, especificar más la pregunta sobre el conjunto de documentos recuperado inicialmente e incluso algunos motores permiten restringir el alcance la operación de búsqueda a alguna de las partes de los documentos contenidos en susíndices (el caso más común es el título o el texto).

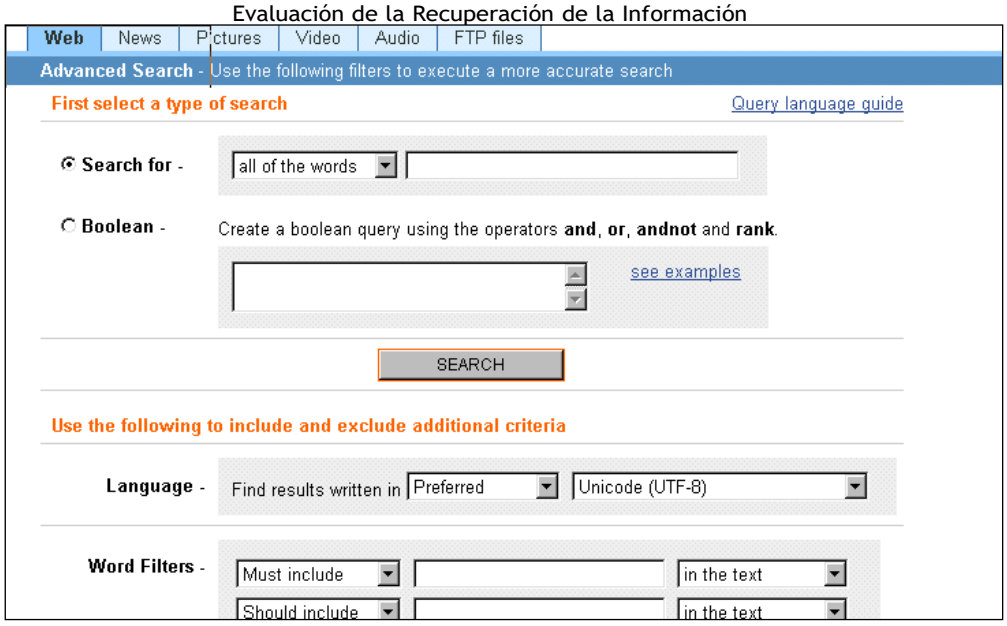

Ilustración 2.5 Sección del formulario de búsqueda avanzada del motor All the Web. Fuente: <http://www.alltheweb.com/advanced>

Son muchos los criterios por los que se pueden identificar diferencias entre las posibilidades de recuperación de información ofrecidaspor cada motor.

La interface de usuario para la realización de las consultas es una de las más empleadas y constituye uno de los parámetros más empleados en los artículos y páginas web dedicadas a la evaluación de las prestaciones decada motor [WIN, 1995], [DAV, 1996], [SLO, 1996], [ZOR, 1996] y [WES, 2001].

Otra diferencia, más interna y no tan explícita como las anteriores, es cómo un sistema interpreta una relación de varios términos como expresión de consulta sin operadores entre ellos (por ejemplo: "Historia Región Murcia"), es decir, cómo descompone la expresión y construye la ecuación de búsqueda. En este punto existen también diferencias entre los sistemas, unos realizan una búsqueda por proximidad de las palabras de la expresión (es el caso de *Overture*), otros motores recuperarán documentos donde aparezcan todas las palabras (Google) y otros recuperarán documentos donde al menos aparezca una de las tres palabras de la ecuación (Alta Vista). Chang identifica cinco tipos de búsqueda [CHA, 2001]:

1. Término simple
2. Términos múltiples
3. Basadas en el contexto

4. Lenguaje natural
5. Correspondencia de patrones

La búsqueda por término simple tiene como objeto devolver una colección de documentos donde al menos se encuentre una ocurrencia de ese término. Algunos sistemas permiten restringir esa búsqueda a un campo determinado (búsqueda por referencia cualificada).

La búsqueda por términos múltiples permite diversas combinaciones basadas en el *Álgebra de Boole*: intersección de los subconjuntos correspondientes a cada término, unión de estos subconjuntos o exclusión de un subconjunto de otro. Algunos sistemas permiten la combinación de los operadores para construir expresiones booleanas complejas.

Las búsquedas basadas en el contexto usan los operadores de proximidad, es decir, localizan documentos donde los términos integrantes de la ecuación de búsqueda se encuentren situados en la misma frase o en el mismo campo (además de, por supuesto, el mismo documento).

El caso más cercano de proximidad es la adyacencia[40] (caso extremo que se produce cuando los términos están escritos en un orden determinado y uno a continuación del otro). Algunos motores permiten la búsqueda en lenguaje natural, que puede resultar especialmente interesante para aquellos usuarios no experimentados en el uso de un motor específico o en el empleo de los operadores booleanos o basados en el contexto.

Estos sistemas interpretan cuestiones del estilo de "¿qué jugador de fútbol es el máximo goleador de la Copa de Europa?" o "¿qué ciudad es la capital de Angola?", devolviendo como resultados un conjunto de documentos que han considerado adecuados con la temática de la pregunta efectuada, tras haber sometido a esta expresión a un procedimiento deanálisis del texto, interpretando la necesidad informativa (*Alta Vista* y *Northern Light* implementan esta modalidad). Un caso extremo de procesamiento de las expresiones en lenguaje natural es el motor *Askjeeves,*que llega a simular una "entrevista" con el usuario ya que, tras recibir la cuestión, la interpreta y extrae de su base de conocimientos una serie de cuestiones que traslada al usuario para refinar su exploración en la base de datos y ajustar mejor la respuesta. Por último, algunos sistemas devuelven

los documentos por correspondencia con un patrón de caracteres

introducido en la interface de consulta[41].

Los índices de los motores.

El índice "es el corazón de un motor de búsqueda" [CHA, 2001], suele consistir en una lista de palabras con cierto valor de discriminación asociadas a sus correspondientes documentos, que en este caso son las descripciones de los contenidos de las URL recopiladas. La mayor parte de los motores de búsqueda emplean como estructura de datos un *fichero inverso* (Baeza-Yates, 1992), (Tramullas Saz, 1997), (Rijsbergen, 1979), [CHA, 2001], [DEL, 2001], basado en la idea general que se muestra en la ilustración siguiente.

Document	Text
1	Pease porridge hot, pease porridge cold,
2	Pease porridge in the pot,
3	Nine days old.
4	Some like it hot, some like it cold,
5	Some like it in the pot,
6	Nine days old.

(a) Example text; each line is one document

Number	Term	Text
1	cold	1,4
2	days	3,6
3	hot	1,4
4	in	2,5
5	it	4,5
6	like	4,5
7	nine	3,6
8	old	3,6
9	pease	1,2
10	porridge	1,2
11	pot	2,5
12	some	4,5
13	the	2,5

(b) Inverted file for text of (a)

Ilustración 2.6 Ejemplo de estructura de un fichero inverso simple. Fuente: Rijsbergen, C.J. Information Retrieval. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 21 de octubre de 2001]

En la práctica el fichero inverso se convierte en una enorme estructura de datos con serios problemas de gestión. Los distintos motores de búsqueda se sirven de distintos esquemas para definir estas estructurasde datos. Se cuenta con un parámetro, denominado *granularidad*, que distingue "la exactitud con la que el índice identifica la localización de una palabra clave; en general, los índices pueden clasificarse con base en este parámetro" [CHA, 2001]. Esta clasificación distingue tres niveles:

Granularidad consistente	Capaz de identificar un conjunto de documentos a partir de una palabra clave
Granularidad media	Capaz de identificar un documento específico a partir de una palabra clave
Granularidad fina	Capaz de identificar la localización de una frase o de una palabra en un documento a partir de una palabra clave

Tabla 2.2 Clasificación de ficheros inversos a partir de la *granularidad* de su índice. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001.

El índice emplea un conjunto de punteros que apuntan a una tabla que contiene las URL en las que aparece una palabra clave. La manera de ordenar estos punteros depende de un mecanismo interno, basado generalmente en su frecuencia o en su peso en el documento. El enorme tamaño de la colección de URL recopiladas por los motores obliga a buscar formas de simplificar al máximo el tamaño de estos índices. En la Tabla 2.3se presentan algunas de las diversas técnicas empleadas.

Conversión de texto a minúsculas	Se convierten todas las palabras a caracteres en minúscula, reduciendo así el número de entradas para un mismo término (Puerto – puerto)
Stemming	Aislamiento de la base de la palabra (por ejemplo: "comprensión" y "comprensivo" se reducirían a "compren").
Supresión de las palabras vacías	Se suprimen del índice todas las palabras porlas que no tiene sentido recuperar información (artículos, preposiciones, adjetivos, interjecciones, por ejemplo)
Comprensión de textos	Técnicas de compactación del tamaño del fichero

Tabla 2.2 Técnicas usadas para reducir el tamaño de los índices de un motor de búsqueda. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell: Kluwer Academic Publishers, 2001.

Las tres primeras reducen hasta un 30% del tamaño y la cuarta llega, a veces, hasta un 90%. Es evidente la tendencia a disminuir el tamaño del índice, ya que cuando las búsquedas constan de varios términos y uno de ellos es muy frecuente, el motor puede tardar varios segundos en responder, hecho no muy bien considerado por muchos usuarios. Los índices con *granularidad* más consistente implican menor tamaño y menos punteros, lo que favorece una simplificación de la estructura de datos. Baeza-Yates expone como ejemplo la idea que Glimpse emplea en *Harvest*: "las preguntas del usuario son resueltas por medio de ficheros inversos, que proporcionan una lista de bloques lógicos, leídos secuencialmente porque su tamaño es menor" (Baeza-Yates, 1999).

Tipos de robots.

Junto a los robots de carácter general dedicados a descubrir recursos, existen otras modalidades específicas de estos sistemas:

- "*Knowbots*: Programados para localizar referencias hipertexto dirigidas hacia un documento, servidor, etc., en particular. Permiten evaluar el impacto de las distintas aportaciones que engrosan las distintas áreas de conocimiento presentes en la Red.
- *Wanderers* (vagabundos): Encargados de realizar estadísticas: crecimiento de la Red, número de servidores conectados, etc.
- *Worms* (gusanos): Encargados de la duplicación de directorios FTP, para incrementar su utilidad a un número mayor de usuarios
- *WebAnts* (hormigas): Conjunto de robots físicamente alejados que cooperan para la consecución de distintos objetivos, como por ejemplo para llevar a cabo una indización distribuida. [DEL, 1998]"

Funcionamiento de los robots.

Se ha comentado anteriormente que, habitualmente, el robot inicia su rastreo a partir de un conjunto de URL muy populares o enviadas explícitamente por los administradores de sitios web, y se siguen los enlaces contenidos en esa relación inicial de páginas evitando repeticiones. El recorrido puede ser de dos modos:

- *breadth-first* (cobertura amplia pero no profunda) y
- *depth-first* (cobertura vertical profunda) (Baeza-Yates, 1999).

La extensión de la web genera problemas al actualizar los índices, ya que transcurre un considerable tiempo entre dos análisis del mismo recurso, intervalo muy variable según el motor. Baeza-Yates esboza una analogía entre el índice y las estrellas del cielo: "lo que vemos en un índice jamás ha existido, es como la luz que ha viajado a lo largo de mucho tiempo hasta llegar a nuestro ojos. Cada página se indexó en un momento distinto del tiempo, pero al ir a ella obtenemos el contenido actual" (Baeza-Yates, 1999). Seguramente por esta razón, algunos motores muestran la fecha de indización de la página. Baeza-Yates estima que alrededor del 9% de los enlaces almacenados son inválidos y Nottes cifra esta cantidad entre el 1% el 13% [NOT, 2000]. Estas cifras son objeto de revisiones frecuentes.Aguilar González resume en una tabla algunas de las principalescaracterísticas de rastreo y los motores que las implementan:

Tipo de rastreo	No	Sí

Rastreo profundo	Excite	El resto
Soporte de marcos	Excite, FAST	El resto
Mapas de imágenes	Excite, FAST	Alta Vista, Northern Light
Robots.txt	Ninguno	Todos
Metadatos	Excite	El resto
Rastreo por popularidad	Ninguno	Todos
Inclusión pagada	Excite, Google[42]	Alta Vista, Inktomi, FAST

Tabla 2.4 Características de rastreo de los robots de los principales motores de búsqueda. Fuente: Aguilar González, R. Monografía sobre motores de búsqueda [En línea]. Yahoo! Geocities, 2002. <http://www.geocities.com/motoresdebusqueda/crawlers.html> [Consulta: 3 de abril de 2002]

Indización de las páginas.

A medida que los robots recopilan páginas, la información contenida en las mismas debe ser indizada, Delgado Domínguez opina que "existen dos estrategias básicas, no mutuamente excluyentes, para realizar este proceso: usar información que provee el creador o editor del documento, o extraerla directamente del documento" [DEL, 1998]. El volumen de información que gestiona un robot obliga a que el motor de búsquedaimplemente algún tipo de *indización automática* [GIL, 1999]. En la práctica, los principales motores emplean ambas estrategias para disponer de una completa descripción del contenido de la página analizada. Aguilar González enumera una serie de criterios utilizados para esta descripción: "el título del documento, los metadatos, el número de veces que se repite una palabra en un documento, algoritmos para valorar el peso del documento, etc." [AGU, 2002].

La mayoría de los motores calculan el número de veces que se repiten las palabras claves en el cuerpo de una página, después escudriñan estas palabras en el nombre del dominio o en la URL, posteriormente en el título de la página, en el encabezado y en los metadatos. El orden en que se busca en cada uno de estos elementos varía en función del motor (cada uno usa sus propios algoritmos con criterios diferentes). Si el motor encuentra las palabras claves en todos estos criterios, entonces posee una razón para asignar un peso mayor al documento. Otra metodología se basa en elnúmero de enlaces que la misma reciba o proporcione.

Aguilar González indica que la primera propuesta en esta línea es de Attardi, de la *Universidad de Pisa*, implementada en el motor *Arianna* y que ha servido de base para el desarrollo de motores que analizan los enlaces (como *Google, WISEnut o Kartoo,* entre otros) [AGU, 2002]. Un ejemplo representativo del comportamiento de un motor clásico a la hora deindizar las páginas web es el motor *Alta Vista:*

- Da prioridad alta a las palabras del título y a las palabras que están localizadas en el comienzo de la página.
- Asigna mayor peso a una palabra en un documento según su frecuencia absoluta.
- El mejor tamaño para una página está entre 4 y 8k. Considera las páginas largas como valiosas en contenido, cuando no están afectadas de "spamming".
- Indexa las palabras claves y la descripción de los metadatos. Si no se tienen metadatos en la página, indexa las primeras 30 o 40 palabras de la página y las toma como descripción.
- Confiere una mayor prioridad a palabras ubicadas en los metadatos o a las palabras con las cuales se registran las páginas, pero no son tan relevantes como el título y el contenido.
- Es sensible a las palabras claves mayúsculas y minúsculas.
- Puede indexar un sitio que contiene marcos. Pero se debe asegurar que todas las páginas enlacen a la página principal.

Google es el mejor ejemplo de uso extensivo de los enlaces como base para mostrar los documentos a los usuarios de un motor. En Google, la indización la realizan dos módulos: el *indexador* y el *clasificador*. El

primero lee las páginas procedentes del *storeserver*[43], descomprime los documentos y selecciona los términos incluidos en los mismos. Cada documento se convierte así en un conjunto de palabras (o '*hits*'), donde se graba la palabra y su posición en el documento, una aproximación de su fuente de texto y otra serie de detalles, por medio del *clasificador*.

El *indexador* analiza también los enlaces incluidos en cada página web, información necesaria para calcular el alineamiento de las páginas a la hora de la recuperación de información [BRI, 1998]. En la siguiente tabla se presentan resumidas algunas de las principales características de la indización y los motores que las implementan.

Características de la indización	**No**	**Si**
Texto completo		Todos
Supresión palabras vacías	FAST, Northern Light	AltaVista, Excite, Inktomi, Google
Meta Descripción	Google, Northern Light	El resto
Meta palabras claves	Excite, FAST, Google, Northern Light	El resto
Texto alternativo	Excite, FAST, Inktomi, Northern Light	AltaVista, Google

Tabla 2.5 Características de la indización realizada por los principales motores de búsqueda. Fuente: Aguilar González, R. Monografía sobre motores de búsqueda [En línea]. Yahoo! Geocities, 2002. <http://www.geocities.com/motoresdebusqueda/crawlers.html> [Consulta: 3 de abril de 2002]

Alineado de los documentos (ranking).

El alineado constituye uno, sino el que más, de los procesos críticos a la hora de valorar la efectividad de un motor de búsqueda, ya que se trata del orden en el que el motor presenta los resultados a sus usuarios, quienes, como es lógico esperan encontrar los documentos más relevantes con sus necesidades situados entre los primeros. El motor debe ordenar el conjunto de documentos constituyente de la respuesta en función de la *relevancia* de estos documentos con el tema de la pregunta realizada.

En función del buen funcionamiento de su algoritmo de alineamiento, el motor será mejor o peor valorado por los usuarios del mismo. Si un motor no discrimina su respuesta en función de la *relevancia* con la temática objeto de la pregunta, el usuario encontrará documentos muy relevantes mezclados con otros menos relevantes e incluso con

muchos nada relevantes, lo que le obligará a consultar un gran número de los documentos devueltos por el motor, teniendo que visitar muchas pantallas y perdiendo, en consecuencia, un cuantioso tiempo.

En esta situación, el usuario seguramente terminará por no recurrir a este motor de búsqueda. Si por el contrario, el motor discriminara ese grado de relación, el usuario encontrará, entre los primeros documentos recuperados, los más relevantes con la temática de la pregunta, por lo que aumentará su grado de satisfacción con el motor y continuará utilizándolo.

Tradicionalmente este procedimiento ha sido uno de los secretos mejor guardados por los responsables de los distintos motores de búsqueda y realmente, no se dispone de una información clara de cómo las motores lo llevan a cabo, con excepción del motor *Google* que ha hecho público su algoritmo *PageRank* [BRI, 1998]. Al igual que ocurría con los criterios de indización existen dos grandes grupos de algoritmos para el alineamiento, los que emplean variantes del modelo de espacio vectorial o del modelo booleano y los que siguen el principio de extensión de los enlaces.

Baeza-Yates cita tres métodos englobados en el primer grupo, en adición al esquema *tf-idf*: *"*se denominan *Booleano extendido, Vectorial extendido* y *Más citado*. Los dos primeros son adaptaciones de los algoritmos normales de alineamiento empleados en estos modelos clásicos de recuperación de información para incluir el hecho de la existencia de enlaces entre las páginas web. El tercero se basa únicamente en los términos incluidos en las páginas que poseen un enlace hacia las páginas de la respuesta" (Baeza-Yates, 1999).

El segundo grupo de algoritmos aporta una de las mayores diferencias conceptuales sobre el alineamiento: el uso de los enlaces decada página (tanto los que recibe una página como los que emanan de ella). El número de enlaces que apuntan a una página sirve como una medida de su popularidad y calidad. La presencia de enlaces comunes entre un conjunto de página es también una medida de relación de los temas tratadosen ellas.

Dentro de esta nueva tipología de técnicas de alineamiento, identificamos tres clases:

- *WebQuery*: da un alineamiento a las páginas que forman la respuesta a una consulta con base en cómo de conectadas están entre ellas. Adicionalmente, extiende el conjunto de páginas de la respuesta a otra serie de páginas altamente conectadas algrupo original de respuestas.

- *HITS*[44]: alinea las páginas Web en dos tipos distintos, que guardan una relación de mutua dependencia: *autoridades*(páginas muy referenciadas desde otras) y *hubs* (o conectores, páginas desde las que se hace referencia a otras consideradas porel autor de calidad en relación a un tema). Esta idea asume que cuando alguien establece un enlace a una página es porque la considera interesante, y que personas con intereses comunes tienden a referirse a las autoridades sobre un tema dentro de una misma página [ARA, 2000]. Conectores y autoridades son conceptos que se retroalimentan: mejores autoridades son inducidas por enlaces desde buenos conectores y buenos conectores vienen de enlaces desde buenas autoridades (Baeza- Yates, 1999).

- *PageRank* asume que el número de enlaces que una páginarecibe desde otras tiene mucho que ver con la calidad de la misma. Este algoritmo se puede resumir de la siguiente manera: "una página A tiene T1....Tn páginas que apuntan a ella por medio de algún enlace (es decir citas). El parámetro **d** es un factor que se puede fijar entre 0 y 1 (generalmente se fija en 0.85). Sea C(A) es número de enlaces que salen de la página A. Entonces, el *PageRank* de la página A vendrá dado por la siguiente expresión: **PR(A) = (1-d) + d(PR(T1)/C(T1) + + PR(Tn)/C(Tn))**". Este cálculo puede realizarse por medio de un algoritmo iterativo y corresponde al vector propio de una matriz normalizada de enlaces en la web. *PageRank* está concebido como un modelo del comportamiento del usuario: si se asume que hay un "navegante aleatorio" que pasa de una página a otra sin presionar nunca el botón de "retroceder" y que, eventualmente nunca se aburriera, la probabilidad de que este navegante visitara una página determinada es precisamente su *PageRank*. Es decir, se trata de un modelo basado en los enlaces de las páginas y que pretende representar la forma de trabajar de los usuarios. Otra justificación intuitiva de PageRank es que una página puede tener un alto coeficiente de PageRank si existen muchas páginas que apuntan a ella, o si hay un número algo menor de páginas que apuntan a ella pero que posean, a su vez, un alto nivel de PageRank. Lo normal es que "aquellas páginas muy citadas son páginas que vale la pena consultar y aquellas que sólo posean un enlace son páginas de poco interés para su consulta" [BRI, 1998].

Confianza en el funcionamiento de los motores de búsqueda.

Tras analizar el funcionamiento de los motores de búsqueda y conocer las particularidades de los problemas que afectan a la calidad de su funcionamiento, llega el momento de establecer si los mismos son fiables o no.

Es seguro que casi todos los usuarios de estos sistemas habrán reflexionado de una manera más o menos análoga al siguiente planteamiento de Manchón: "los resultados de algunos estudios indican que muchos usuarios prefieren la búsqueda jerárquica frente al motor de búsqueda. Ello puede ser causado por la proliferación de motores de búsqueda muy defectuosos que en la práctica no encuentran nunca la información deseada. Los usuarios se han acostumbrado a desconfiar de losmotores de búsqueda, ya que excepto en contadas ocasiones no funcionan bien. Por ejemplo, todos los usuarios muestran incredulidad y sorpresa mayúscula al usar *Google* y comprobar que realmente funciona bien" [MAN, 2002].

Esta confianza en *Google* puede deberse a que utiliza la estructura hipertextual de la web de dos maneras: primero para establecer el alineamiento de los documentos recuperados a través del algoritmo y segundo, para extender las búsquedas textuales. También emplea esta estructura para extender la búsqueda a documentos que no han sido o no pueden ser indexados. Para ello, complementa la información con el texto que acompaña al ancla del enlace [ARA, 2000].

Grado-Caffaro opina que "es fácil ver que el problema fundamental, en este contexto de los motores de búsqueda, es que no existe modo de garantizar, de momento, en el mercado, que las páginas que se han obtenido sean realmente las más relevantes y que el ranking obedezca a la realidad en términos de la *relevancia* de la información que se proporciona" [GRA, 2000].

Es decir, el problema planteado es explicar razonadamente por qué el motor de búsqueda proporciona unas páginas y no otras, o lo que es lo mismo, se trata de resolver el problema de la asignación de *relevancia* a las páginas devueltas con respecto a la temática de la pregunta planteada.

A pesar de las altas dosis de subjetividad que puedan estar presentes en estos postulados anteriores, no dejan de reflejar un problema muy común en el uso de los motores de búsqueda: estos sistemas muchas veces no proporcionan información verdaderamente relevante sobre un tema, a pesar de devolvernos una ingente cantidad de documentos en un tiempo relativamente escaso y de disponer el motor de una enorme base de datos con varios millones de documentos indexados. Este hecho provoca que

surjan opiniones tan rotundas como la anterior, que descalifica por completo la operatoria de estos sistemas.

Partiendo de una postura mucho más positivista, hay que intentar diferenciar los problemas que padecen estos sistemas a la hora de llevar a cabo correctamente su tarea, e intentar aislarlos en su contexto, exponiendo claramente su alcance y sus posibles soluciones. Esta serie de problemas podrían clasificarse de la siguiente manera:

1. Formulación adecuada de la pregunta
2. Interactividad con la interface de usuario
3. Inadecuada indización de los documentos
4. Actualización de los índices del motor

El primer grupo de problemas se encuentra muy ligado, la mayor parte de las veces, a una inadecuada formulación de la ecuación de búsqueda. Este problema, típico en la recuperación de información, cobra más importancia si cabe, en el contexto de los motores de búsqueda cuyos usuarios no tienen por qué disponer de unos conocimientos mínimos en técnicas de recuperación de información.

Este problema intenta ser paliado por los responsables de los propios motores quienes, en mayor o menor medida, insertan en la ayuda deestos sistemas explicaciones de cómo sacar el mejor partido al motor para recuperar información. También es frecuente encontrar publicaciones impresas y páginas web que realizan esta labor de asesoramiento al usuario no iniciado. Paralelamente al problema de los legos en la materia, surge el problema de adaptación que sufren algunos usuarios al cambiar de un motor a otro, aunque este problema es de menor incidencia. Con ello, el problema de la formulación inadecuada de las ecuaciones de búsqueda subyace y va a estar, de alguna manera, siempre presente en toda operación de recuperación de información.

El segundo problema es la interactividad con la interface del motor de búsqueda. En algunos casos esta interface ofrece escasas prestaciones a los usuarios para mejorar la calidad de sus operaciones de búsqueda y, en otros casos, resultan confusas e inducen a error a los usuarios, lo que tampoco contribuye a mejorar la efectividad del sistema.

El tercero de los problemas es el de la inadecuada indización de las páginas web. A la ausencia de una estructura básica de los documentos analizados por los robots, hay que unir lo reciente de esta tecnología (algunos motores con más de cien millones de páginas en su base de datos yaún se consideran “prototipos”).

En este punto confluyen muchas circunstancias, “además de por las propias limitaciones de la tecnología en su estado actual, existen también claros intereses, por parte de los propietarios de las páginas web, en que sus páginas aparezcan en la búsqueda y que aparezcan en la mejor posición.

Este interés, legítimo en principio, puede dejar de serlo cuando se utilizan mecanismos que distorsionan la realidad en ese afán por aparecer

en los procesos de búsqueda[45]. Además de este problema, no hay que olvidar que las técnicas de indización automática ofrecen un rendimiento en absoluto cercano a la perfección, por lo que los algoritmos que implementan los motores cometen fallos que se trasladan al conjunto de resultados.

Se ha comentado varias veces el problema que representa la naturaleza dinámica de la web para la actualización de los índices de los motores. Pero esta situación no puede dejar de ser óbice para que los administradores de los distintos sistemas, además de seguir recopilando nuevos recursos, presten la debida atención a mantener adecuadamente los índices de sus bases de datos. Hay que unir a esta tesitura el factor no sólo de la importancia/*relevancia* de la página sino de la importancia/*relevancia* del cambio que se ha podido producir lo que introduce un nuevo y adicionalnivel de complejidad a la efectividad de la recuperación de información.

Ante esta amplia serie de problemas, es por lo que estos sistemas precisan de herramientas de medida de su efectividad, que analicen de forma objetiva su rendimiento y establezcan enunciados sobre su funcionamiento que sean objetivos y fundamentados, alejados de opiniones personales e intuitivas, como las que se han recogido al principio de este apartado. Es por ello que, casi al mismo tiempo que surgieron los SRI se desarrollaron diversas técnicas para medir su rendimiento, tanto en el contexto tradicional como en el más reciente de la web.

3 Evaluación de la Recuperación de la Información.

La naturaleza determinista de los SRI propicia su necesidad intrínseca de evaluación. Por ello y paralelamente al desarrollo de su tecnología, surge un amplio campo de trabajo dedicado específicamente a la determinación de medidas que permitan valorar su efectividad. Un repaso exhaustivo de la bibliografía especializada permite identificar varios grupos de evaluaciones: las basadas en la *relevancia* de los documentos, las basadas en los usuarios y un tercer grupo de medidas alternativas a la realización de los juicios de *relevancia*, que pretenden evitar afectarse de las dosis de subjetividad que estos juicios poseen de forma inherente.

Necesidad de la evaluación de los SRI.

Los SRI, como cualquier otro sistema, son susceptibles de ser sometidos a evaluación, con el fin de que sus usuarios se encuentren en

---

[38] Delgado Domínguez comenta que a los robots se les denomina también "spiders" ("arañas") o "web crawlers" ("quienes andan a gatas por la web"). Baeza-Yates aporta otra denominación:"walkers" ("andadores"). Todos estos términos definen un movimiento lento y continuo entre las distintas páginas que conforman la web (que se puede traducir como "tela de araña").

[39] Muchos textos dicen erróneamente que el robot "se mueve a lo largo de la web", como si tuviera vida. Realmente es una aplicación informática que solicita una serie de transacciones a los servidores web que alojan las páginas analizadas.

[40] Esta modalidad también es conocida por "búsqueda por frase literal" o "búsqueda exacta".

[41] Es el caso de aquellos motores que permiten hacer uso del operador de truncamiento, como es el caso de *Alta Vista*.

[42] *Google* dispone de un "sistema de publicidad autoadministrada". En la práctica es una inclusión pagada de referencias a sitios web.

[43] Este módulo es repositorio de la relación inicial de páginas a analizar por el robot. En el mismo se almacena, en formato comprimido, el contenido de las mismas.

[44] HITS: Hypertext Induced Topic Search. Se puede traducir al Español como "Búsqueda de temas hipertextual inducida".

[45] Se han citado varias de estas técnicas de indebido uso, aunque desgraciadamente frecuente, en la página 21.

condiciones de valorar su efectividad y, de este modo, adquieran confianza en los mismos. Borlund opina que "la tradición de la evaluación de los SRI fue establecida desde la realización de los experimentos de *Cranfield*, seguido de los resultados y experiencias que Lancaster desarrolló en la evaluación de *MEDLARS* y los diversos proyectos SMART, de Salton; y hoy poseen vigencia con los experimentos TREC. Las evaluaciones de los SRI se encuentran estrechamente vinculadas con la investigación y el desarrollo de la recuperación de información" [BOR, 2000].

Blair afirma que "es la propia naturaleza de los SRI la que propicia su necesidad crítica de evaluación, justo como cualquier otro campo de trabajo que aspire a ser clasificado como campo científico" [BLA, 1990].

Baeza-Yates manifiesta que "un SRI puede ser evaluado por diversos criterios, incluyendo entre los mismos: la eficacia en la ejecución, el efectivo almacenamiento de los datos, la efectividad en la recuperación de la información y la serie de características que ofrece el sistema al usuario" (Baeza-Yates, 1992).

Estos criterios no deben confundirse, la *eficacia en la ejecución* es la medida del tiempo que se toma un SRI para realizar una operación. Este parámetro ha sido siempre la preocupación principal en un SRI, especialmente desde que muchos de ellos son interactivos, y un largo tiempo de recuperación interfiere con la utilidad del sistema, llegando a alejar a los usuarios del mismo. La *eficiencia del almacenamiento* es medida por el espacio que se precisa para almacenar los datos. Una medida común de medir esta eficiencia, es la ratio del tamaño del fichero índice unido al tamaño de los archivos de documentos, sobre el tamaño de los archivos de documentos, esta ratio es conocida como *exceso de espacio*. Los valores de esta ratio comprendidos entre 1,5 y 3 son típicos de los SRI basados en los ficheros inversos.

Para finalizar, Baeza-Yates subraya que "de forma tradicional se ha conferido mucha importancia a la efectividad de la recuperación, normalmente basada en la *relevancia* de los documentos recuperados" (Baeza-Yates, 1992).

Bors considera que "los experimentos, evaluaciones einvestigaciones tienen una larga tradición en la investigación de la recuperación de la información, especialmente los relacionados con el paradigma de comparación exacta, concentrados en mejorar el acierto entre los términos de una pregunta y la representación de los documentos para facilitar el aumento de la *exhaustividad* y de la *precisión* de las búsquedas" [BOR, 2000]. Este mismo autor sugiere una diferencia, a la hora de esa evaluación, entre la evaluación del "acceso físico" y la evaluación del "acceso lógico" (o "intelectual"); considerando que las evaluaciones que se

lleven a cabo, deben centrarse en el segundo tipo. El acceso físico es el que concierne a cómo la información demandada es recuperada y representada de forma física al usuario. Tiene que ver con la manera en la que un SRI (manual o automatizado) encuentra dicha información, o indica ciertas directrices al usuario sobre cómo localizarla, una vez que le proporciona su dirección. Este acceso se encuentra muy vinculado con las técnicas de recuperación y de presentación de la información. El acceso lógico está relacionado con la localización de la información deseada.

Para ilustrar las reflexiones anteriores, Blair propone el siguiente ejemplo: "consideremos una biblioteca: descubrir dónde se encuentra un libro con una signatura determinada es un problema relacionado con el acceso físico al objeto informativo (el libro); descubrir qué libro puede informarnos sobre una determinada materia es un problema relacionado con el acceso lógico" [BLA, 1990]. Este segundo caso tiene que ver con la *relevancia* del objeto localizado con una determinada petición de información. Es por ello que Blair considera que los problemas del acceso lógico cobran mucho más protagonismo frente a los problemas del acceso físico, que deben resolverse una vez solucionado los anteriores.

Estas afirmaciones, realizadas a principio de la década de los años noventa, siguen enteramente vigentes diez años después. Borlund distingue entre "aproximaciones al funcionamiento del sistema y aproximaciones centradas en el usuario" [BOR, 2000], plenamente coincidentes con el "acceso físico" y el "acceso lógico" de Blair. Otro hecho que no debemos perder de vista es la actual competencia establecida entre los desarrolladores de los algoritmos que emplean los directorios, motores de búsqueda o metabuscadores de la red Internet, sistemas que rivalizan sobre cómo facilitar al usuario más documentos en el menor tiempo posible, sin entrar a considerar que, quizá el usuario prefiera que la información entregada como respuesta le sea verdaderamente útil para sus necesidades, aunque tenga que esperar algunas milésimas de segundo más para recibirla.

Esta tendencia a avanzar en el desarrollo del aspecto físico certifica los temores de Blair, quien además considera que se llevan a cabo excesivas evaluaciones sobre determinados aspectos relacionados con el acceso físico, cuando donde deberían realizarse más evaluaciones es sobre el lógico. Siguiendo esta línea de razonamiento, ¿qué debería ser evaluado con el fin de determinar con certeza que la información que un SRI proporciona es válida para los usuarios del mismo? Para Blair, evidentemente, debería evaluarse el acceso lógico por medio del análisis de la *relevancia* o *no relevancia* del documento recuperado.

Baeza-Yates afirma que existen dos tipos de evaluaciones a efectuar: "cuando se analiza el tiempo de respuesta y el espacio requerido para la

gestión se estudia el rendimiento de las estructuras de datos empleadas en la indexación de los documentos, la interacción con el sistema, los retrasos de las redes de comunicaciones y cualquier otro retardo adicionalmente introducido por el software del sistema. Esta evaluación podría denominarse simplemente como *evaluación del funcionamiento del sistema"* (Baeza-Yates, 1999). En un SRI, los documentos recuperados no serán respuestas exactas a esta petición (a veces, porque los usuarios plantean las preguntas de una forma algo vaga). Los documentos recuperados se clasifican de acuerdo a su *relevancia* con la pregunta. Los SRI requieren evaluar cómo de relacionado con la temática objeto de la pregunta es el conjunto de documentos que forman la respuesta, "este tipo de evaluación, se conoce como *evaluación del funcionamiento de la recuperación*" (Baeza-Yates, 1999)

Relevancia vs Pertinencia.

En este punto surge una nueva cuestión, si bien podría parecertrivial a primera vista, puede marcar en gran medida el resultado de un proceso de evaluación. Esta cuestión no es otra que responder con certeza a la pregunta "¿cuándo un documento es relevante?".

El término *relevancia* significa "calidad o condición de relevante, importancia, significación", y el término "relevante" lo define como "importante o significativo". Entendemos, por extensión de las definiciones anteriores, que un documento recuperado se puede considerar relevante cuando el contenido del mismo posee alguna significación o importancia con motivo de la pregunta realizada por el usuario, es decir, con su necesidad de información. Conocer el significado del término no nos ayuda, desgraciadamente en demasía, ya que surgen nuevos problemas a la hora de determinar con exactitud cuándo un documento puede ser considerado relevante o no. No debemos olvidar que estos problemas se encuentran estrechamente entroncados con la naturaleza cognitiva de este proceso, destacando los siguientes a continuación:

- Un mismo documento puede ser considerado relevante, o no relevante, por dos personas distintas en función de los motivos que producen la necesidad de información o del grado de conocimiento que sobre la materia posean ambos. Llegados a un caso extremo, un mismo documento puede parecer relevante o no a la misma persona en momentos diferentes de tiempo [LAN, 1993].
- Resulta difícil definir, a priori, unos criterios para determinar cuándo un documento es relevante e incluso resulta complicado explicitarlos de forma clara y concisa, "es más fácil proceder a

la determinación de la *relevancia* que explicar cómo la misma se ha llevado a cabo" [BLA, 1990]. De hecho, considera que "el concepto de la *relevancia* no está claro en un primer término, se nos presenta como un concepto afectado de una gran dosis de subjetividad que puede ser explicado de múltiples maneras por distintas personas y, por tanto, dentro del contexto de una búsqueda en un sistema de recuperación de información este concepto puede ser precisado de muchas maneras distintas por todos aquellos que realicen búsquedas en el mismo. En un segundo lugar, no debe sorprendernos que un usuario afirme que unos determinados documentos son relevantes a sus necesidades de información y que, en cambio, no sea capaz de precisarnos con exactitud qué significa ser relevante para él" [BLA, 1990].

- Esto no quiere decir que el concepto carezca de importancia, sino que la realización de un juicio de *relevancia* viene a formar parte de ese amplio conjunto de tareas cotidianas que llevamos acabo los seres humanos (procesos cognitivos por tanto) pero que generalmente, no podemos encontrar las palabras adecuadas para proceder a su descripción" [BLA, 1990].

- Por último, puede resultar aventurado calificar de forma categórica a un documento como relevante con un tema, o por el contrario, calificarlo como no relevante de igual manera. Resulta muy normal encontrar documentos que, en alguno de sus apartados resulta relevante con una materia determinada pero que no en el resto de sus contenidos. Para algunos autores, surge entonces el concepto de "*relevancia* parcial", debido a que, en realidad, la *relevancia* no puede medirse en términos binarios (sí/no), sino que puede adquirir muchos valores intermedios (muy relevante, relevante, escasamente relevante, mínimamente relevante, etc.), lo que propicia que la *relevancia* pueda medirse en términos de función continua en lugar de una función binaria (que sólo admite dos estados).

Todos estos impedimentos condicionan, en cierto grado, la viabilidad de la *relevancia* para constituirse en un criterio de evaluación de la recuperación de la información. Cooper introduce la idea de "utilidad de un documento", considerando que es mejor definir a la *relevancia* en términos de la percepción que un usuario posee ante un documento recuperado, es decir: "si el mismo le va a ser útil o no" [COO, 1973]. Este nuevo punto de vista presenta, esencialmente, una ventaja: emplaza la estimación de la adecuación o no de un documento recuperado dentro del juicio que llevará a cabo el usuario, en tanto que, tal como hemos comentado anteriormente enumerando los problemas de la *relevancia*,

podemos asumir que un usuario tendrá problemas a la hora de definir qué es relevante y qué no lo es, pero tendrá pocos problemas a la hora de decidir si el documento le parece o no útil.

Es el usuario quién va a analizar el documento y quien lo va a utilizar si le conviene, por lo que los juicios de *relevancia* van a ser realizados por él, y son esos juicios de *relevancia* los que van propiciar que un SRI sea considerado bueno o malo. La importancia del concepto de "utilidad" lleva a Blair a concluir que "la misma simplifica el objetivo deun SRI y, aunque su evaluación es subjetiva, es posible medirla de mejor manera que si no se aplica este criterio, en tanto que es una noción primitiva que denota la realización de una actividad" [BLA, 1990].

Frants plantea otra acepción de *relevancia*, muy similar a la anterior, en términos de "eficiencia funcional", un SRI alcanzará altos niveles de este valor cuando la mayoría de los documentos recuperados satisfagan la demanda de información del usuario, es decir, le resulten útiles. [FRA, 1997].

Lancaster introduce un pensamiento muy interesante sobre esta cuestión: "aunque puede usarse otra terminología, la voz *relevancia* parece la más apropiada para indicar la relación entre un documento y una petición de información efectuada por un usuario, aunque puede resultar erróneo asumir que ese grado de relación es fijo e invariable, siendo mejor decir, que un documento ha sido juzgado como relevante a una específica petición de información". [LAN, 1993].

Prolongando este pensamiento, Lancaster reflexiona de una manera muy paralela a los planteamientos de Blair y considera que la *relevancia* de un documento con respecto a una necesidad de información planteada por un usuario no tiene por qué coincidir con los juicios de valor que emitan muchos expertos sobre el contenido del documento sino con la satisfacción de ese usuario y la "utilidad" que estos contenidos van a tener para él, opinando que "es mejor, en este segundo caso, hacer uso de la palabra *pertinencia*". Es decir, *relevancia* va a quedar asociada con el concepto de la relación existente entre los contenidos de un documento con una temáticadeterminada y *pertinencia* va a restringirse a la "relación de utilidad" existente entre un documento recuperado y una necesidad de información individual.

Para Salton: "el conjunto pertinente de documentos recuperados puede definirse como el subconjunto de los documentos almacenados en el sistema que es apropiado para la necesidad de información del usuario" (Salton & McGill, 1983). El término *pertinencia* significa "calidad de pertinente", entendiéndose como "pertinente" a todo lo que viene a propósito o resulta oportuno, es decir que podemos decir que un documento

pertinente es un documento que resulta oportuno, porque le proporciona al usuario final la información que a él le cumple algún propósito.

Opiniones similares de esta distinción se recogen en el trabajo de Foskett, quien define como *documento relevante* a aquel "documento perteneciente al campo/materia/universo del discurso delimitado por los términos de la pregunta, establecido por el consenso de los trabajadores en ese campo", igualmente define como documento pertinente a "aquel documento que añade nueva información a la previamente almacenada en la mete del usuario, que le resulta útil en el trabajo que ha propiciado la pregunta" [FOS, 1972].

Verdaderamente, "las diferentes aproximaciones que se desarrollan para evaluar los SRI poseen todas los mismos objetivos finales, porque los procesos de evaluación están relacionados con la capacidad del sistema de satisfacer las necesidades de información de sus usuarios" [BOR, 2000]. Bibliografía adicional sobre esta acepción de la *relevancia* es propuesta por Lancaster [LAN, 1993] recogiendo citas de Cooper (ya citado anteriormente), Goffman, Wilson, Bezer y O'Connor. Otra recopilación de citas sobre este concepto la realiza Mizzaro, citando, entre otros, a Vickery, Rees y Schultz, Cuadra y Katter, Saracevic y Schamber [MIZ, 1998].

Este conjunto de opiniones ha sido aceptado por los autores que han trabajado con posterioridad en este campo, de hecho, en un número especial de la revista *Informing Science* dedicado a la investigación en nuestra área (volumen 3 del año 2000), encontramos la siguiente frase de Greisdorf: "en los últimos treinta años no se ha encontrado sustituto práctico para el concepto de *relevancia* como criterio de medida y cuantificación de la efectividad de los SRI" [GRE, 2000].

Incidiendo en esta opinión, Gordon y Pathak opinan: "si los juicios de *relevancia* son llevados a cabo por expertos en la materia, decidiendo por su cuenta qué páginas web son relevantes y cuáles no, se introducirían muchas disfunciones debidas a la familiaridad con la materia o al desconocimiento exacto de las necesidades de información del usuario, de las motivaciones que provocan la necesidad de información y de otra serie de detalles más o menos subjetivos que escapan a la percepción del localizador de información. No podemos enfatizar suficientemente la importancia de los juicios de *relevancia* realizados por aquellos quienes verdaderamente necesitan la información" [GOR, 1999].

En realidad, los documentos recuperados no son relevantes o no relevantes, propiamente hablando, es decir, que no se trata de una decisión binaria, en tanto que los contenidos de los documentos pueden coincidir en mayor o menor parte con las necesidades de información. Lo que sí podemos determinar es si son o no relevantes para una determinada

persona. Desde un punto de vista pragmático, el mismo documento puede significar varias cosas para personas diferentes; los juicios de *relevancia* pueden sólo realizar evaluaciones semánticas o incluso sintácticas de documentos o preguntas. Pero estos juicios fallan al involucrar a usuarios particulares, y también fallan al identificar dónde el usuario realmente encuentra a un documento particularmente relevante. Estos autoresbromean un poco al respecto, parafraseando, ligeramente modificada, una vieja frase: "la *relevancia* reside en los detalles".

Asumimos, por tanto, el planteamiento de que un documento será relevante para nuestra necesidad de información, cuando el mismo verdaderamente nos aporte algún contenido relacionado con nuestra petición, con lo cual, realmente, cuando hablemos de *relevancia* podemos estar hablando de *pertinencia*, siempre que estemos refiriéndonos al punto de vista del usuario final que realiza una operación de recuperación de información.