

# 2

## La Recuperación y los Sistemas de Recuperación de Información.

**RESUMEN:** Este capítulo constituye una presentación del concepto de recuperación de información, y de la amplia serie de diferencias que lo delimitan de otras aplicaciones de la informática en lo relacionado con la gestión de datos. Al mismo tiempo se exponen los distintos modelos sobre los que se basan los sistemas que permiten esa recuperación, confiriéndosele especial importancia a todo lo relacionado con la recuperación de información en la web, contexto sobre el cual se va a desarrollar nuestra propuesta de evaluación.



## La Recuperación y los Sistemas de Recuperación de Información.

<b>Hacia una definición de la Recuperación de Información.....</b>	<b>11</b>
<b>Sistemas de Recuperación de Información. ....</b>	<b>17</b>
Vista funcional de un SRI.....	17
Evolución de los SRI.....	19
<b>Modelos para la recuperación de información.....</b>	<b>20</b>
<b>La recuperación de información en la web.....</b>	<b>22</b>
Breve perspectiva histórica de la web.....	23
Métodos de recuperación de información en la web.....	25
<b>Los motores de búsqueda como paradigma de la recuperación de información en Internet.....</b>	<b>34</b>
Funcionamiento de un motor de búsqueda.....	35
Arquitectura de un motor de búsqueda.....	35
Los índices de los motores.....	42
Tipos de robots.....	44
Funcionamiento de los robots.....	44
Indización de las páginas.....	46
Alineado de los documentos (ranking).....	48
<b>Confianza en el funcionamiento de los motores de búsqueda.....</b>	<b>50</b>
<b>Tablas e Ilustraciones.....</b>	<b>53</b>

### Hacia una definición de la Recuperación de Información.

Resulta cuando menos curioso el hecho de que un concepto tan empleado como el de recuperación de información presente cierta confusión a la hora de establecer una definición que lo sitúe adecuadamente dentro del campo de las *Ciencias de la Información*. Rijsbergen es el autor que mejor introduce este problema al considerar que "se trata de un término que suele ser definido en un sentido muy amplio" [RIJ, 1999], y Lancaster avisa al indicar que "el concepto de recuperación de información es de aquellos que

pueden resultar sencillos de definir, llevado a ello por la gran profusión de veces en las que es empleado". [LAN, 1993]

El profuso uso de este término, al igual que ocurre en otras disciplinas con otros vocablos que también pueden parecer básicos, ha propiciado que el mismo no se encuentre bien empleado en muchas ocasiones, ya que unas veces los autores lo presentan como sinónimo de la recuperación de datos llevada desde la perspectiva de las base de datos.

Otro conjunto de autores 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 anterior), un tercer grupo de autores la define de forma muy genérica sin entrar en mayores consideraciones sobre estas diferencias, y un cuarto y último grupo pasa de largo sobre este problema, profundizando más en la explicación de los sistemas de recuperación de información<sup>3</sup> (SRI en adelante).

El primer grupo de definiciones debe su naturaleza a la clara influencia de la tecnología informática, cuya evolución ha inducido a muchos autores a cometer el error de considerar sinónimos ambos conceptos, llegándose a olvidar que se puede recuperar información sin procedimientos informáticos (aunque no es lo más común hoy en día, evidentemente), pero el hecho del frecuente y necesario empleo de una tecnología no debe sustituir el adecuado uso de los conceptos terminológicos. Un claro ejemplo de este desacierto lo encontramos en el *Glosario de la Asociación de Bibliotecarios Americanos*<sup>4</sup>, que define el término inglés "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*<sup>5</sup>. De la misma opinión es el *Diccionario Mac Millan de Tecnología de la Información*, que considera a la recuperación de información como el conjunto de "técnicas empleadas para almacenar y buscar grandes cantidades de datos y ponerlos a disposición de los usuarios" [LON, 1989]

---

<sup>3</sup> Quizás el esfuerzo realizado por los autores en definir a estos sistemas ha favorecido que el concepto de recuperación haya quedado relegado a un segundo plano.

<sup>4</sup> A.L.A. : American Library Association.

<sup>5</sup> Este Glosario indica que "document retrieval" puede considerarse término sinónimo de "information retrieval".

Un segundo grupo de autores establecen diferencias entre ambos conceptos. Meadow piensa que la recuperación de la información “se trata de una disciplina que involucra la localización de una determinada información dentro de un almacén de información o base de datos” [MEA, 1992]. Este autor, implícitamente, establece que el concepto de recuperación de información se encuentra asociado con el concepto de *selectividad*, ya que la información específica ha de extraerse siguiendo algún tipo de criterio discriminatorio (selectivo por tanto). Pérez-Carballo y Strzalkowski redundan en esta tesis, en tanto que “una típica tarea de la recuperación de información es *traer* documentos relevantes desde una gran archivo en respuesta a una pregunta formulada por un usuario y ordenar estos documentos de acuerdo con su *relevancia*” [PER, 2000]. Igualmente, Grossman y Frieder indican que “la recuperación de información es encontrar documentos relevantes, no encontrar simples correspondencias a unos patrones de bits” [GRO, 1998]

Meadow afirma igualmente que no es lo mismo la recuperación de información entendida como traducción del término inglés *information recovery* que cuando se traduce el término *information retrieval*, ya que “en el primer caso no es necesario proceso de selección alguno” [MEA, 1992].

De esta misma opinión es Blair, quien dedica gran parte de la presentación de su libro<sup>6</sup> a establecer una clara diferencia entre el término *data retrieval* y el término *information retrieval*, utilizando como criterios distintivos, entre otros [BLA, 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 pueden contener, sólo probablemente, lo deseado, con un evidente factor de indeterminación.
2. De lo anterior, se deduce que según la relación entre el requerimiento al sistema y la satisfacción de usuario, la recuperación de datos es

---

<sup>6</sup> Blair, D.C. *Language and representation in information retrieval*. Amsterdam [etc.]: Elsevier Science Publishers, 1990.

*determinista* y en recuperación de información es *posibilista*, por causa del 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 obtenida satisface las necesidades de información del usuario, es decir, su percepción personal de utilidad.

Tramullas Sáez destaca un aspecto importante de las reflexiones de Blair, "la importancia, en ocasiones ignorada, que tiene el factor de predicción. Predicción por parte del usuario, ya que éste debe intuir, en numerosas ocasiones, los términos que han sido utilizados para representar el contenido de los documentos, independientemente de la presencia de mecanismos de control terminológico. Este criterio de predicción es otro de los elementos que desempeñan un papel fundamental en el complejo proceso de la recuperación de información" [TRA, 1997] y que no se presenta en el campo de la recuperación de datos.

La Tabla 2.1 sintetiza las diferencias fundamentales existentes entre *recuperación de datos* y *recuperación de información* para Rijsbergen [RIJ, 1999]:

	<b>Recuperación de datos</b>	<b>Recuperación de información</b>
Acierto (correspondencia)	Exacta	Parcial, la mejor
Inferencia	Algebraica	Inductiva
Modelo	Determinístico	Posibilístico
Lenguaje de consulta	Fuertemente Estructurado	Estructurado o Natural
Especificación de la consulta	Precisa	Imprecisa
Error en la respuesta	Sensible	Insensible

Tabla 2.1 Diferencias entre recuperación de datos y 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 expone las diferencias entre ambos tipos de recuperación con argumentos algo menos abstractos que los anteriores, destacando 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" [BAE, 1999].

Para este autor, el problema de la recuperación de información se puede definir como “dada una necesidad de información (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*” [BAE, 1999]. 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 ha venido preocupándose especialmente del tema de las estructuras de datos y de los métodos de acceso a los mismos [BAE, 1992], [BAE, 1999], siendo este autor una verdadera referencia en esta materia.

A la hora de definir la recuperación de información, en lugar de proponer una definición propia, hace uso de la elaborada por Salton: “la recuperación de la información tiene que ver con la representación, almacenamiento, organización y acceso a los ítem de información” [SAL, 1983].

Salton indica que, en principio, no deben existir limitaciones a la naturaleza del objeto informativo y 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. Desafortunadamente, la caracterización de la necesidad informativa de un usuario no es un problema sencillo de resolver” [BAE, 1999].

En el tercer grupo de autores encontramos definiciones esencialmente iguales a la realizada por Salton (que puede considerarse la base del resto de definiciones que pueden encontrarse en la bibliografía especializada en la materia y la que mejor refleja la definición de recuperación de información), aunque, en este caso, el rasgo diferenciador de estos trabajos reside en que sus autores no profundizan en las diferencias entre “recuperación de datos” y “recuperación de información”, bien por no ser objeto de sus trabajos o 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 estima que la recuperación de información es “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.” [CRO, 1987].

Tramullas Sáez impregna su definición del carácter selectivo comentado anteriormente al afirmar que “el planteamiento de la recuperación de información en su moderno concepto y discusión, hay que buscarlo en la realización de los tests de *Cranfield* y en la 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” [TRA, 1997].

El cuarto y último grupo de autores caracterizan sus trabajos porque eluden llevar a cabo una definición de recuperación de la información. Este grupo tiene como máximo exponente a Chowdhury, quien simplemente dedica el primer párrafo de su libro<sup>7</sup> a decir que “el término recuperación de la información fue acuñado en 1952 y fue ganando popularidad en la comunidad científica de 1961 en adelante<sup>8</sup>” [CHO, 1999], pasando inmediatamente a mostrar los propósitos, funciones y componentes de los SRI.

Otro autor de esta tendencia es Korfhage, quien se centra en el almacenamiento y recuperación de la información, considerando a estos procesos como las dos caras de una moneda, aunque no entra a definirlos. Para Korfhage, “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” [KOR, 1997]. Realmente este autor se dedica más en presentar las diferencias existentes entre dato, información, señal<sup>9</sup>, conocimiento y sabiduría [MEA, 1992], [MAR, 1999].

---

<sup>7</sup> Chowdhury, G. G. *Introduction to modern information retrieval*. London: Library Association, 1999.

<sup>8</sup> Chowdhury introduce una cita de Rijsbergen y Aogsti, correspondiente al artículo “The Context of Information”, *The Computer Journal*, vol 35 (2), 1992.

<sup>9</sup> Para entender el concepto de señal, Korfhage remite a la *Teoría Matemática de la Comunicación* de Shannon (más conocida como *Teoría de la Información*). Fuente: ‘A mathematical theory of communication’ *Bell Systems Technical Journal* 27, 1948.

## Sistemas de Recuperación de Información.

Tomando como base de partida la definición de recuperación de información concebida por Salton [SAL, 1983], unida ésta a las aportaciones de Rijsbergen<sup>10</sup> [RIJ, 1999], correspondería ahora, siguiendo la opinión de Baeza-Yates [BAE, 1999], elegir el mejor modelo para el diseño de un sistema de recuperación de información (SRI en adelante), aunque antes resulta necesario proceder a una adecuada conceptualización de qué se entiende por “sistema de recuperación de información” y cuál es su utilidad.

### Vista funcional de un SRI.

Las manifiestas similitudes existentes entre la recuperación de información y otras áreas vinculadas al procesamiento de la información, propician que las mismas se trasladen hacia el campo de los sistemas encargados de llevar a cabo esta tarea.

Salton opina que “la recuperación de información se entiende mejor cuando uno recuerda que la información que se procesa consiste en 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 entiende que “cualquier SRI puede ser descrito como un conjunto de ítem 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” [SAL, 1983]

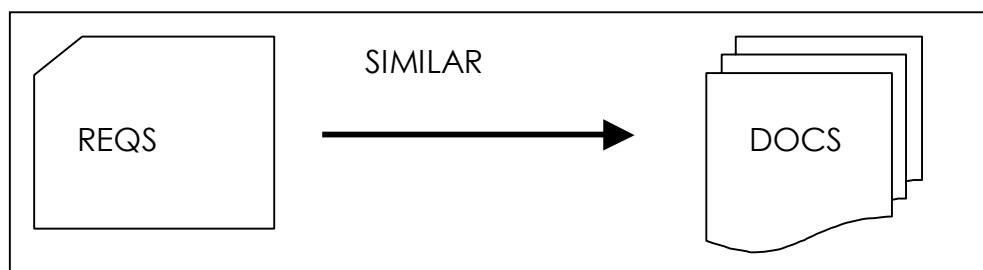


Ilustración 2.1 Esquema simple de un SRI. Fuente Salton , G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983.

---

<sup>10</sup> Recogidas en la Tabla 2.1

Salton opina que, en la práctica, este esquema inicial reflejado en la Ilustración 2.1, resulta algo simple y es necesario ampliarlo, “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” [SAL, 1983]

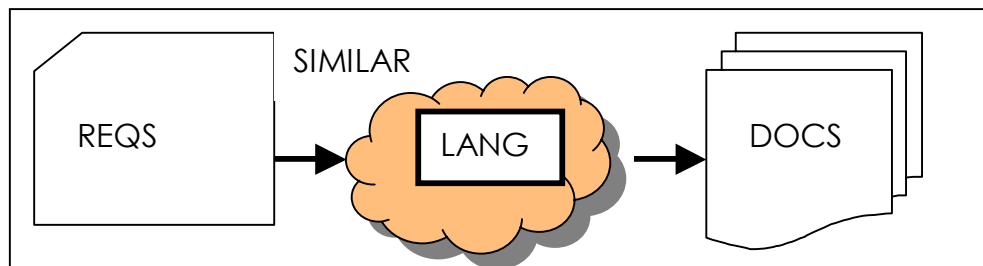


Ilustración 2.2 Esquema avanzado de un SRI. Fuente Salton , G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983.

En la Ilustración 2.2, se observa que el proceso establecido entre la entrada REQS y SIMILAR es el proceso de formulación de la búsqueda, y el establecido entre SIMILAR y el conjunto de documentos DOCS es el proceso de 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. 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 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.**

Muchos autores presentan la evolución de estos sistemas desde la puesta en marcha del primero de ellos, pero el que mejor simplifica este progreso es Baeza-Yates, quien muestra tres fases fundamentales [BAE, 1999]:

1. *Desarrollos iniciales.* El autor refleja que ya existían métodos de recuperación de información con las antiguas colecciones de papiros. Otro ejemplo típico sería la *tabla de contenidos* de un libro, sustituida por otras estructuras algo más complejas a medida que ha crecido el volumen de información a gestionar. La evolución lógica de la tabla de contenidos fue el *índice*, estructura núcleo de los SRI actuales.
2. *Recuperación de información en las bibliotecas.* Estas instituciones fueron de las primeras en adoptar estos sistemas. Originalmente eran desarrollados por las propias bibliotecas y posteriormente se ha creado un mercado de aplicaciones informáticas altamente especializadas en este sector. Se identifican varias generaciones: mecanización de los catálogos manuales, aumento de las posibilidades de búsqueda y una tercera generación se encuentra trabajando en el desarrollo de interfaces gráficas, características de hipertexto, arquitecturas de sistemas abiertos y automatización de procesos.
3. *La World Wide Web.* La evolución lógica de los SRI ha sido hacia la web, donde han encontrado gran aplicación práctica y un aumento del número de usuarios, especialmente en el campo de los directorios y motores de búsqueda<sup>11</sup>. El alto grado de consolidación de la web, con apenas diez años transcurridos desde su desarrollo, se ha visto favorecido por el abaratamiento de la tecnología informática, por el alto grado de desarrollo de las telecomunicaciones y por la facilidad que posee cualquier usuario de este sistema de hacer público cualquier documento que considere interesante, sin tener que pasar el filtro de los tradicionales círculos editoriales. De esta manera, han aumentado los usuarios y lo que es más importante, los contenidos.

Los SRI también han evolucionado para adaptarse a este nuevo entorno, aunque su novedad, no permite disponer aún de unas definidas metodologías que evalúen su efectividad. Esta evolución no es un proceso finalizado, sino más bien un proceso en realización, que lleva al

---

<sup>11</sup> Estos sistemas se presentan posteriormente en el apartado dedicado a la *recuperación de la información en la web*, dentro de este mismo capítulo.

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 ha de quedar 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, los métodos para establecer la importancia (orden) de los documentos de salida y los mecanismos que permiten una realimentación por parte del usuario para mejorar la consulta” [VIL, 1997]. Existen varias propuestas de clasificación de los modelos de recuperación, una de las más completas la realiza Dominich, quien establece cinco grupos [DOM, 2000]:

Modelo	Descripción
Modelos clásicos	Incluye los tres más comúnmente citados: <i>booleano</i> , <i>espacio vectorial</i> y <i>probabilístico</i> .
Modelos alternativos	Están basados en la Lógica Fuzzy
Modelos lógicos	Desarrollados en la década de los noventa, basados en la Lógica Formal. La recuperación de información se entiende como un proceso inferencial a través del cual se puede estimar la probabilidad de que una necesidad de información de un usuario, expresada como una o más consultas, sea satisfecha ofreciendo un documento como “prueba” [VIL, 1997].
Modelos basados en la interactividad	Incluyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la <i>relevancia</i> de los documentos recuperados [SAL, 1989]
Modelos basados en la Inteligencia Artificial <sup>12</sup>	Bases de conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural.

Tabla 2.2 Clasificación de los Modelos de Recuperación de Información según Dominich. Fuente: Dominich, S. ‘A unified mathematical definition of classical information retrieval’. Journal of the American Society for Information Science, 51 (7), 2000. p. 614-624.

<sup>12</sup> Respetando los grupos establecidos por Dominich, surgen serias dudas a la hora de no considerar los Modelos Lógicos parte de los Modelos basados en la Inteligencia Artificial, realmente podrían englobarse en el mismo grupo de modelos.

Baeza-Yates lleva a cabo una clasificación taxonómica de los distintos modelos de recuperación de información, a partir de la tarea inicial que realiza el usuario del sistema, que puede consistir en recuperar información por medio de una ecuación de búsqueda (*retrieval*) que se inserta en un formulario destinado a ello, o bien dedicar un tiempo a consultar (*browse*<sup>13</sup>) los documentos en la búsqueda de referencias interesantes [BAE, 1999], dando entrada en su clasificación al *hipertexto* [CON, 1988] [NIE, 1990], modelo en el cual se basa la web [BER, 1992].

Este 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 uno de estos modelos: 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 [BAE, 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:

---

<sup>13</sup> El término inglés "browse" suele traducirse como "navegar" u "hojear" las páginas web a través de su estructura hipertextual. Del mismo modo, "browsing" puede traducirse como "navegación" u "hojeo". Este término también se empleaba dentro de la terminología del Hipertexto y ha sido adoptado por la terminología propia de la web.

## Vista lógica de los documentos

<b>Modalidad</b>		<b>Términos Índice</b>	<b>Texto Completo</b>	<b>Texto Completo + Estructura</b>
	<b>Recuperación</b>	Clásicos Conjuntos teóricos Algebraicos Probabilísticos	Clásicos Conjuntos teóricos Algebraicos Probabilísticas	Estructurados
	<b>Navegación</b>	Estructura plana	Estructura plana Hipertexto	Estructura guiada Hipertexto

Tabla 2.3 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 [BAE, 1999] como Villena Román [VIL, 1997] llevan a cabo una presentación detallada de cada uno de los modelos, siendo también interesante la lectura de Grossman y Frieder<sup>14</sup> [GRO, 1998], para conocer las alternativas a los modelos clásicos.

### La recuperación de información en la web.

Hu recuerda que “el primer motor de búsqueda desarrollado en la red Internet fue ARCHIE<sup>15</sup>, creado en 1990, aunque no fue hasta la creación del primer navegador, *Mosaic*<sup>16</sup>, que propició el crecimiento de los documentos y la gestión de información multimedia hasta que se expandió el uso de estos sistemas” [HU, 2001].

<sup>14</sup> Estos autores prefieren emplear el término “estrategias de recuperación de información” en lugar del término “modelo”.

<sup>15</sup> 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.

<sup>16</sup> *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.

La web<sup>17</sup> 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 en la que difundía la invención de la *World Wide Web* y contactó con Tim Berners-Lee, becario británico del CERN<sup>18</sup>. Berners-Lee estaba decidido a desarrollar un método eficiente y rápido para intercambiar datos científicos 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.

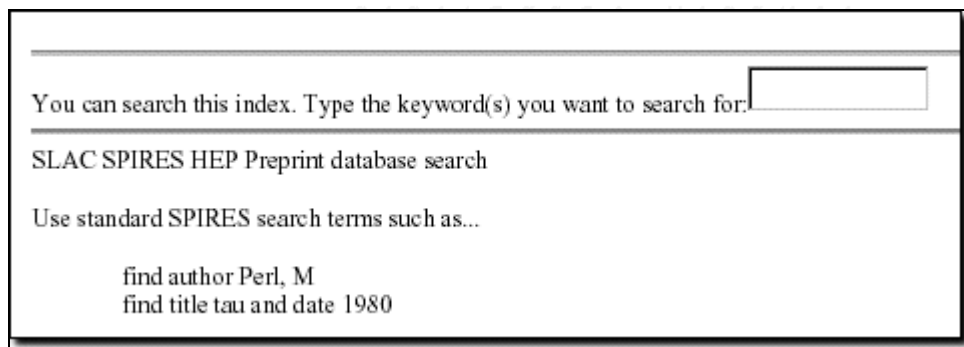
---

<sup>17</sup> En el resto de este trabajo se va a utilizar esta manera de denominar al sistema “World Wide Web” para simplificar la lectura.

<sup>18</sup> 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”).

- Integración de un amplio rango de tecnologías y distintos tipos de documentos.
- Proporcionar una herramienta que permita leer los documentos desde cualquier lugar de Internet y por cualquier individuo que este navegando dentro de este almacén, permitiendo accesos simultáneos.

Berners-Lee propuso a Kunz que almacenara la información de su departamento en Stanford en un servidor, con el fin de que otros científicos pudieran consultarla a través de Internet a través del formulario que se recoge en la ilustración siguiente. El propio Berners-Lee fue el primero en probarlo.



The image shows a screenshot of a web search interface. At the top, there is a horizontal line followed by the text "You can search this index. Type the keyword(s) you want to search for:" and a small rectangular input box. Below this, the text "SLAC SPIRES HEP Preprint database search" is displayed. Underneath, it says "Use standard SPIRES search terms such as...". At the bottom, two example search terms are listed: "find author Perl, M" and "find title tau and date 1980".

Ilustración 2.3 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*, permitió aumentar el número de usuarios de la red Internet que accedía a este novedoso sistema, crecimiento que continuó con el desarrollo de nuevos navegadores: *Netscape* e *Internet Explorer*<sup>19</sup>. A principios del año 1995 se formó la organización *Consortio World Wide Web*<sup>20</sup> (o *W3C*) que está bajo la dirección del fundador de la Web

Cabe reflexionar en este punto que si bien el ritmo de introducción del ordenador (como vehículo transmisor de la Informática) en todos los ámbitos

---

<sup>19</sup> Aplicaciones informáticas que han desencadenado una batalla legal de lucha antimonopolio sólo comparable a la entablada a principios del siglo XX para la extracción del petróleo, que se encontraba en manos del magnate Rockefeller y su compañía 'Standard Oil Company'.

<sup>20</sup> Traducción del término inglés "World Wide Web Consortium". La URL de esta organización es <<http://www.w3c.org/>>. Actualmente dispone de más de 500 organizaciones asociadas.

de nuestra sociedad fue extraordinario, el desarrollo y expansión de la web ha rebasado exponencialmente esos valores y, según datos de noviembre de 2001, hay activos más de treinta y seis millones de servidores web y existen unos 1.600 millones de páginas web, justo al cumplirse diez años de la puesta en marcha de la primera página.

### **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 es por ello que han comenzado a surgir grandes problemas cuando se realizan operaciones de recuperación de información con ellos, 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, como por ejemplo el peculiar fenómeno denominado *spamming*<sup>21</sup> o los relacionados 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*<sup>22</sup>, 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*<sup>23</sup>, sistemas que clasifican documentos web seleccionados por materia y que nos permiten navegar por sus secciones o

---

<sup>21</sup> Los constructores de páginas web insertan en la descripción de las mismas términos que nada tienen que ver con el contenido de las mismas, por ejemplo: "mp3", "sex", "pokemon", "Microsoft" (términos todos ellos de uso muy frecuente por todos aquellos usuarios de los motores de búsqueda), provocando que estos usuarios recuperen esas páginas "trucadas", cuando pretenden recuperar documentos de otra temática.

<sup>22</sup> Se va a emplear el término “motor de búsqueda” como traducción del término “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.

<sup>23</sup> Se va a emplear 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.

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<sup>24</sup>)" [BAE, 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<sup>25</sup>), 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 hojear de los recursos descendiendo desde los temas más generales a los más 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 través de dos sistemas: navegación a través de la estructura de las categorías y búsqueda por palabras claves sobre el conjunto de referencias contenidas en el índice. El directorio más grande y famoso es *Yahoo!*, aunque existen otros bastante

---

<sup>24</sup> En algunos textos se emplea el término "hiperenlace" como traducción de "hyperlink".

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

conocidos: *Dmoz* (un directorio alimentado por miles de colaboradores), *Looksmart*, *Infospace* e *Hispanicista*. Con mucha diferencia, el más utilizado es *Yahoo!*. Los directorios son más usados que los motores, especialmente cuando “no se conoce exactamente el objetivo de la búsqueda” [MAN, 2002], ya que resulta difícil acertar con los términos de búsqueda adecuados.

Los motores de búsqueda son aplicaciones que manejan también grandes bases de datos de referencias a páginas web recopiladas automáticamente, sin intervención humana. Uno o varios agentes de búsqueda recorren la web, a partir de una lista inicial de direcciones 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 y proporcionan el resultado más adecuado a una búsqueda. También almacenan direcciones que les son remitidas por los usuarios<sup>26</sup>. Entre los motores más populares destacan *Altavista*, *Lycos*, *Alltheweb*, *Hotbot*, *Overture*, *Askjeeves*, *Direct Hit*, *Google*, *Microsoft Network*, *Terra* y *WISEnut*, entre otros. Delgado Domínguez resume en la Tabla 2.4 las características básicas de estos dos métodos de recuperación de información en la web:

	<b>Descubrimiento de recursos</b>	<b>Representación del contenido</b>	<b>Representación de la consulta</b>	<b>Presentación de los resultados</b>
<b>Directorios</b>	Lo realizan personas	Clasificación manual	Implícita (navegación por categorías)	Páginas creadas antes de la consulta. Poco exhaustivos, muy precisos
<b>Motores de búsqueda</b>	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 2.4 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/adelaide/tice/modul6/memfin.pdf>> [Consulta: 18 de septiembre de 2001]

Es oportuno puntualizar que el razonamiento que lleva a la autora a considerar un directorio más preciso que un motor, se basa, sin duda alguna,

<sup>26</sup> Aunque algunas fuentes cifran entre el 95% y 97% el número de URL que son rechazadas por estos motores por diversos motivos.

en la fiabilidad de la descripción del registro, realizada manualmente de forma detallada y ajustada, entendiéndose 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<sup>27</sup>. 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 [BAE, 1999].

Los *lenguajes de consulta a la web*<sup>28</sup> 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 [BAE, 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

---

<sup>27</sup> Más vinculada a términos como *relevancia* o *pertinencia* del documento recuperado con respecto de la temática objeto de la pregunta.

<sup>28</sup> Traducción literal del término inglés “web query languages”.

desarrollados pueden entenderse como extensiones del lenguaje SQL<sup>29</sup> para el contexto de la web empleado en los tradicionales gestores relacionales.

Todos estos modelos constituyen adaptaciones o propuestas de desarrollo de sistemas navegacionales para la consulta de hipertextos [CAN, 1990], [NIE, 1990], [BAE, 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” [BAE, 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, lo que 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, *metabuscaadores*<sup>30</sup> y técnicas de *filtrado de información*<sup>31</sup> [CHA, 2001].

Los *metabuscaadores* 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 principal ventaja 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” [BAE, 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

---

<sup>29</sup> SQL: Structured Query Language.

<sup>30</sup> “Metabuscador” es la traducción más aceptada del término “meta-search engine”.

<sup>31</sup> Traducción literal del término inglés “information filtering”.

y crawlers<sup>32</sup>) que tienen registrados y obtienen los resultados que les devuelven. 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 hacer uso de varios de ellos" [BRA, 2000].

Estos sistemas se diferencian unos de otros en la manera en que llevan a cabo el alineamiento de los resultados<sup>33</sup> en el conjunto unificado<sup>34</sup>, y cómo de bien traducen estos sistemas la pregunta formulada por el usuario a los lenguajes específicos de interrogación que maneja cada sistema, 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 (un ejemplo es *All4One* que busca en una gran cantidad de motores de búsqueda y directorios); los meta buscadores funcionan de manera similar a los 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].

Algunos de estos sistemas presentan los resultados en diferentes ventanas, correspondiendo cada una de ellas a una fuente distinta (*Oneseek* o *Proteus*, por ejemplo).

---

<sup>32</sup> Este término se refiere al robot que recopila páginas web para el índice de los motores de búsqueda.

<sup>33</sup> Aunque el Diccionario de la R.A.E. admite el uso del vocablo "ranking", preferimos emplear el término "alineamiento", al tratarse el anterior de un anglicismo.

<sup>34</sup> En algunos casos este proceso no se realiza [BAE, 1999].

Uno de los mayores inconvenientes de estos sistemas es que el resultado no tiene porqué ser necesariamente todo el conjunto de páginas sobre la materia preguntada que se encuentran almacenadas en las fuentes del metabuscador, ya que el número de documentos recuperados de cada una de estas fuentes se encuentra generalmente limitado, “sin embargo, el resultado devuelto por un metabuscador suele ser más relevante en su conjunto” [BAE, 1999]. Puede sorprender la existencia de esta limitación, pero no se debe olvidar uno de los elementos que, tradicionalmente, más han incidido en las evaluaciones de los SRI: el *tiempo de respuesta del sistema* [LAN, 1993].

Si un metabuscador devolviera todas las referencias de todos los motores y directorios que le sirven de fuente en relación con la materia objeto de una búsqueda, el tiempo de respuesta del sistema alcanzaría valores que seguramente alejarían a los usuarios del metabuscador por excesivo. Es por ello que 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) [BAE, 1999].

Las técnicas de filtrado de la información que comenta Chang son más un complemento de los motores de búsqueda 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 “se basan en una combinación de sistemas de autoaprendizaje y sistemas de recuperación de información, y han sido

empleadas en la construcción de motores de búsqueda especializados" [CHA, 2001].

Tal como se observa en la Ilustración 2.4, el filtrado de términos mejora la calidad del índice del motor, rechazando términos de escaso o nulo valor de discriminación y contribuyendo a acrecentar la velocidad de la recuperación de información, al aligerar las dimensiones del fichero índice.

Según el esquema de la Ilustración 2.4, el agente (o los agentes) encargados de recopilar la información por la web, someten las páginas que recuperan al sistema de filtrado, el cual si las acepta, las almacenará en el índice del motor, mientras que las rechazadas son descartadas. Entre los motores de búsqueda más conocidos que emplean estas técnicas destaca *Northern Light*.

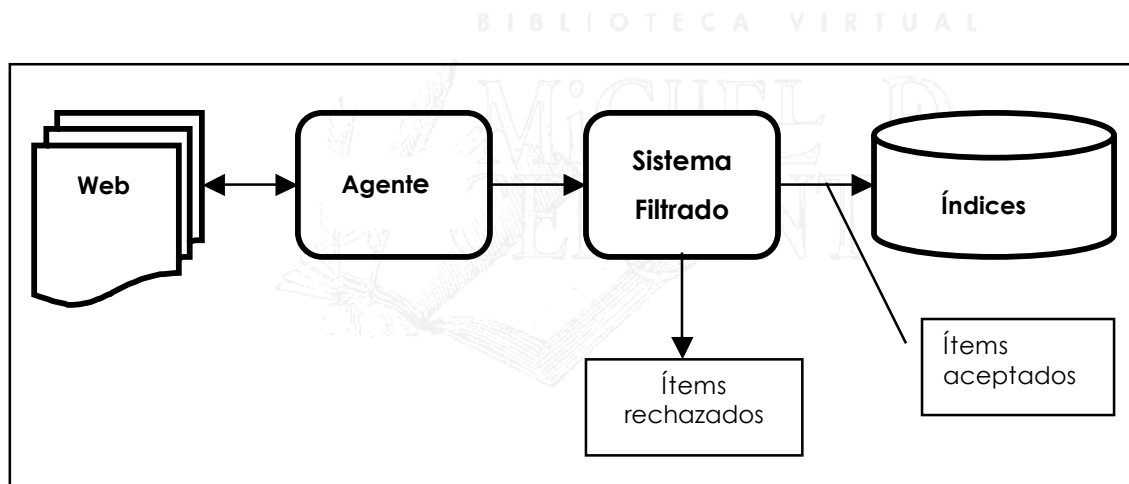


Ilustración 2.4 El proceso de construcción de un motor de búsqueda específico a partir de un filtrado de documentos. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001.

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 opina 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<sup>35</sup>. La exploración de los enlaces<sup>36</sup> insertados en una página web y, la exploración de los enlaces que apuntan hacia esa misma página, ha propiciado la creación de una nueva familia de motores de búsqueda, de la que *Google*<sup>37</sup> es su mejor exponente [BRI, 1998]. *Google* 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 hace referencia, en segundo, tercer y cuarto lugar, a los directorios, motores de búsqueda, metabuscadores y lenguajes de consulta para la web, sistemas que han sido presentados anteriormente.

En quinto lugar menciona la búsqueda de documentos multimedia, campo que se encuentra 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 (seguramente debido a que ha mejorado la capacidad de su conexión a Internet). Para este autor su desarrollo “está considerado uno de los mayores 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

---

<sup>35</sup> 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”.

<sup>36</sup> No se debe confundir “exploración de los enlaces” con la “explotación de la estructura hipertextual” que mencionaba Baeza-Yates. Este autor incluye a *Google* y a su algoritmo de alineamiento *Page Rank* en el 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*.

<sup>37</sup> *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>> [Consulta: 21 de enero de 2002].

lenguaje natural y los aún incipientes desarrollos de sistemas de recuperación de documentos en formato XML.

### **Los motores de búsqueda como paradigma de la recuperación de información en Internet.**

De la totalidad de los SRI que se han desarrollado en Internet, los motores de búsqueda son los que más se incardinan con la naturaleza dinámica del contexto de la web, siendo unos 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* [SAL, 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: "spamming", inaccesibilidad de páginas, deficientes o inexistentes descripciones de las páginas, volatilidad, etc.

Independientemente de su método de rastreo y de los posteriores criterios y algoritmos empleados para el alineamiento de los documentos, todos los motores de búsqueda parten de una situación inicial parecida: una lista de direcciones que sirve de punto de partida para el robot (o los robots). Esta similitud de condiciones iniciales propicia, ineludiblemente, una posterior comparación del resultado final, es decir, de la porción de web indexada y de la calidad de esta indexación. Otro factor que contribuye a esta serie de comparaciones es el cierto ocultismo de los métodos seguidos por cada motor en la realización de sus tareas, lo que conlleva, al igual que en el caso anterior, a la necesidad de comparar el resultado obtenido con el fin de poder apreciar cuál de esos sistemas es de uso más recomendable.

Si se asumen que de lo completa, representativa y actualizada que sea la colección de un motor de búsqueda, depende su calidad; en un directorio, en cambio, la misma reside en la capacidad de los gestores en la realización de las descripciones y en el número de estos gestores, ambos motivos 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 de un motor debe estudiarse desde dos perspectivas complementarias: la recopilación y la recuperación de información. Un motor compila de forma automática las direcciones de las páginas que van a formar parte de su índice tras realizar sobre su contenido un proceso de indización. Una vez se encuentren estos registros debidamente depositados en la base de datos del motor, los usuarios buscarán en su índice por medio de una interface de consulta, que puede ser más o menos avanzada en función del grado de desarrollo del sistema. Al módulo encargado de la recopilación de las páginas se le conoce comúnmente como robot<sup>38</sup>, “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 pueden recopilar varios millones de páginas por día, y actualizar la información recogida en los índices en períodos de tiempo extremadamente pequeños si consideramos la extensión del espacio al que nos estamos refiriendo. Por regla general, se parte de una lista inicial de direcciones de sitios web, que son visitados por el robot, y a partir de ahí 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 de indización o indexación (“indexing”), con lo cual él habla de dos módulos independientes, el “crawler” o robot y el indexador [BAE, 1999].

### **Arquitectura de un motor de búsqueda.**

La mayoría de los motores de búsqueda emplean una arquitectura de tipo robot-indexador centralizada, que se muestra en la Ilustración 2.5. A pesar de lo que puede inducir su nombre y de una amplia serie de definiciones

---

<sup>38</sup> Delgado Domínguez nos dice que a los robots se les denomina también “spiders” (que podríamos traducir como “arañas”) o “web crawlers” (una posible traducción sería “gateadores por la web” o “quienes andan a gatas por la web”). Baeza-Yates aporta otra denominación: “walkers” (“andadores”). Todos estos términos hacen referencia a un movimiento lento y continuado entre los distintos elementos (páginas en este caso) que conforman la web (que se puede traducir como “tela de araña”).

incorrectas<sup>39</sup>, el robot no se mueve por la red, ni se ejecuta sobre las máquinas remotas que visita, ya que realmente el robot funciona sobre el sistema local del motor de búsqueda y envía una serie de peticiones a los servidores web remotos (donde se 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 memoria RAM y sobre 500 Gb de espacio en disco; sólo el módulo de interrogación del índice consume más del 75% de estos recursos" [BAE, 1999].

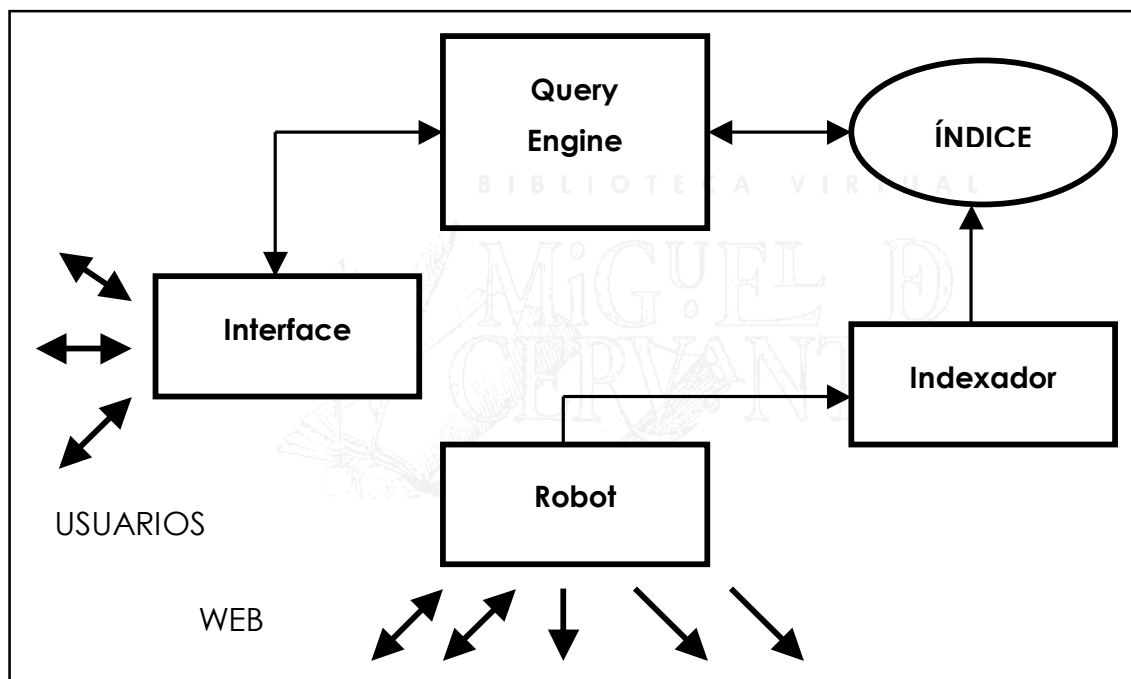


Ilustración 2.5 Arquitectura simple de un motor de búsqueda. Fuente: o a partir de un filtrado de documentos. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p.

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.

<sup>39</sup> En muchos artículos divulgativos y textos educativos se difunde la equivocada idea de que el robot "se mueve a lo largo de la web", como si tuviera vida, cuando se trata realmente de una aplicación informática que solicita una serie de transacciones a los servidores web donde se alojan las páginas analizadas.

- 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 en la web, sea el desarrollo de una serie de alternativas a este modelo de arquitectura simple, para procurar paliar estos defectos. Baeza-Yates destaca la arquitectura del sistema *Harvest*<sup>40</sup> como 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.



---

<sup>40</sup> *Harvest* significa "cosecha".

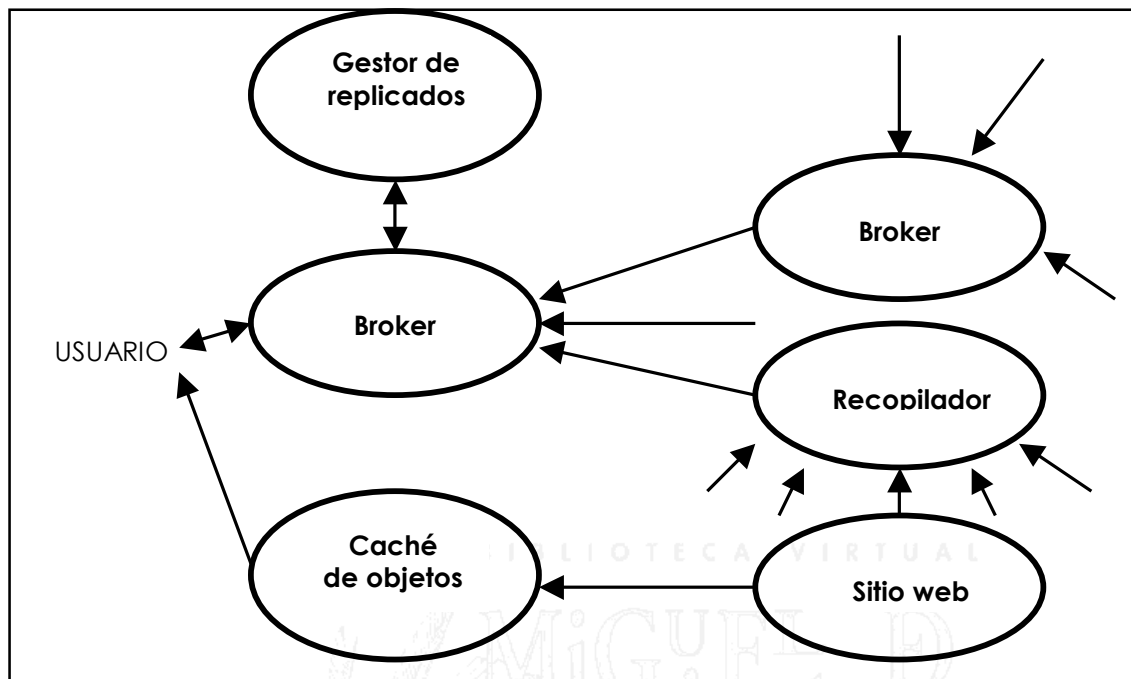


Ilustración 2.6 Arquitectura Harvest. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p.

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 considerablemente la carga de tráfico.

En último lugar, la información es recopilada 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 de ellos es el módulo que colecciona páginas y extrae de ellas toda la información necesaria para crear el índice del motor de búsqueda. El segundo de estos componentes, el *broker*, es el módulo encargado de proporcionar el mecanismo de indexación de las páginas recopiladas y la interface de consulta para los datos recopilados. 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 [BAE, 1999].

No todos los sistemas poseen iguales prestaciones en lo relacionado con la recuperación de información. La clásica y más simple interface de usuario es la típica caja de formulario web que se muestra en la Ilustración 2.7. En esa caja el usuario inserta el conjunto de términos integrantes de su ecuación de búsqueda. Algunas veces, ese formulario le permite insertar alguna restricción a la búsqueda: el idioma de las páginas a recuperar, el tipo de objeto, si se desea emplear la búsqueda por frase literal o "búsqueda exacta", etc.



The image shows a search interface for 'All the Web'. At the top left, it says 'Search All The Web, All the Time:'. To the right are links for 'Help', 'Customize', and 'Advanced Search'. Below this is a search bar with a dropdown menu set to 'Any language' and a 'Search' button. To the right of the search bar is a checkbox labeled 'Exact phrase'. At the bottom, there are links for 'Search for: Web pages | News | Pictures | Videos | MP3 files | FTP files'.

Ilustración 2.7 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 2.8), 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 sus índices (el caso más común es el título o el texto).

The image shows a screenshot of the 'All the Web' advanced search interface. At the top, it says 'Search All The Web, All the Time:' with links for 'Help', 'Customize', and 'Simple Search'. Below this is a search input field with a dropdown menu set to 'all of the words' and a 'Search' button. A secondary search bar is labeled 'Search for:' with links for 'Web pages', 'News', 'Pictures', 'Videos', 'MP3 files', and 'FTP files'. The 'Language' section allows filtering by language (set to 'Any language') and character set (set to 'Unicode (UTF-8)'). The 'Word Filters' section includes three rows for 'Should include', 'Must include', and 'Must not include', each with a text input field and a dropdown menu set to 'in the text'. The 'Domain Filters' section has 'Only Include' and 'Exclude' text input fields. The 'IP-address Filters' section has an 'Address(es) and/or range(s)' text input field. The 'Result Restrictions' section includes 'Pages updated' (set to 'anytime') and 'Document size' (set to 'exactly' and 'bytes').

Ilustración 2.8 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 ofrecidas por 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 de cada motor [WIN, 1995], [DAV, 1996], [SLO, 1996], [ZOR, 1996] y [WES, 2001].

Otra de las diferencias, algo más interna y no tan explícita, es la forma en la que 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 el motor 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 pueda encontrar 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<sup>41</sup> (cuando los términos están escritos en un orden determinado, por ejemplo, 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 de análisis del

---

<sup>41</sup> También es conocida por “búsqueda por frase literal” o “búsqueda exacta”.

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

### Los índices de los motores.

El índice "es el corazón de un motor de búsqueda" [CHA, 2001], generalmente consiste en una lista de palabras con 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* [BAE, 1992], [TRA, 1997], [RIJ, 1999], [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 2.9 Ejemplo de la estructura de un fichero inverso (tabla de la derecha). 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 estructuras de 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:

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

Tabla 2.5 Clasificación de los 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 donde se recogen todas las URL en las que aparece una palabra clave. La manera en la que se ordenan estos punteros depende de un mecanismo interno de ordenación basado, generalmente, en criterios de frecuencias o pesos 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 2.6 se presentan algunas de las diversas técnicas empleadas.

<b>Conversión de texto a minúsculas</b>	Se convierten todas las palabras a caracteres en minúscula, reduciendo así el número de entradas para un mismo término (Puerto – puerto)
<b>Stemming</b>	Aislamiento de la base de la palabra (por ejemplo, comprensión y comprensivo se reducirían a "compren"), reduciéndose así el número de entradas en el índice
<b>Supresión de las palabras vacías<sup>42</sup></b>	Se suprimen del índice todas aquellas palabras por las que no tiene sentido recuperar información (artículos, preposiciones, adjetivos o interjecciones, por ejemplo)
<b>Comprensión de textos</b>	Técnicas de compactación del tamaño del fichero

Tabla 2.6 Técnicas empleadas 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, Massachusetts: Kluwer Academic Publishers, 2001.

---

<sup>42</sup> Las *palabras vacías* son palabras sin valor de discriminación. En Inglés se conocen como "stopwords".

Las tres primeras técnicas consiguen reducir hasta un 30% del tamaño del índice y la cuarta llega, en algunos casos, a comprimir hasta un 90% el tamaño [BAE, 1999]. Resulta clara 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 autores y profesionales. El uso de índices con *granularidad* más consistente implica menor tamaño de índice y menos punteros, lo que favorece una simplificación de la estructura de datos. Baeza-Yates expone como ejemplo la idea que Glimpse emplea para el sistema *Harvest*: "las preguntas del usuario son resueltas por medio de ficheros inversos, que proporcionan una lista de bloques lógicos que son leídos de forma secuencial, ya que su tamaño es menor" [BAE, 1999].

### Tipos de robots.

Junto a los robots de carácter general, existen otras modalidades de estos sistemas, más específicas:

- *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) [BAE, 1999].

La extensión de la web genera problemas con el refresco de los índices de los motores, ya que transcurre un necesario período de tiempo entre dos análisis del mismo recurso, intervalo que varía mucho según el motor. Analizando esta problemática, Baeza-Yates esboza una analogía entre el índice y las estrellas del cielo: "lo que vemos en un índice jamás ha existido, ya que la luz 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 [BAE, 1999].

Por ello, algunos motores muestran en la respuesta la fecha de indización de la página. Baeza-Yates estima que alrededor del 9% de los enlaces almacenados son inválidos y Nottes cifra este porcentaje en el rango comprendido entre el uno y el trece por ciento, según el motor analizado [NOT, 2000c]. Esta cifra es objeto de estudio por varios analistas y precisa de continuas revisiones ante la naturaleza dinámica de la web, constituyendo uno de los criterios más significativos a la hora de ponderar la calidad de un motor de búsqueda frente a otro. Aguilar González resume en una tabla algunas de las principales características de rastreo:

Característica de rastreo	No	Si
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 <sup>43</sup>	Alta Vista, Inktomi, FAST

Tabla 2.7 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]

<sup>43</sup> Google cuenta con un "sistema de publicidad autoadministrada" que en la práctica es una inclusión pagada de referencias a sitios web. No obstante, este motor muestra estos enlaces de forma separada al resto de los documentos recuperados.

### 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 búsqueda implemente 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 el número 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* o *WISEnut*) [AGU, 2002].

Un ejemplo representativo del comportamiento de un motor clásico a la hora de indizar 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 este motor, la función de indización la llevan a cabo dos módulos: el *indexador* y el *clasificador*. El primero lee las páginas procedentes del *storeserver*<sup>44</sup>, descomprime los documentos y selecciona los términos incluidos en los mismos.

Cada documento se convierte 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].

La Tabla 2.8 resume algunas de las principales características de la indización y los motores que las implementan.

---

<sup>44</sup> Este módulo es el repositorio de la relación inicial de páginas que debe analizar el robot. En el mismo se almacena, en formato comprimido, el contenido de las mismas.

Características de la indexación	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 2.8 Características de la indexación 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 terminará por no recurrir a este motor de búsqueda. Si, en cambio, el motor discrimina ese grado de relación, el usuario encontrará entre los primeros documentos a los más relevantes con la temática objeto 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 clásico 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” [BAE, 1999].

El segundo grupo de algoritmos aporta una de las mayores diferencias conceptuales sobre el alineamiento: el uso de los enlaces de cada 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 tratados en 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 al grupo original de respuestas.
- *HITS*<sup>45</sup>: 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 por el autor de calidad en relación a un tema). Esta idea asume que cuando alguien

---

<sup>45</sup> HITS: Hypertext Induced Topic Search. Se puede traducir al Español como “Búsqueda de temas hipertextual inducida”.l

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 [BAE, 1999].

- *PageRank* asume que el número de enlaces que una página proporciona tiene mucho que ver con la calidad de la misma, es por ello que 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 expresión:  $PR(A) = (1-d) + d(PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$ ". 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 aburriría, 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, en cambio, 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. 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 los motores 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 descalificando 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 de estos 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 de usuario del motor de búsqueda. En algunos casos esa 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 y aún 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. A modo de ejemplo de esas malas prácticas se puede citar el conocido 'spamming'" [GRA, 2000]. 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 adicional nivel 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.

## Tablas e Ilustraciones.

Tabla 2.1 Diferencias entre recuperación de datos y recuperación de información. Fuente: Rijsbergen, C.J. <i>Information Retrieval</i> . [En línea]. Glasgow, University, 1999. < <a href="http://www.dcs.gla.ac.uk/~iain/keith/">http://www.dcs.gla.ac.uk/~iain/keith/</a> > [Consulta: 21 de octubre de 2001].....	14
Ilustración 2.1 Esquema simple de un SRI. Fuente Salton , G. and Mc Gill, M.J. <i>Introduction to Modern Information Retrieval</i> . New York: Mc Graw-Hill Computer Series, 1983.....	17
Ilustración 2.2 Esquema avanzado de un SRI. Fuente Salton , G. and Mc Gill, M.J. <i>Introduction to Modern Information Retrieval</i> . New York: Mc Graw-Hill Computer Series, 1983.....	18
Tabla 2.2 Clasificación de los Modelos de Recuperación de Información según Dominich. Fuente: Dominich, S. 'A unified mathematical definition of classical information retrieval'. <i>Journal of the American Society for Information Science</i> , 51 (7), 2000. p. 614-624. ....	20

Tabla 2.3 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. ....	22
Ilustración 2.3 Sección de la primera página web diseñada por Kunz. Esta página sigue activa en la dirección < <a href="http://www.slac.stanford.edu/spires/hep/">http://www.slac.stanford.edu/spires/hep/</a> > de la Universidad de Stanford. ....	24
Tabla 2.4 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. < <a href="http://dmi.uib.es/people/adelaide/tice/modul6/memfin.pdf">http://dmi.uib.es/people/adelaide/tice/modul6/memfin.pdf</a> > [Consulta: 18 de septiembre de 2001] .....	27
Ilustración 2.4 El 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.....	32
Ilustración 2.5 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.....	36
Ilustración 2.6 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.....	38
Ilustración 2.7 Formulario de búsqueda simple del motor All the Web. Fuente: < <a href="http://www.alltheweb.com">http://www.alltheweb.com</a> > .....	39
Ilustración 2.8 Sección del formulario de búsqueda avanzada del motor All the Web. Fuente: < <a href="http://www.alltheweb.com/advanced">http://www.alltheweb.com/advanced</a> > .....	40
Ilustración 2.9 Ejemplo de la estructura de un fichero inverso (tabla de la derecha). Fuente: Rijsbergen, C.J. Information Retrieval. [En línea]. Glasgow, University, 1999. < <a href="http://www.dcs.gla.ac.uk/~iain/keith/">http://www.dcs.gla.ac.uk/~iain/keith/</a> > [Consulta: 21 de octubre de 2001].....	42
Tabla 2.5 Clasificación de los ficheros inversos a partir de la <i>granularidad</i> de su índice. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001. ....	43
Tabla 2.6 Técnicas empleadas 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, Massachusetts: Kluwer Academic Publishers, 2001. ....	43
Tabla 2.7 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. < <a href="http://www.geocities.com/motoresdebusqueda/crawlers.html">http://www.geocities.com/motoresdebusqueda/crawlers.html</a> > [Consulta: 3 de abril de 2002] .....	45
Tabla 2.8 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. < <a href="http://www.geocities.com/motoresdebusqueda/crawlers.html">http://www.geocities.com/motoresdebusqueda/crawlers.html</a> > [Consulta: 3 de abril de 2002] .....	48