RECUPERACIÓN DE INFORMACIÓN: MODELOS, SISTEMAS Y EVALUACIÓN.

Francisco Javier Martínez Méndez

'Since the 1940s the problem of information storage and retrieval has attracted increasing attention'.

Keith Rijsbergen.

Information Retrieval, chapter 1. 1979

http://www.dcs.gla.ac.uk/Keith/Chapter.1/Ch.1.html

MARTÍNEZ MÉNDEZ, Francisco Javier

Recuperación de información: modelos, sistemas y evaluación / Francisco Javier Martínez Méndez. – Murcia: KIOSKO JMC, 2004.

106 p.

ISBN:

1. Recuperación de Información. I. Título.

025.04.03

RECUPERACIÓN DE INFORMACIÓN: MODELOS, SISTEMAS Y EVALUACIÓN.

ISBN: 84-932537-7-4

Depósito legal: MU-729-2004

Edita, Distribuye y Vende: EL KIOSKO JMC

Imprime: EL KIOSKO JMC

RECUPERACIÓN DE INFORMACIÓN: MODELOS, SISTEMAS Y EVALUACIÓN.

Prólogo.

Si considerásemos la capacidad del ser humano para almacenar información en su cerebro, nos daríamos cuenta de hasta que extremos resulta ridículo construir dispositivos artificiales para el almacenamiento de esa información. Claro está que esta reflexión encierra su propia paradoja, ya que a pesar de la gigantesca capacidad de nuestro cerebro, se trata a su vez de un órgano en extremo complejo del que en estos albores del siglo XXI comenzamos apenas a comprender su funcionamiento. Por otro lado, las computadoras son simples máquinas -por complejas que nos parezcan- y acarrean un proceso, sino inverso estrictamente hablando, contrario; de la simplicidad, casi infantil en los inicios de su aparición a la actualidad y de ahora en adelante con el desarrollo de la nanotecnología, hasta alcanzar un régimen de complejidad en estos momento difícil de adivinar -hoy ya, señalan los constructores de computadores, resulta casi imposible seguir un error dentro de la lógica de un microprocesador-.

Con este panorama, ¿de qué disponemos realmente?, pues de un cerebro al que no podemos explotar más allá de un pequeño porcentaje de sus posibilidades -desconocemos las estructura del sistema operativo que lo rige- y de máquinas que no resuelven más que un pequeño porcentaje de nuestras necesidades.

Con estas perspectivas tenemos que abordar el almacenamiento de datos en las computadoras, ocioso resulta señalar que el objeto de este almacenamiento es el poder encontrar eso datos cuando a cualquier persona le surja una necesidad que pueda ser cubierta con uno o un conjunto de esos datos previamente almacenados.

El desarrollo de las bases de datos, en un principio, pareció la solución a todos nuestros problemas gracias a su capacidad ilimitada de almacenar datos y sus fáciles mecanismos de extracción. Y, ¿cuál es la causa de que estas bases de datos no resolviesen el problema?, si bien si han resuelto algunos. Todo responde a una cuestión crucial de reduccionismo del problema, considerar que una parte de la realidad circundante se puede reducir a una ristra de caracteres sintácticamente simples y sin ninguna ambigüedad semántica, es querer estar alejado de esa realidad.

Por todo ello, se inició el desarrollo de sistemas algo más complejos, pero aun ciertamente alejados de la solución del problema, con el fin de ir acercándose a una meta más consecuente con las necesidades de los seres humanos. Estos sistemas son los conocidos como *Sistemas de Recuperación de Información*. Desde su creación se ha sido consciente de esa distancia entre realidad y solución propuesta entre necesidad y respuesta, de ahí que desde sus inicios se idearan de

forma paralela mecanismos para medir esa distancia: los métodos de evaluación de los sistemas de recuperación de información.

En el momento actual existe un ingenio: la *web,* sistema con una dinámica de funcionamiento y estructura para la cual muchas de las cosas desarrolladas con anterioridad están resultando estériles, entre otras las medidas de evaluación de su comportamiento. Es por eso que los investigadores de todo el mundo se encuentran en plena efervescencia creativa presentando nuevas orientaciones, añadiendo, en definitiva, más complejidad al problema de reducir la distancia, de aproximar las necesidades de los usuarios -que emergen de la mente- con la respuesta de los ingenios artificiales.

Pues bien, el presente trabajo fruto de una dilatada investigación avalada por no una menos dilatada experiencia del autor, enfatiza en todos estos elementos de una forma clara y eficaz, y con un estilo sencillo pero completo. No podemos presumir, por desgracia, de abundancia de textos en español sobre el tema, pero es que, además, sobre las medidas de evaluación ni siquiera era fácil encontrar hasta ahora una monografía más o menos dedicada a este tema.

Por ello se trata de un texto oportuno y necesario para tener conciencia real de la actual problemática de los sistemas de recuperación de información: son ingenios inacabados.

Murcia, Marzo de 2004

Dr. José Vicente Rodríguez Muñoz

Facultad de Comunicación y Documentación

Índice General.

1 La Recuperación y los Sistemas de Recuperación de Información.

.

Este capítulo representa una presentación del concepto de recuperación de información, y del conjunto de diferencias que posee con otras aplicaciones de la Informática en lo relacionado con la gestión y recuperación de datos. Al mismo tiempo se exponen los distintos modelos sobre los que se basan los sistemas que permiten la recuperación de información.

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]. En realidad, 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*[1] (SRI en adelante).

El primer grupo de definiciones están muy influenciadas por la tecnología informática, cuya evolución ha inducido a considerar sinónimos ambos conceptos, llegándose a olvidar que se puede recuperar información sin procedimientos informáticos (aunque no es lo más común hoy en día). Aun así, el frecuente y necesario empleo de una tecnología no debe sustituir el adecuado uso de los conceptos terminológicos. Un claro ejemplo de este desacierto es el *Glosario de la Asociación de*

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

Bibliotecarios Americanos, que define el término "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[2]. Igualmente, el *Diccionario Mac Millan de Tecnología de la Información* considera a la recuperación de información como las "técnicas empleadas para almacenar y buscar grandes cantidades de datos y ponerlos a disposición de los usuarios" [LON, 1989]

Un segundo grupo de autores fijan diferencias. Meadow piensa que la recuperación de la información es "una disciplina que involucra la localización de una determinada información dentro de un almacén de información o base de datos" [MEA, 1992]. Este autor, implícitamente, establece que la recuperación de información se encuentra asociada 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: "una típica tarea de la recuperación de información es *traer* documentos relevantes desde una gran archivo en respuesta a una pregunta formulada y ordenarlos de acuerdo con su *relevancia*" [PER, 2000].

Del mismo modo, Grossman y Frieder indican que recuperar información es "encontrar documentos relevantes, no encontrar simples correspondencias a unos patrones de bits" [GRO, 1998]. Meadow considera 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 similar opinión es Blair, quien dedica gran parte de la presentación de su libro '*Language and representation in information retrieval*' a asentar las diferencias entre *data retrieval* e *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 probablemente contendrá lo deseado, con un evidente factor de indeterminación.
2. Según la relación entre el requerimiento al sistema y la satisfacción de usuario, la recuperación de datos es *determinista* y la

[2] Este Glosario indica que "document retrieval" es un término sinónimo de "information retrieval".

recuperación de información es *posibilista*, debido al nivel de incertidumbre presente en la respuesta.

3. Éxito de la búsqueda. En recuperación de datos el criterio a emplear es la exactitud de lo encontrado, mientras que en recuperación de información, el criterio de valor es el grado en el que la respuesta satisface las necesidades de información del usuario, es decir, su percepción personal de utilidad.

Tramullas Saz destaca un aspecto importante de las reflexiones de Blair, la importancia, en ocasiones ignorada, que tiene el factor de 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 "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 siguiente tabla compendia las diferencias fundamentales existentes entre *recuperación de datos* y *recuperación de información* a juicio de Rijsbergen [RIJ, 1999]:

	Recuperación de datos	**Recuperación de información**
Acierto	Exacto	Parcial, el mejor
Inferencia	Algebraica	Inductiva
Modelo	Determinístico	Posibilístico
Lenguaje de consulta	Fuertemente Estructurado	Estructurado o Natural
Especificación consulta	Precisa	Imprecisa
Error en la respuesta	Sensible	Insensible

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

Baeza-Yates plantea las diferencias entre ambos tipos de recuperación con argumentos quizá algo menos abstractos que los anteriormente empleados por otros autores, 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 define 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*". En la solución de este problema se identifican dos grandes etapas:

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

Baeza-Yates se preocupa especialmente de las estructuras de datos y métodos de acceso a los mismos [BAE, 1992], [BAE, 1999], siendo este autor una verdadera referencia en esta materia. Curiosamente, a la hora de definir la recuperación de información, en lugar de proponer una definición propia, emplea la elaborada por Salton: "la recuperación de la información tiene que ver con la representación, almacenamiento, organización y acceso a los ítem de información" [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].

El tercer grupo de autores emplea la definición formulada por Salton (base de la mayoría de definiciones de a bibliografía especializada), añadiendo el rasgo diferenciador en que estos autores no profundizan en escrutar las diferencias entre "recuperación de datos" y "recuperación de información", bien por no ser objeto de sus trabajos o 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 Saz 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 se distinguen porque eluden definir la recuperación de la información. Su máximo exponente a Chowdhury, quien simplemente dedica el primer párrafo de su libro

'Introduction to modern information retrieval' a señalar que "el término recuperación de la información fue acuñado en 1952 y fue ganando popularidad en la comunidad científica de 1961 en adelante[3]", mostrando después los propósitos, funciones y componentes de los SRI [CHO, 1999]. Otro autor perteneciente a esta corriente es Korfhage, quien se centra en el almacenamiento y recuperación de la información, considerando a estos procesos como las dos caras de una moneda. Para este autor, "un usuario de un sistema de información lo utiliza de dos formas posibles: para almacenar información en anticipación de una futura necesidad, y para encontrar información en respuesta una necesidad" [KOR, 1997].

Sistemas de Recuperación de Información.

Tomando como base de partida la definición propuesta por Salton y uniéndola a las aportaciones de Rijsbergen [RIJ, 1999], correspondería ahora (siguiendo la opinión de Baeza-Yates [BAE, 1999]), elegir el mejor modelo para el diseño de un SRI, aunque antes resulta necesario definir adecuadamente "sistema de recuperación de información".

Vista funcional de un SRI.

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

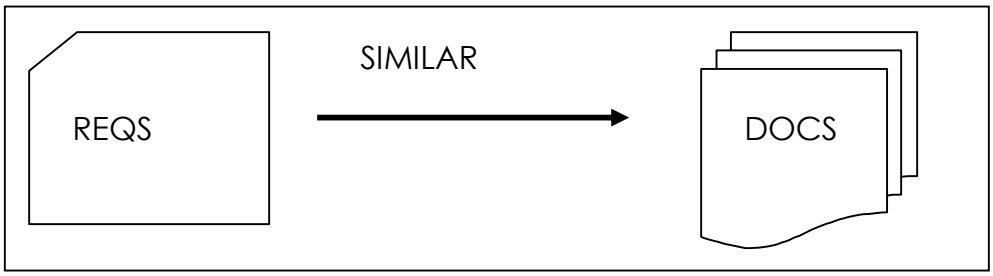


Ilustración 1.1 Esquema simple de un SRI. Fuente Salton , G. and Mc Gill, M.J. *Introduction to Modern Information Retrieval*. New York: Mc Graw-Hill Computer Series, 1983.

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

Es el mismo Salton quien reconoce que en la práctica este esquema resulta muy simple y precisa ampliación, "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]

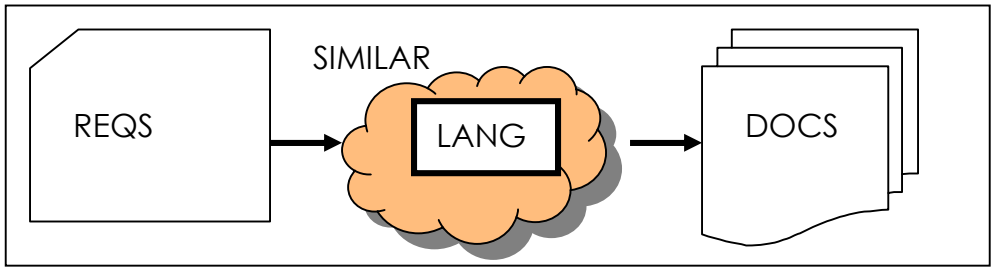


Ilustración 1.2 Esquema avanzado de un SRI. Fuente Salton , G. and Mc Gill, M.J. *Introduction to Modern Information Retrieval*. New York: Mc Graw-Hill Computer Series, 1983.

En la Ilustración 1.2, 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 que sea adecuada para compararlas con las preguntas de los usuarios.
4. Analizar las preguntas de los usuarios y representarlas de una forma que sea adecuada para compararlas con las representaciones de los documentos de la base de datos.
5. Realizar la correspondencia entre la representación de la búsqueda y los documentos almacenados en la base de datos.
6. Recuperar la información relevante
7. Realizar los ajustes necesarios en el sistema basados en la retroalimentación con los usuarios

Evolución de los SRI.

Muchos autores presentan la evolución de estos sistemas, pero quien mejor simplifica este progreso es Baeza-Yates, destacando tres fases fundamentales [BAE, 1999]:

1. *Desarrollos iniciales*. Ya existían métodos de recuperación de información en las antiguas colecciones de papiros. Otro ejemplo clásico que se ha venido utilizando es la *tabla de contenidos* de un libro, sustituida por otras estructuras más complejas a medida que ha crecido el volumen de información. La evolución lógica de la tabla de contenidos es el *índice*, estructura que aún constituye el núcleo de los SRI actuales.
2. *Recuperación de información en las bibliotecas*. Fueron las primeras instituciones en adoptar estos sistemas. Originalmente fueron desarrollados por ellas mismas y posteriormente se ha creado un mercado informático altamente especializado, en el que participan empresas e instituciones.
3. *La World Wide Web*. La evolución lógica de los SRI ha sido hacia la web, donde han encontrado una alta aplicación práctica y un aumento del número de usuarios, especialmente en el campo de los directorios y motores de búsqueda[4]. El alto grado de consolidación de la web está siendo favorecido por el vertiginoso abaratamiento de la tecnología informática, por el espectacular o desarrollo de las telecomunicaciones y por la facilidad de publicación de cualquier documento que un autor considere interesante, sin tener que pasar por el filtro de los tradicionales círculos editoriales.

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

[4] Estos sistemas se presentan posteriormente en el apartado dedicado a la *recuperación de la información en la web*, dentro de este mismo capítulo.

Modelos para la recuperación de información.

El diseño de un SRI se realiza bajo un modelo, donde queda definido "cómo se obtienen las representaciones de los documentos y de la consulta, la estrategia para evaluar la *relevancia* de un documento respecto a una consulta y los métodos para establecer la importancia (orden) de los documentos de salida" [VIL, 1997].

Existen varias propuestas de clasificación de modelos, una de las síntesis más completas la realiza Dominich en cinco grupos [DOM, 2000]:

Modelo	**Descripción**
Modelos clásicos	Incluye los tres más comúnmente citados: *booleano*, *espacio vectorial* y *probabilístico*.
Modelos alternativos	Están basados en la Lógica Fuzzy
Modelos lógicos	Basados en la Lógica Formal. La recuperación de información es un proceso inferencial.
Modelos basados en la interactividad	Incluyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la *relevancia* de los documentos recuperados [SAL, 1989]
Modelos basados en la Inteligencia Artificial[5]	Bases de conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural.

Tabla 1.2 Clasificación de los Modelos de Recuperación de Información según Dominich. Fuente: Dominich, S. 'A unified mathematical definition of classical information retrieval'. Journal of the American Society for Information Science, 51 (7), 2000. p. 614-624.

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

Este mismo autor divide a los modelos basados en la recuperación en dos grupos: clásicos y estructurados. En el primero de ellos incluye a los modelos booleano, espacio vectorial y probabilístico. Posteriormente, presenta una serie de paradigmas alternativos a cada modelo: teoría de conjuntos (conjuntos difusos y booleano extendido), algebraicos (vector generalizado, indización por semántica latente y redes neuronales), y por último, probabilísticos (redes de inferencia y redes de conocimiento); los

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

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:

Modalidad	Vista lógica de los documentos: Términos Índice	Vista lógica de los documentos: Texto Completo	Vista lógica de los documentos: Texto Completo y Estructura
Recuperación	Clásicos Conjuntos teóricos Algebraicos Probabilísticos	Clásicos Conjuntos teóricos Algebraicos Probabilísticas	Estructurados
Navegación	Estructura plana	Estructura plana Hipertexto	Estructura guiada Hipertexto

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

Tanto Baeza-Yates [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[6] [GRO, 1998], para conocer las alternativas a los modelos clásicos.

Modelo del Espacio Vectorial.

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

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

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

Si nuestro SRI contiene los siguientes cuatro documentos:

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

D2: el caudal de un río asciende en Invierno

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

D4: si un río es navegable, es porque tiene mucho caudal

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

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

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

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

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

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

de 2 es 200% más importante que una frecuencia de 1, y la diferencia aritmética es sólo de una unidad).

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

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

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

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

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

la matriz de nuestro ejemplo, proporcionaría la siguiente matriz de pesos (a la que añadimos una fila con el vector pregunta).

Cálculo de frecuencias inversas

Idf (río) = Log (4/4) = log (1) = 0

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

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

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

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

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

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

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

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

Matriz ***tf-idf.***

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

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

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

Cálculo de similitudes

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

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

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

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

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

2 La Recuperación de Información en la world wide web.

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

Recuperar información en la web.

Hu recuerda que "el primer motor de búsqueda desarrollado en la red Internet fue ARCHIE[7], creado en 1990, aunque no fue hasta la creación del primer navegador, *Mosaic*[8], que propició el crecimiento de los documentos y la gestión de información multimedia cuando se expandió el uso de estos sistemas" [HU, 2001].

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

Breve perspectiva histórica de la web.

El nacimiento y crecimiento exponencial de la web es un hecho suficientemente conocido y cuyo alcance ha traspasado los límites de la comunidad científica hasta llegar a todo el entorno social. En agosto de 1991, Paul F. Kunz, físico de la *Universidad de Stanford* leyó una noticia en la que difundía la invención de la *World Wide Web* y contactó con Tim

[7] 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.

[8] *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.

Berners-Lee, becario británico del CERN[9]. 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.
- Integración de un amplio rango de tecnologías y distintos tipos de documentos.
- Proporcionar una herramienta para los documentos desde cualquier lugar de Internet y por cualquier individuo que este navegando dentro de este almacén, permitiendo accesos simultáneos.

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

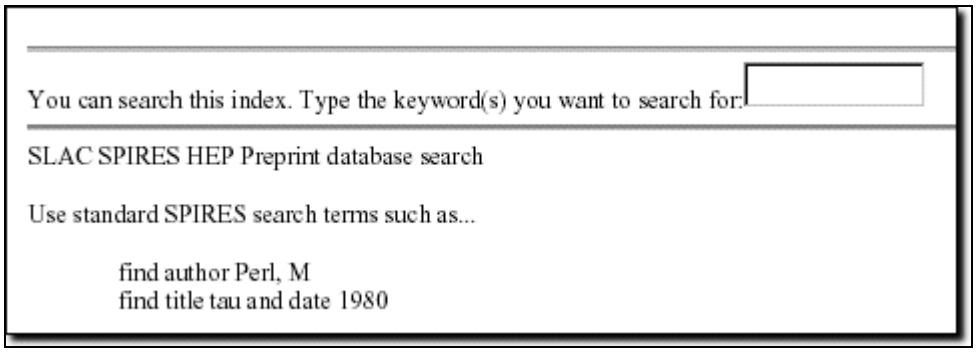


Ilustración 2.1 Sección de la primera página web diseñada por Kunz. Esta página sigue activa en la dirección <http://www.slac.stanford.edu/spires/hep/> de la Universidad de Stanford.

El posterior desarrollo de *Mosaic* aumentó el número de usuarios que accedía a este novedoso sistema, crecimiento que continuó con el

[9] 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").

desarrollo de nuevos navegadores: N*etscape* e *Internet Explorer*[10]. A principios del año 1995 se formó la organización *Consorcio World Wide Web*[11] (o *W3C*), que está bajo la dirección del fundador de la Web

Métodos de recuperación de información en la web.

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

Al mismo tiempo, en la web surgen nuevos problemas, tales como los populares fenómenos denominados *cloaking*[12]*, links farms*[13] y *guest spamming*[14]*,* o los vinculados con el enorme tamaño del índice de estos SRI, que poco a poco llega a alcanzar magnitudes impresionantes, muy difíciles de gestionar adecuadamente con los modelos tradicionales.

Baeza-Yates afirma que hay básicamente tres formas de buscar información en la web: "dos de ellas son bien conocidas y frecuentemente usadas. La primera es hacer uso de los *motores de búsqueda*[15], 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*[16]*,* sistemas

---

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

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

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

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

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

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

[16] Empleamos el término "directorio" como traducción del término "web directory". En términos coloquiales se utiliza también la palabra "índice", aunque esta última

que clasifican documentos web seleccionados por materia y que nos permiten navegar por sus secciones o buscar en sus índices. La tercera, que no está del todo disponible actualmente, es buscar en la web a través de la *explotación de su estructura hipertextual* (de los enlaces de las páginas web[17])" [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[18]), o bien un motor también permite la búsqueda por categorías, como si fuera un directorio (*Microsoft Network*, por ejemplo). Estas situaciones no contribuyen a superar ese grado de confusión.

Los directorios son aplicaciones controladas por humanos que manejan grandes bases de datos con direcciones de páginas, títulos, descripciones, etc. Estas bases de datos son alimentadas cuando sus administradores revisan las direcciones que les son enviadas para luego ir clasificándolas en subdirectorios de forma temática. Los directorios más amplios cuentan con cientos de trabajadores y colaboradores revisando nuevas páginas para ir ingresándolas en sus bases de datos. Los directorios están "organizados en categorías temáticas, que se organizan jerárquicamente en un árbol de materias de información que permite el hojeo de los recursos descendiendo desde los temas más generales a 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 referenciados a través de dos sistemas: navegación a través de la estructura de las categorías temáticas y búsquedas por palabras claves sobre el conjunto de referencias contenidas en el índice. El directorio más grande y famoso es *Yahoo!*, aunque existen otros bastante conocidos: *Dmoz* (un directorio alimentado por miles de colaboradores), *Looksmart*,

---

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.

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

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

Infospace e *Hispavista*. Con mucha diferencia, el más utilizado es *Yahoo!*. Los directorios son más usados que los motores de búsqueda especialmente cuando "no se conoce exactamente el objetivo de la búsqueda" [MAN, 2002], ya que resulta difícil acertar con los términos de búsqueda adecuados.

Los motores de búsqueda son robustas aplicaciones que manejan también grandes bases de datos de referencias a páginas web recopiladas por medio de un proceso automático, sin intervención humana. Uno o varios agentes de búsqueda recorren la web, a partir de una relación de direcciones inicial y recopilan nuevas direcciones generando una serie de etiquetas que permiten su indexación y almacenamiento en la base de datos. Un motor no cuenta con subcategorías como los directorios, sino con avanzados algoritmos de búsqueda que analizan las páginas que tienen en su base de datos y proporcionan el resultado más adecuado a una búsqueda. También almacenan direcciones que les son remitidas por los usuarios[19]. Entre los motores más populares destacan *Altavista*, *Lycos*, *Alltheweb*, *Hotbot*, *Overture, Askjeeves, Direct Hit*, *Google*, *Microsoft Network*, *Terra* y *WISEnut,* entre otros.

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

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

Delgado Domínguez resume en la tabla 2.1 las características básicas de estos dos métodos de recuperación de información en la web. Es oportuno puntualizar que el razonamiento que lleva a la autora a considerar un directorio más preciso que un motor, se basa, sin duda

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

alguna, en la fiabilidad de la descripción del registro, realizada manualmente de forma detallada y ajustada, entendiendo en este caso *precisión* como *ajuste* o *correspondencia* de la descripción realizada con el contenido de la página referenciada, en lugar de la acepción del mismo término empleada para medir el acierto de una operación de búsqueda[20]. 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*[21] 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 desarrollados pueden entenderse como extensiones del lenguaje SQL[22] 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

---

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

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

[22] SQL: Structured Query Language.

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, *metabuscadores*[23] y técnicas de *filtrado de información*[24] [CHA, 2001].

Los *metabuscadores* son sistemas desarrollados para mitigar el problema de tener que acceder a varios motores de búsqueda con el fin de recuperar una información más completa sobre un tema, siendo estos mismos sistemas los que se encargan de efectuarlos por el usuario. Un metabuscador colecciona las respuestas recibidas y las unifica, "la 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 y crawlers[25]) 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].

---

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

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

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

Estos sistemas se diferencian en la manera en que llevan a cabo el alineamiento de los resultados[26] en el conjunto unificado[27], y cómo de bien traducen estos sistemas la pregunta formulada por el usuario a los lenguajes específicos de interrogación de sus sistemas fuentes, ya que el lenguaje común a todos será más o menos reducido. Algunos metabuscadores se instalan como cliente en entorno local (*Webcompass* o *Copernic*, por ejemplo), o bien se consultan en línea (*Buscopio*, por ejemplo). Otra diferencia sustancial existente entre estos sistemas es la presentación de los resultados, "los llegan a clasificar en dos tipos, los multi buscadores y los meta buscadores: los multi buscadores ejecutan la consulta contra varios motores de forma simultánea y presentan los resultados sin más organización que la derivada de la velocidad de respuesta de cada motor (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]. Otros metabuscadores muestran los resultados en diferentes ventanas, correspondiendo cada una de ellas a una fuente distinta (*Oneseek* o *Proteus*, por ejemplo).

Generalmente, el resultado obtenido con un metabuscador no es todo el conjunto de páginas sobre una materia que se almacenan en las fuentes del metabuscador, ya que suele limitarse el número de documentos recuperados de cada fuente. Aún así, "el resultado de un metabuscador suele ser más relevante en su conjunto" [BAE, 1999]. Puede sorprender esta limitación, pero no debemos olvidar uno de los elementos más considerados en las evaluaciones de los SRI, el *tiempo de respuesta del sistema* [LAN, 1993]. Si un metabuscador devolviera todas las referencias de todos los motores y directorios fuentes en relación con la materia objeto de una búsqueda, el tiempo de respuesta del sistema alcanzaría valores que seguramente alejarían a los usuarios del metabuscador por excesivo. Por ello resulta necesario establecer un número límite de documentos recuperados por motor, con el fin de que el tiempo de respuesta, que de por sí, ya sería siempre mayor que el precisado por un único motor, no aumente excesivamente.

Actualmente, se encuentra en desarrollo una nueva generación de metabuscadores, destacando *Inquirus* como prototipo que emplea "búsqueda de términos en contexto y análisis de páginas para una más eficiente y mejor búsqueda en la web, también permite el uso de los operadores booleanos" [INQ, 2002]. Este sistema muestra los resultados

---

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

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

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

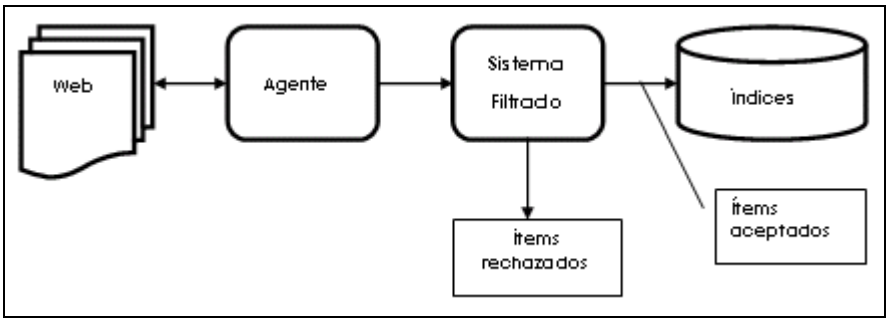


Ilustración 2.1 Proceso de construcción de un motor de búsqueda específico a partir de un filtrado de documentos. Fuente: Chang, G. et al. Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Academic Publishers, 2001.

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

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

Hu considera que "los enlaces establecidos entre las páginas web pueden resultar de tremenda utilidad como fuentes de información para los indicadores" [HU, 2001]. El hecho de que el autor de una página web introduzca un enlace en la misma, representa, implícitamente, un respaldo para la página enlazada[28]. La exploración de los enlaces[29] insertados en una página web y la exploración de los enlaces que apuntan hacia esa misma página, ha respaldado la creación de una nueva familia de motores de búsqueda, de la que *Google*[30] 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 cita en segundo, tercer y cuarto lugar, a los directorios, motores de búsqueda, metabuscadores y lenguajes de consulta para la web, sistemas todos presentados anteriormente. En quinto lugar menciona la búsqueda de documentos multimedia, área en plena expansión, máxime cuando cada vez es mayor el número de documentos de esta naturaleza en la web y el número de usuarios que los demandan (gracias a las mejores conexiones a Internet). Para este autor, este desarrollo "está considerado uno de 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 lenguaje natural y los aún incipientes desarrollos de sistemas de recuperación de documentos en formato XML.

---

[28] 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".

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

[30] *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].

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

De la totalidad de los SRI que se han desarrollado en la web, los motores de búsqueda son los que más se incardinan con su naturaleza dinámica, siendo sistemas de evolución paralela al crecimiento de la web y al aumento del número de usuarios. Constituyen además uno de los desarrollos más consolidados de las técnicas de *Indización Automática* [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 y que no tenían lugar en los tradicionales SRI.

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

Si se asume que de lo completa, representativa y actualizada que sea la colección de un motor de búsqueda, depende su calidad; en un directorio, en cambio, esta calidad depende de la capacidad de los indicadores y en su número, motivos ambos más relacionados con capacidades presupuestarias que con prestaciones tecnológicas. En cambio, los motores representan un claro ejemplo de la aplicación de las técnicas de recuperación de información a la resolución de un reto, tan antiguo como moderno, en el campo de la Información y la Documentación: disponer en un índice las referencias a la mayor parte de los documentos existentes.

Funcionamiento de un motor de búsqueda.

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

El módulo que realiza esta recopilación es conocido comúnmente como *robot*[31], es un programa que "rastrea la estructura hipertextual de la

[31] Delgado Domínguez comenta que a los robots se les denomina también "spiders" ("arañas") o "web crawlers" ("quienes andan a gatas por la web").

web, recogiendo información sobre las páginas que encuentra. Esa información se indiza y se introduce en una base de datos que será explorada posteriormente utilizando un motor de búsqueda" [DEL, 1998]. Estos robots recopilan varios millones de páginas por día, y actualizan la información depositada en los índices en períodos de tiempo muy pequeños (si se considera la extensión del espacio al que nos estamos refiriendo). Parten generalmente de una lista inicial de direcciones de sitios web, que son visitados y a partir de los cuales cada robot rastrea a su manera la web, de ahí que la información almacenada en cada base de datos de cada motor sea diferente. A diferencia de Delgado Domínguez, Baeza-Yates distingue en un robot las funciones de análisis o rastreo ("crawling") de las de indexación ("indexing"), con lo cual 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 usan una arquitectura de tipo robot-indexador centralizada (ver Ilustración 2.2). A pesar de lo que puede inducir su nombre[32], el robot no se mueve por la red, ni se ejecuta sobre las máquinas remotas cuyas páginas consulta. El robot funciona sobre el sistema local del motor de búsqueda y envía una serie de peticiones a los servidores remotos que alojan las páginas a analizar. El índice también se gestiona localmente. Esta arquitectura clásica es la que implementa, entre otros, el motor *Alta Vista*, "precisando para ello, en 1998, de 20 ordenadores multiprocesadores, todos con más de 130 Gb de 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].

Baeza-Yates aporta otra denominación: "walkers" ("andadores"). Todos estos términos definen un movimiento lento y continuo entre las distintas páginas que conforman la web (que se puede traducir como "tela de araña").

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

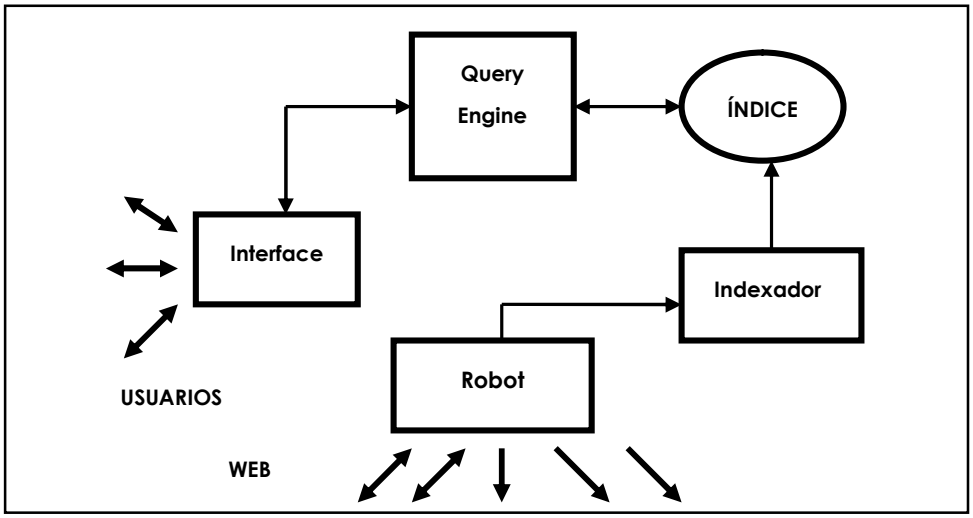


Ilustración 2.2 Arquitectura simple de un motor de búsqueda. Fuente: o a partir de un filtrado de documentos. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p.

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

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

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

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

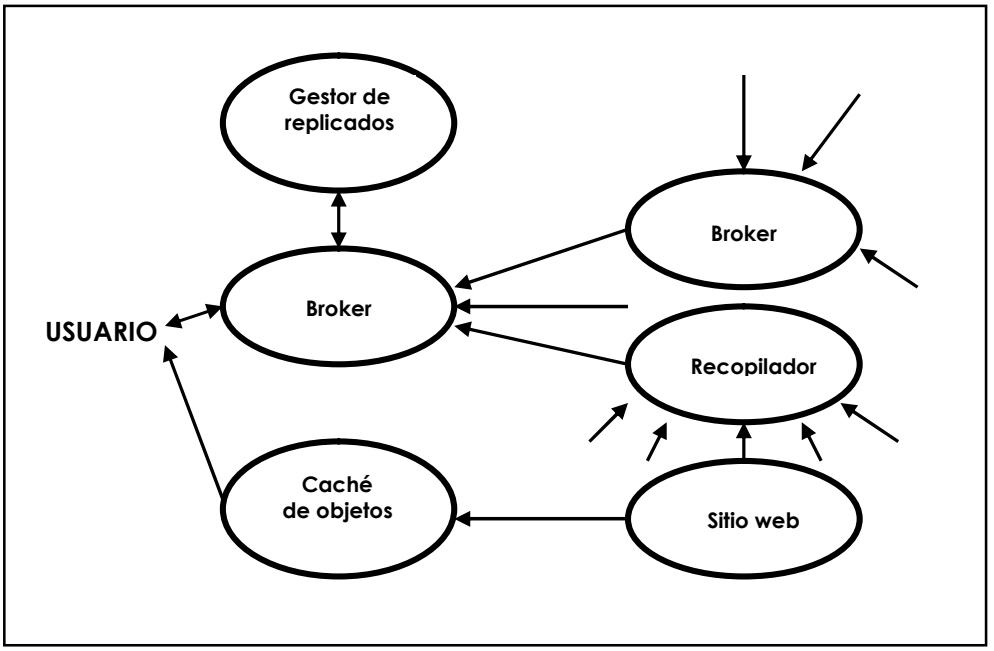


Ilustración 2.3 Arquitectura Harvest. Fuente Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p.

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

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

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

Interface de usuario.

El estudio de la interface debe abordarse bajo dos perspectivas: la interface que el sistema dispone para que el usuario exprese sus necesidades de información (Chang la denomina "interface de consulta" [CHA, 2001]) y la interface de respuesta que dispone el sistema para

mostrar al usuario el resultado de su operación de búsqueda [BAE, 1999]. No todos los sistemas poseen iguales prestaciones en la recuperación de información. La más simple interface de usuario es la clásica caja dentro de un formulario web (ver Ilustración 2.4). En esta caja, el usuario inserta su ecuación de búsqueda, pudiendo, a veces, introducir alguna restricción a la búsqueda, como por ejemplo: el idioma de las páginas a recuperar, el tipo de objeto, si se desea emplear la búsqueda por frase literal o "búsqueda exacta", etc.

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

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

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

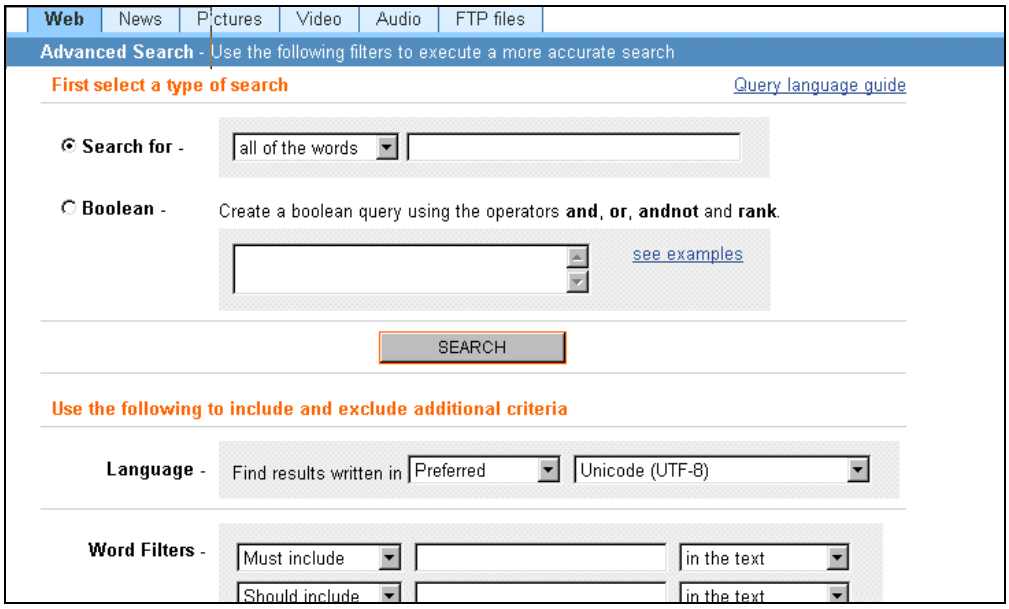


Ilustración 2.5 Sección del formulario de búsqueda avanzada del motor All the Web. Fuente: <http://www.alltheweb.com/advanced>

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

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

5. Correspondencia de patrones

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

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

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

El caso más cercano de proximidad es la adyacencia[33] (caso extremo que se produce cuando los términos están escritos en un orden determinado y uno a continuación del otro). Algunos motores permiten la búsqueda en lenguaje natural, que puede resultar especialmente interesante para aquellos usuarios no experimentados en el uso de un motor específico o en el empleo de los operadores booleanos o basados en el contexto.

Estos sistemas interpretan cuestiones del estilo de "¿qué jugador de fútbol es el máximo goleador de la Copa de Europa?" o "¿qué ciudad es la capital de Angola?", devolviendo como resultados un conjunto de documentos que han considerado adecuados con la temática de la pregunta efectuada, tras haber sometido a esta expresión a un procedimiento de análisis del texto, interpretando la necesidad informativa (*Alta Vista* y *Northern Light* implementan esta modalidad). Un caso extremo de procesamiento de las expresiones en lenguaje natural es el motor *Askjeeves,* que llega a simular una "entrevista" con el usuario ya que, tras recibir la cuestión, la interpreta y extrae de su base de conocimientos una serie de cuestiones que traslada al usuario para refinar su exploración en la base de datos y ajustar mejor la respuesta. Por último, algunos sistemas devuelven los documentos por correspondencia con un patrón de caracteres introducido en la interface de consulta[34].

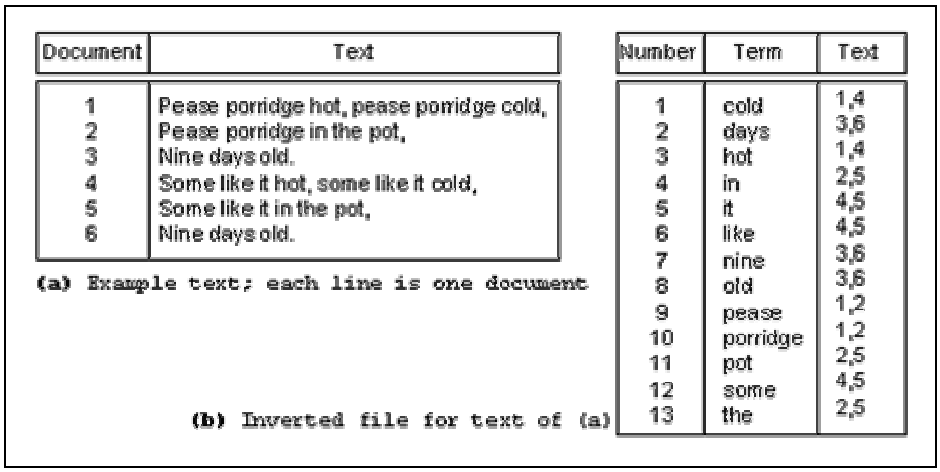
---

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

[34] 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], suele consistir en una lista de palabras con cierto valor de discriminación asociadas a sus correspondientes documentos, que en este caso son las descripciones de los contenidos de las URL recopiladas. La mayor parte de los motores de búsqueda emplean como estructura de datos un *fichero inverso* [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 2.6 Ejemplo de estructura de un fichero inverso simple. Fuente: Rijsbergen, C.J. Information Retrieval. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 21 de octubre de 2001]

En la práctica el fichero inverso se convierte en una enorme estructura de datos con serios problemas de gestión. Los distintos motores de búsqueda se sirven de distintos esquemas para definir estas 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:

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

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

El índice emplea un conjunto de punteros que apuntan a una tabla que contiene las URL en las que aparece una palabra clave. La manera

de ordenar estos punteros depende de un mecanismo interno, basado generalmente en su frecuencia o en su peso en el documento. El enorme tamaño de la colección de URL recopiladas por los motores obliga a buscar formas de simplificar al máximo el tamaño de estos índices. En la Tabla 2.3 se presentan algunas de las diversas técnicas empleadas.

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

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

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

Tipos de robots.

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

- "*Knowbots*: Programados para localizar referencias hipertexto dirigidas hacia un documento, servidor, etc., en particular. Permiten evaluar el impacto de las distintas aportaciones que engrosan las distintas áreas de conocimiento presentes en la Red.
- *Wanderers* (vagabundos): Encargados de realizar estadísticas: crecimiento de la Red, número de servidores conectados, etc.
- *Worms* (gusanos): Encargados de la duplicación de directorios FTP, para incrementar su utilidad a un número mayor de usuarios

- *WebAnts* (hormigas): Conjunto de robots físicamente alejados que cooperan para la consecución de distintos objetivos, como por ejemplo para llevar a cabo una indización distribuida. [DEL, 1998]"

Funcionamiento de los robots.

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

- *breadth-first* (cobertura amplia pero no profunda) y
- *depth-first* (cobertura vertical profunda) [BAE, 1999].

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

Tipo de rastreo	**No**	**Sí**
Rastreo profundo	Excite	El resto
Soporte de marcos	Excite, FAST	El resto
Mapas de imágenes	Excite, FAST	Alta Vista, Northern Light
Robots.txt	Ninguno	Todos
Metadatos	Excite	El resto
Rastreo por popularidad	Ninguno	Todos
Inclusión pagada	Excite, Google[35]	Alta Vista, Inktomi, FAST

Tabla 2.4 Características de rastreo de los robots de los principales motores de búsqueda. Fuente: Aguilar González, R. Monografía sobre motores de búsqueda [En línea]. Yahoo! Geocities, 2002.

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

<http://www.geocities.com/motoresdebusqueda/crawlers.html> [Consulta: 3 de abril de 2002]

Indización de las páginas.

A medida que los robots recopilan páginas, la información contenida en las mismas debe ser indizada, Delgado Domínguez opina que "existen dos estrategias básicas, no mutuamente excluyentes, para realizar este proceso: usar información que provee el creador o editor del documento, o extraerla directamente del documento" [DEL, 1998]. El volumen de información que gestiona un robot obliga a que el motor de 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, WISEnut o Kartoo,* entre otros) [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 Google, la indización la realizan dos módulos: el *indexador* y el *clasificador*. El primero lee las páginas procedentes del *storeserver*[36], descomprime los documentos y selecciona los términos incluidos en los mismos. Cada documento se convierte así en un conjunto de palabras (o '*hits*'), donde se graba la palabra y su posición en el documento, una aproximación de su fuente de texto y otra serie de detalles, por medio del *clasificador*.

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

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

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

Alineado de los documentos (ranking).

El alineado constituye uno, sino el que más, de los procesos críticos a la hora de valorar la efectividad de un motor de búsqueda, ya que se trata del orden en el que el motor presenta los resultados a sus usuarios,

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

quienes, como es lógico esperan encontrar los documentos más relevantes con sus necesidades situados entre los primeros. El motor debe ordenar el conjunto de documentos constituyente de la respuesta en función de la *relevancia* de estos documentos con el tema de la pregunta realizada.

En función del buen funcionamiento de su algoritmo de alineamiento, el motor será mejor o peor valorado por los usuarios del mismo. Si un motor no discrimina su respuesta en función de la *relevancia* con la temática objeto de la pregunta, el usuario encontrará documentos muy relevantes mezclados con otros menos relevantes e incluso con muchos nada relevantes, lo que le obligará a consultar un gran número de los documentos devueltos por el motor, teniendo que visitar muchas pantallas y perdiendo, en consecuencia, un cuantioso tiempo.

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

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

Baeza-Yates cita tres métodos englobados en el primer grupo, en adición al esquema *tf-idf*: "se denominan *Booleano extendido*, *Vectorial extendido* y *Más citado*. Los dos primeros son adaptaciones de los algoritmos normales de alineamiento empleados en estos modelos clásicos de recuperación de información para incluir el hecho de la existencia de enlaces entre las páginas web. El tercero se basa únicamente en los términos incluidos en las páginas que poseen un enlace hacia las páginas de la respuesta" [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*[37]: 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 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 recibe desde otras tiene mucho que ver con la calidad de la misma. Este algoritmo se puede resumir de la siguiente manera: "una página A tiene T1....Tn páginas que apuntan a ella por medio de algún enlace (es decir citas). El parámetro **d** es un factor que se puede fijar entre 0 y 1 (generalmente se fija en 0.85). Sea C(A) es número de enlaces que salen de la página A. Entonces, el *PageRank* de la página A vendrá dado por la siguiente expresión: **PR(A) = (1-d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))**". Este cálculo puede realizarse por medio de un algoritmo iterativo y corresponde al vector propio de una matriz normalizada de enlaces en la web. *PageRank* está concebido como un modelo del comportamiento del usuario: si se asume que hay un "navegante aleatorio" que pasa de una página a otra sin presionar nunca el botón de "retroceder" y que, eventualmente nunca se aburriera, la probabilidad de que este navegante visitara una página determinada es precisamente su *PageRank*. Es decir, se trata de un modelo basado en los enlaces de las páginas y que pretende representar la forma de trabajar de los usuarios. Otra justificación intuitiva de PageRank es que una página puede tener un alto coeficiente de PageRank si existen muchas páginas que apuntan a ella, o si hay un número algo menor de páginas que apuntan a ella pero que posean, a su vez, un alto nivel de PageRank. Lo normal es que "aquellas páginas muy citadas son páginas que

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

vale la pena consultar y aquellas que sólo posean un enlace son páginas de poco interés para su consulta" [BRI, 1998].

Confianza en el funcionamiento de los motores de búsqueda.

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

Es seguro que casi todos los usuarios de estos sistemas habrán reflexionado de una manera más o menos análoga al siguiente planteamiento de Manchón: "los resultados de algunos estudios indican que muchos usuarios prefieren la búsqueda jerárquica frente al motor de búsqueda. Ello puede ser causado por la proliferación de motores de búsqueda muy defectuosos que en la práctica no encuentran nunca la información deseada. Los usuarios se han acostumbrado a desconfiar de 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, que descalifica por completo la operatoria de estos sistemas.

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

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

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

Este problema intenta ser paliado por los responsables de los propios motores quienes, en mayor o menor medida, insertan en la ayuda 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 del motor de búsqueda. En algunos casos esta interface ofrece escasas prestaciones a los usuarios para mejorar la calidad de sus operaciones de búsqueda y, en otros casos, resultan confusas e inducen a error a los usuarios, lo que tampoco contribuye a mejorar la efectividad del sistema.

El tercero de los problemas es el de la inadecuada indización de las páginas web. A la ausencia de una estructura básica de los documentos analizados por los robots, hay que unir lo reciente de esta tecnología (algunos motores con más de cien millones de páginas en su base de datos 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[38]. 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.

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

3 Evaluación de la Recuperación de la Información.

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

Necesidad de la evaluación de los SRI.

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

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

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

Estos criterios no deben confundirse, la *eficacia en la ejecución* es la medida del tiempo que se toma un SRI para realizar una operación. Este parámetro ha sido siempre la preocupación principal en un SRI, especialmente desde que muchos de ellos son interactivos, y un largo tiempo de recuperación interfiere con la utilidad del sistema, llegando a alejar a los usuarios del mismo. La *eficiencia del almacenamiento* es

medida por el espacio que se precisa para almacenar los datos. Una medida común de medir esta eficiencia, es la ratio del tamaño del fichero índice unido al tamaño de los archivos de documentos, sobre el tamaño de los archivos de documentos, esta ratio es conocida como *exceso de espacio*. Los valores de esta ratio comprendidos entre 1,5 y 3 son típicos de los SRI basados en los ficheros inversos.

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

Bors considera que "los experimentos, evaluaciones e investigaciones tienen una larga tradición en la investigación de la recuperación de la información, especialmente los relacionados con el paradigma de comparación exacta, concentrados en mejorar el acierto entre los términos de una pregunta y la representación de los documentos para facilitar el aumento de la *exhaustividad* y de la *precisión* de las búsquedas" [BOR, 2000]. Este mismo autor sugiere una diferencia, a la hora de esa evaluación, entre la evaluación del "acceso físico" y la evaluación del "acceso lógico" (o "intelectual"); considerando que las evaluaciones que se lleven a cabo, deben centrarse en el segundo tipo. El acceso físico es el que concierne a cómo la información demandada es recuperada y representada de forma física al usuario. Tiene que ver con la manera en la que un SRI (manual o automatizado) encuentra dicha información, o indica ciertas directrices al usuario sobre cómo localizarla, una vez que le proporciona su dirección. Este acceso se encuentra muy vinculado con las técnicas de recuperación y de presentación de la información. El acceso lógico está relacionado con la localización de la información deseada.

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

Estas afirmaciones, realizadas a principio de la década de los años noventa, siguen enteramente vigentes diez años después. Borlund distingue entre "aproximaciones al funcionamiento del sistema y aproximaciones centradas en el usuario" [BOR, 2000], plenamente coincidentes con el "acceso físico" y el "acceso lógico" de Blair. Otro hecho que no debemos perder de vista es la actual competencia establecida entre los desarrolladores de los algoritmos que emplean los directorios, motores de búsqueda o metabuscadores de la red Internet, sistemas que rivalizan sobre cómo facilitar al usuario más documentos en

el menor tiempo posible, sin entrar a considerar que, quizá el usuario prefiera que la información entregada como respuesta le sea verdaderamente útil para sus necesidades, aunque tenga que esperar algunas milésimas de segundo más para recibirla.

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

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

Relevancia vs Pertinencia.

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

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

- Un mismo documento puede ser considerado relevante, o no relevante, por dos personas distintas en función de los motivos que producen la necesidad de información o del grado de conocimiento que sobre la materia posean ambos. Llegados a un caso extremo, un mismo documento puede parecer relevante o no a la misma persona en momentos diferentes de tiempo [LAN, 1993].
- Resulta difícil definir, a priori, unos criterios para determinar cuándo un documento es relevante e incluso resulta complicado explicitarlos de forma clara y concisa, "es más fácil proceder a la determinación de la *relevancia* que explicar cómo la misma se ha llevado a cabo" [BLA, 1990]. De hecho, considera que "el concepto de la *relevancia* no está claro en un primer término, se nos presenta como un concepto afectado de una gran dosis de subjetividad que puede ser explicado de múltiples maneras por distintas personas y, por tanto, dentro del contexto de una búsqueda en un sistema de recuperación de información este concepto puede ser precisado de muchas maneras distintas por todos aquellos que realicen búsquedas en el mismo. En un segundo lugar, no debe sorprendernos que un usuario afirme que unos determinados documentos son relevantes a sus necesidades de información y que, en cambio, no sea capaz de precisarnos con exactitud qué significa ser relevante para él" [BLA, 1990].
- Esto no quiere decir que el concepto carezca de importancia, sino que la realización de un juicio de *relevancia* viene a formar parte de ese amplio conjunto de tareas cotidianas que llevamos a cabo los seres humanos (procesos cognitivos por tanto) pero que generalmente, no podemos encontrar las palabras adecuadas para proceder a su descripción" [BLA, 1990].
- Por último, puede resultar aventurado calificar de forma categórica a un documento como relevante con un tema, o por el contrario, calificarlo como no relevante de igual manera. Resulta muy normal encontrar documentos que, en alguno de sus apartados resulta relevante con una materia determinada pero que no en el resto de sus contenidos. Para algunos autores, surge entonces el concepto de "*relevancia* parcial", debido a que, en realidad, la *relevancia* no puede medirse en términos binarios (sí/no), sino que puede adquirir muchos valores intermedios (muy relevante, relevante, escasamente relevante, mínimamente relevante, etc.), lo que propicia que la *relevancia* pueda medirse en términos de función continua en lugar de una función binaria (que sólo admite dos estados).

Todos estos impedimentos condicionan, en cierto grado, la viabilidad de la *relevancia* para constituirse en un criterio de evaluación

de la recuperación de la información. Cooper introduce la idea de "utilidad de un documento", considerando que es mejor definir a la *relevancia* en términos de la percepción que un usuario posee ante un documento recuperado, es decir: "si el mismo le va a ser útil o no" [COO, 1973]. Este nuevo punto de vista presenta, esencialmente, una ventaja: emplaza la estimación de la adecuación o no de un documento recuperado dentro del juicio que llevará a cabo el usuario, en tanto que, tal como hemos comentado anteriormente enumerando los problemas de la *relevancia*, podemos asumir que un usuario tendrá problemas a la hora de definir qué es relevante y qué no lo es, pero tendrá pocos problemas a la hora de decidir si el documento le parece o no útil.

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

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

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

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

Para Salton: "el conjunto pertinente de documentos recuperados puede definirse como el subconjunto de los documentos almacenados en

el sistema que es apropiado para la necesidad de información del usuario" [SAL, 1983]. El término *pertinencia* significa "calidad de pertinente", entendiéndose como "pertinente" a todo lo que viene a propósito o resulta oportuno, es decir que podemos decir que un documento pertinente es un documento que resulta oportuno, porque le proporciona al usuario final la información que a él le cumple algún propósito.

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

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

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

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

En realidad, los documentos recuperados no son relevantes o no relevantes, propiamente hablando, es decir, que no se trata de una decisión binaria, en tanto que los contenidos de los documentos pueden

coincidir en mayor o menor parte con las necesidades de información. Lo que sí podemos determinar es si son o no relevantes para una determinada persona. Desde un punto de vista pragmático, el mismo documento puede significar varias cosas para personas diferentes; los juicios de *relevancia* pueden sólo realizar evaluaciones semánticas o incluso sintácticas de documentos o preguntas. Pero estos juicios fallan al involucrar a usuarios particulares, y también fallan al identificar dónde el usuario realmente encuentra a un documento particularmente relevante. Estos autores bromean un poco al respecto, parafraseando, ligeramente modificada, una vieja frase: "la *relevancia* reside en los detalles".

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

Las primeras evaluaciones.

Es norma común en toda la bibliografía consultada, hacer referencia a una serie de evaluaciones pioneras llevadas a cabo a partir de la mitad de la década de los años cincuenta, que marcaron el camino de los trabajos en este campo, llegando a aportar (a pesar de lo primitivos que eran esos SRI), una serie de medidas que siguen estando vigentes hoy en día. Ya se ha indicado al principio de este capítulo que Borlund vincula la existencia de los SRI a la necesidad de evaluarlos, manifestando que algunos estudios de evaluación son casi tan antiguos como los propios sistemas: *Cranfield*, *MEDLARS* o *SMART* [BOR, 2000].

Algunos de los manuales consultados presentan estos estudios con amplio detalle, por lo que no parece oportuno llevar a cabo una detallada exposición de los mismos, sino que se considera más interesante exponer los objetivos y logros alcanzados en los procesos de evaluación más destacados y resaltar qué ha trascendido posteriormente de cada uno de ellos, remitiendo a la bibliografía consultada para una mayor amplitud en las exposiciones. En lo que sí coinciden la mayoría de los manuales consultados es remitir al lector hacia trabajos previos de Sparck-Jones, quien recopila todos los estudios sobre la medida de la efectividad de la recuperación de información realizados entre los años 1958 y 1978.

Proyectos Cranfield.

Si bien Lancaster afirma que los primeros estudios datan del año 1953, verdaderamente "los primeros estudios significativos fueron los *Proyectos Cranfield*, que proporcionaron una nueva dimensión a la investigación en SRI" [CHO, 1999]. Estos estudios se llevaron a cabo en el *Instituto Cranfield de Tecnología* y representan el arranque de la investigación en la evaluación de la recuperación de información, tal

como dice López Huertas para quien "los tests *Cranfield* son el punto de partida de las investigaciones empíricas y experimentales sobre la recuperación de la información, estudios que, hasta ese momento, se desenvolvían en un ámbito filosófico o especulativo" [LOP, 1998]. Originariamente, este estudio proyectaba evaluar el funcionamiento de varios sistemas de indización y el rendimiento de los SRI basados en ellos, aunque su repercusión ha sido mayor, ya que algunas de las medidas más comúnmente empleadas en la evaluación de los SRI (*precisión*[39], *exhaustividad*[40], tasa de fallo[41], etc.) fueron establecidas a partir de la realización de este estudio.

Son dos los estudios *Cranfield* más importantes. El primero de ellos, dirigido por Cleverdon, comenzó en 1957 y tenía como objetivos comparar la efectividad de cuatro sistemas de indización: un catálogo alfabético de materias basado en una lista de encabezamientos; una clasificación CDU; un catálogo basado en una clasificación por facetas y, finalmente, un catálogo compilado por un índice coordinado de unitérminos.

Los resultados de este estudio proporcionan unos valores de *exhaustividad* altos (entre el 60-90% con un promedio del 80%), favorecidos por el tiempo dedicado a la indización, y proporcionó algunos datos sobre el sistema de indización: "primeramente, el test probó que el rendimiento de un sistema no depende de la experiencia del indizador; en segundo lugar, mostró que los sistemas donde los documentos se organizan por medio de una clasificación facetada rendían menos que los basados en un índice alfabético" [CHO, 1999].

López Huertas cita a Belkin para resaltar la repercusión de este experimento en los sistemas de indización y en los lenguajes documentales[42] [LOP, 1998], aunque donde realmente ha sido realmente importante es en el área de la evaluación de los SRI, principalmente por dos razones: estableció los factores más implicados en el funcionamiento de los SRI y refrendó el desarrollo de la primera metodología de evaluación, sobresaliendo la introducción de las medidas de *exhaustividad* y *precisión*, que, bien ligeramente modificadas, bien

---

[39] Medida que se presenta en el apartado siguiente y que determina el porcentaje de acierto de una operación de recuperación de información.

[40] Medida que se presenta en el apartado siguiente y que determina la profundidad de una operación de recuperación de información, es decir, el porcentaje de documentos recuperados válidos para el usuario comparado con el total de documentos interesantes para el usuario que hay en la base de datos.

[41] Medida que se presenta en el apartado siguiente y que determina el porcentaje de error cometido en una operación de recuperación de información.

[42] Belkin considera esta experiencia el origen del *paradigma físico* y del *paradigma lógico o cognitivo* que tanta fama ha proporcionado a este autor.

provistas de nuevas interpretaciones o bien asociadas a exacerbadas críticas, nos siguen acompañando hasta nuestros días.

El segundo proyecto *Cranfield* fue un experimento controlado destinado a fijar los efectos de los componentes de los lenguajes de indización en la ejecución de los SRI. Este test también pretendía ofrecer datos sobre la naturaleza de los fallos de un SRI[43]. Sus resultados mostraron resultados contradictorios, principalmente a la hora de seleccionar los términos más adecuados para representar los conceptos contenidos en los documentos, ya que los sistemas de indización libre (no controlados) ofrecieron mejor rendimiento que los controlados, obteniéndose mejores resultados con lenguajes de indización basados en los títulos de los artículos que en los basados en los resúmenes, hecho sorprendente cuando menos. Vickery comenta "que las medidas usadas en el segundo experimento *Cranfield* no caracterizaron adecuadamente los aspectos operativos de un SRI" [CHO, 1999]. Por lo tanto, este segundo proyecto no ha tenido tanta repercusión como el primero.

MEDLARS.

El funcionamiento del sistema MEDLARS[44] fue evaluado entre los años 1966 y 1967 por Lancaster. Se trata, probablemente, "del SRI más famoso disponible de la *Biblioteca Nacional de Medicina*" [SAL, 1983]. La experiencia pretendía observar la efectividad de la recuperación de información de esta base de datos y averiguar la manera de mejorarla. Los resultados emanados proporcionaron valores medios de *exhaustividad* más bajos que los obtenidos en el primer test de *Cranfield*, cifrándose aproximadamente en torno al 57%, y valores medios de *precisión* del 50%. A diferencia del test de *Cranfield*, sí proporcionó pormenores sobre las razones de los fallos producidos en la recuperación de información, centrándose la mayor proporción de los problemas en la indización, en la realización de las búsquedas y en la interacción del usuario con el sistema (estas tres razones totalizan un 87% de los fallos) [CHO, 1999].

SMART.

El sistema SMART, diseñado en 1964 por Salton, fue concebido como una herramienta experimental de la evaluación de la efectividad de muchos tipos de análisis y procedimientos de búsqueda. Este sistema se distingue a si mismo del resto de los SRI convencionales en cuatro aspectos fundamentales: (1) usa métodos de indización automática para asignar descriptores a los documentos y a la peticiones de información; (2) agrupa documentos relacionados dentro de clases comunes de materias, siendo posible comenzar a estudiar un término específico y obtener sus términos

[43] Este problema no fue considerado en el primero de los tests de *Cranfield*.

[44] MEDLARS: Medical Literature Analysis and Retrieval System, sistema de recuperación de información de la Biblioteca Nacional de Medicina (<http://www.nlm.nih.gov/>).

asociados; (3) identifica los documentos a recuperar por similitud con la pregunta realizada por el usuarios y finalmente, (4) incluye procedimientos automáticos para generar mejores ecuaciones de búsqueda basadas en la información obtenida de búsquedas anteriores" [SAL, 1983].

Los experimentos originales realizados con SMART intentaban desarrollar un prototipo de SRI íntegramente automatizado. Se propusieron medidas suplementarias a las usadas en los tests *Cranfield*, algunas de las cuales se han incorporado a posteriores evaluaciones de SRI[45]. En total, se determinaron cuatro grupos de evaluación:

Grupo de evaluación	**Función**
A	Autores de las búsquedas, cada uno de ellos lleva a cabo juicios de *relevancia* de las búsquedas que realiza.
B	Enjuiciadores de la *relevancia* de búsquedas realizadas por otros autores, y sólo una de cada una de ellas. Son personas distintas.
C	El documento es relevante para una búsqueda determinada si el juicio A o el juicio B es relevante.
D	El documento es relevante para una búsqueda determinada si el juicio A y el juicio B es relevante.

Tabla 3.1 Grupos de evaluación formados para el experimento del sistema SMART. Fuente: Salton, G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983

SMART incorpora tres procedimientos diferentes de análisis del lenguaje, conocidos como *palabra*, *lema* y *tesauro*. El primero de estos métodos emplea palabras comunes reducidas a su forma singular a las que se les asigna un peso. El segundo método extrae la base de la palabra, desprendiéndola de los sufijos, de manera que se agrupan varias palabras en un mismo lema, que es a quien se le asigna el peso. Por último, con el tesauro se asignan los términos descriptores que mejor representan a los conceptos de los documentos y se les asigna un peso. Tras la realización de los experimentos sobre este sistema, se obtuvieron dos series de resultados principales:

1. El procedimiento de análisis del texto que hace uso del tesauro mejora ligeramente al de los lemas y ambos resultan bastante mejores que el los términos simples o palabras.
2. Los mejores resultados en términos de *exhaustividad* y *precisión* se obtienen en el cuarto grupo de evaluaciones, es decir, cuando

[45] Se refieren a los gráficos *Exhaustividad-Precisión* promedio de ambos valores y a sus versiones normalizadas. Estas medidas se presentan posteriormente.

tanto los usuarios que realizan las preguntas como los evaluadores ajenos a esas preguntas están de acuerdo [CHO, 1999].

Salton realiza una interesante comparación de su sistema con el sistema MEDLARS [SAL, 1983], a partir de una subcolección de documentos extraída del sistema *Science Citation Index* (SCI). Los resultados de *exhaustividad* y *precisión* de MEDLARS son ligeramente inferiores que los obtenidos por SMART cuando se aplica el método del tesauro para reconocer el texto. En cambio, MEDLARS supera a los otros dos procedimientos de SMART.

El proyecto STAIRS.

La evaluación del funcionamiento de STAIRS[46] fue un proyecto desarrollado en los años ochenta por Blair y Maron [BLA, 1990], quienes evaluaron la efectividad en la recuperación de información de este sistema examinando alrededor de 40.000 documentos legales (unas 350.000 páginas de texto completo), lo que representa un sistema de tamaño real. Los juicios de *relevancia* fueron llevados a cabo por los usuarios que realizaron las consultas.

El número de documentos útiles no recuperados fue estimado a través de la aplicación de una serie de técnicas de muestreo estadístico. Los usuarios utilizaron, durante el ensayo, los SRI de similar manera que si estuvieran realizando una consulta normal y corriente.

El experimento proporcionó unos resultados de *precisión* que rondaban valores cercanos al 75%, y unos valores de *exhaustividad* que oscilaban alrededor del 20%, cuantías algo más bajas que las obtenidas en otros estudios anteriores[47], especialmente en el caso de la *exhaustividad*. Esta medida se analizó dependiendo de si el juicio de valor lo realizaba el abogado (experto) o el pasante (abogado también, pero menos experto). Los resultados mostraron que la media de las *exhaustividade*s obtenida por los abogados (9,73%) superaba a la media de los pasantes (7,56%).

En cambio, las diferencias en *precisión* son mayores, alcanzando los trece puntos (85% frente a un 82%). Basándose en el valor de la *precisión*, este experimento propugna que los juicios de *relevancia* deben llevarlo a cabo un grupo de expertos, aunque esta conclusión parece muy simple, en tanto que las diferencias, tanto de *precisión* como de *exhaustividad* no son muy grandes.

---

[46]STAIRS: Storage And Information Retrieval System.

[47]Estos bajos valores de *exhaustividad* planteaban una duda a considerar en el seno del contexto de una base de datos de legislación, ya que este sistema debería entregar al abogado la mayor parte de documentos relacionados con un tema. De hecho, los usuarios pensaban que estaban recuperando alrededor del 75% de los documentos relevantes con un tema.

Otra conclusión importante de este estudio es que existen grandes diferencias entre la percepción de los usuarios del SRI y la realidad en todo lo relacionado con la *exhaustividad* de una búsqueda.

De hecho, este estudio muestra la dificultad de estimar fiablemente el nivel de esta medida, es decir, demostró lo difícil que resulta estimar cuántos documentos relacionados con la temática de la pregunta van a estar incluidos en la base de datos, a partir de una muestra. Por último, los autores achacan los pobres resultados obtenidos en este estudio a razones de "desinformación", en este caso incide mucho que el sistema STAIRS sea de texto completo y el usuario se siente incapaz de predecir las palabras y frases que representan a los documentos útiles de la colección.

Conferencias TREC.

Las conferencias TREC[48] se han convertido en el foro de intercambio científico más prestigioso del campo de la recuperación de información. TREC reúne a creadores de diferentes sistemas y compara los resultados que éstos obtienen en diferentes pruebas, previamente estandarizadas y acordadas por todos. Este foro se viene celebrando anualmente desde 1991. TREC nació con la idea de resolver uno de los mayores problemas de las evaluaciones de los SRI: las mismas suelen llevarse a cabo sobre pequeñas colecciones de documentos, y sus resultados resultan de difícil extrapolación a la totalidad de la colección almacenada. En 1991, con la idea de salvar este problema, DARPA[49] propuso poner en marcha los experimentos TREC en el Instituto Nacional de Ciencia y Tecnología (NIST), para propiciar que los investigadores en recuperación de información probaran sus sistemas en una gran colección de documentos. Chowdhury sintetiza los cinco objetivos de estos experimentos:

1. Aumento de la investigación en recuperación de información sobre grandes colecciones de documentos.
2. Desarrollo de la comunicación entre los entornos académicos, industrial y gubernamentales a través de la realización de un foro abierto.
3. Incremento de la transferencia de tecnología.
4. Presentación del estado de la investigación y desarrollo en este campo de forma anual
5. Perfeccionamiento de las técnicas de evaluación [CHO, 1999].

---

[48] TREC: Text REtrieval Conferences. Se puede ampliar información en <http://trec.nist.gov/>

[49] DARPA: *Defence Advanced Research Projects Agency*. Es la Agencia de Proyectos de Investigación del Departamento de Defensa de Estados Unidos de América. Es la organización a la que debe su nacimiento la Red Internet.

La primera conferencia, TREC-1 (1992), ofreció como resultado principal el hecho de la existencia de una amplia similitud entre los SRI que hacen uso de técnicas basadas en lenguaje natural y los basados en los modelos probabilístico y los basados en el modelo del vector. En la conferencia TREC-2 (1993), se detectó una significativa mejoría de la recuperación de información, con respecto a la anterior.

Las siguientes conferencias aportaron nuevas prestaciones a los experimentos: localización de información en varias bases de datos de forma simultánea, presencia de errores ortográficos con el fin de valorar el comportamiento de los SRI ante ellos y recuperación de información en idiomas distintos del Inglés (se eligieron el Español y el Chino) para valorar los posibles cambios de comportamiento de los SRI.

Estas conferencias han aportado la evaluación de variadas modalidades de recuperación de información (desde el clásico modelo booleano a la búsqueda por cadenas de texto o las búsquedas basadas en diccionarios), y han demostrado hasta qué punto pueden alcanzarse resultados significativos de investigación a través de la cooperación entre investigadores en el ámbito mundial.

De hecho, en palabras de Sparck Jones, "la comunidad investigadora debe estar muy agradecida a las conferencias TREC, en tanto que han revitalizado la investigación en recuperación de información y también ha demostrado la importancia de este campo de investigación en áreas afines, tales como el procesamiento del lenguaje natural y la inteligencia artificial" [CHO, 1999].

Medidas tradicionalmente empleadas.

Tras haber delimitado conceptualmente qué se entiende por resultado relevante y por resultado pertinente, se procederá a la identificación de los parámetros utilizados en la medida de la efectividad de los SRI. Es decir, a partir de este punto, se abandona la discusión sobre qué se debe considerar relevante y qué no se debe considerar, y bajo qué circunstancias se desarrolla ese juicio o evaluación, para adentrarnos en el estudio de las variables a medir en la recuperación de información. Es conveniente recordar que, de un lado, existen una serie de medidas orientadas a analizar el acceso físico a los datos, y por el otro, existen otras que intentarán analizar si el contenido es o no pertinente.

Rijsbergen se pregunta *qué evaluar*, y se responde citando a Cleverdon (tests *Cranfield*) con seis medidas principales: "la cobertura de una colección; el tiempo de respuesta del sistema; la forma de presentación de los resultados; el esfuerzo realizado por el usuario; la *exhaustividad* y la *precisión* del sistema" [RIJ, 1999]. Para este autor, las cuatro primeras medidas son intuitivas y fácilmente estimables. Considera además que la *exhaustividad* y la *precisión*, son las que verdaderamente pretenden medir la *efectividad* del SRI: "la efectividad es puramente una

medida de la capacidad del sistema para satisfacer al usuario en términos de la *relevancia* de los documentos recuperados" [RIJ, 1999].

Chowdhury presenta las medidas de Cleverdon y cita a Vickery, quien propone seis medidas divididas en dos grupos: "el primero lo forman la cobertura (proporción de las referencias que potencialmente podrían haberse recuperado), la *exhaustividad* y el tiempo de respuesta; el segundo lo forman la *precisión*, la usabilidad (el valor de las referencias considerado en términos de fiabilidad, comprensión, actualidad, etc.) y la presentación (la forma en la que los resultados de la búsqueda son presentados al usuario)" [CHO, 1999].

Las medidas de Cleverdon es también usado por Salton, quien tiene similares dudas sobre el cálculo de los valores de la *precisión* y de la *exhaustividad* [SAL, 1983]. Junto a las medidas basadas en la *relevancia*, diversos autores proponen una amplia serie de medidas basadas en otros criterios. Meadow las sintetiza en tres grupos según su base: la *relevancia*, el resultado y el proceso [MEA, 1992].

Medidas basadas en la *Relevancia*	
Precisión	Documentos relevantes recuperados divididos entre el total de documentos recuperados
Exhaustividad	Documentos relevantes recuperados dividido entre el total de documentos relevantes
Promedio de la efectividad E-P	Promedios de la efectividad en pares de valores de *exhaustividad* y *precisión*

Tabla 3.2. Medidas basadas en la relevancia. Fuente: Meadow, C. T. Text Information retrieval Systems. San Diego: Academic Press, 1993.

En esta primera tabla aparecen las medidas de Precisión y Exhaustividad que serán, a pesar de las críticas que han recibido por su subjetividad, las más empleadas en todos los estudios de evaluación.

Medidas de Resultado	
Precisión	-- ya definida anteriormente --
Exhaustividad	-- ya definida anteriormente --
Promedio efectividad E-P	-- ya definida anteriormente --
Medidas promedio de la satisfacción del usuario	Medidas que pretenden medir la reacción de los usuarios ante el resultado de una búsqueda

Tabla 3.3 Medidas basadas en los resultados. Fuente: Meadow, C. T. Text Information retrieval Systems. San Diego: Academic Press, 1993.

Prácticamente se trata del mismo conjunto de medidas. A este conjunto se le une una nueva, la medida de la satisfacción del usuario ante el resultado obtenido.

Medidas basadas en el Proceso	
Selección	Mide cuántos documentos hay en la base de datos, el grado de solapamiento con otras relacionadas, qué se espera de la base de datos antes de las búsquedas
Contenido	Tipo de documentos de la base de datos, temática de los documentos, frecuencia de actualización
Traducción de una consulta	Se verifica si el usuario puede plantear la consulta directamente o precisa de intermediación
Errores en establecimiento de la consulta	Media de errores sintácticos en la escritura de la búsqueda que propician la recuperación de conjuntos vacíos y erróneos
Tiempo medio de realización de la búsqueda	Tiempo medio de realización de una estrategia de búsqueda
Dificultad en la realización de la búsqueda	A la ratio anterior habrá que añadir los problemas que usuarios inexpertos se pueden encontrar
Número de comandos precisos para una búsqueda	Promedio de instrucciones necesarias para realizar una búsqueda
Coste de la búsqueda	Costes directos e indirectos en su realización
Nº docs recuperados	Extensión del resultado de una búsqueda
Número de documentos revisados por el usuario	Promedio de documentos que los usuarios están dispuestos a revisar

Tabla 3.4 Medidas basadas en el proceso de la recuperación de la información. Fuente: Meadow, C. T. Text Information retrieval Systems. San Diego: Academic Press, 1993.

Medidas basadas en la *relevancia*.

Se considera de una mayor importancia el conjunto de las medidas basadas en la *relevancia* que el conjunto de las medidas basadas en el proceso y en el resultado.

Las medidas correspondientes al segundo grupo sirven para diferenciar unos sistemas de otros con base en las prestaciones de la aplicación informática subyacente, y no permiten la evaluación de aspectos relacionados con el contenido de los documentos.

Las medidas del tercer grupo se encuentran muy relacionadas con las basadas en la *relevancia*, aunque introducen algunos aspectos diferenciadores que presentaremos en el desarrollo de este capítulo, pero, en términos generales, hemos de decir que, son más frecuentes las coincidencias que las discrepancias.

En una búsqueda de información, un usuario obtiene un conjunto de documentos, de los cuales unos formarán parte del subconjunto de

documentos relevantes con la temática objeto de la búsqueda y otros van a formar parte del subconjunto de documentos que no lo van a ser.

Asimismo, este usuario dejará de recuperar otro conjunto de documentos igualmente relevantes con esa temática, y otro conjunto de documentos no relevantes, tal como muestra la ilustración 3.1.

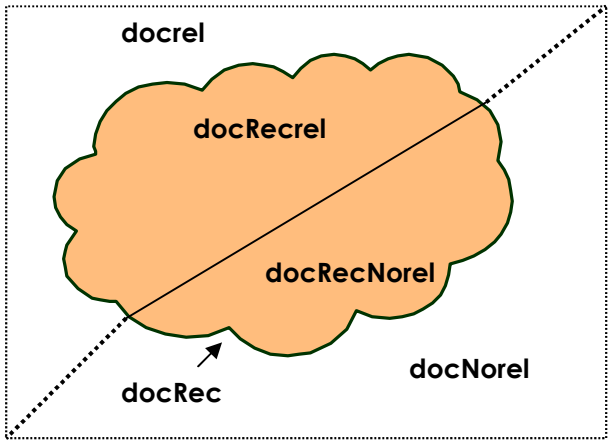


Ilustración 3.1 Distribución de los resultados de una operación de búsqueda. docRec: documentos recuperados: docRecrel: documentos recuperados relevantes; docRecNorel: documentos recuperados no relevantes; docrel: documentos relevantes; docNorel: documentos no relevantes.

Esta distribución de los resultados de una operación de búsqueda, conforma la especificación de una serie de subconjuntos de la base de datos en relación con la pregunta realizada, que nos muestra Rijsbergen en la tabla 3.5, más conocida como *Tabla de Contingencia*:

	RELEVANTES	NO-RELEVANTES	
RECUPERADOS	$A \cap B$	$\neg A \cap B$	B
NO-RECUPERADOS	$A \cap \neg B$	$\neg A \cap \neg B$	$\neg B$
	A	$\neg A$	N

N = número documentos en el sistema

Tabla 3.5 Tabla de contingencia de Rijsbergen. 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]

Esta tabla, que también encontramos en [SAL, 1983], [FRA, 1997], [LAN, 1993], [MEA, 1992], [CHO, 1999] y [KOR, 1997], sirve de base para formular una definición de las medidas de *exhaustividad*, de *precisión* así como una tercera medida, que Rijsbergen destaca como muy

interesante, la *tasa de fallo*. Estas especificaciones se recogen en la siguiente tabla:

Precisión	$\frac{\lvert A \cap B \rvert}{\lvert B \rvert}$
Exhaustividad	$\frac{\lvert A \cap B \rvert}{\lvert A \rvert}$
Fallo	$\frac{\lvert \overline{A} \cap B \rvert}{\lvert \overline{A} \rvert}$

Tabla 3.6 Fórmulas de *Precisión*, *Exhaustividad* y Fallo. 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]

La primera medida basada en la *relevancia* propuesta es la *precisión,* la cual es, quizás, la medida más intuitiva y más sencilla de recordar: La *precisión* mide el porcentaje de documentos recuperados que resultan relevantes con el tema de la pregunta y su cálculo es verdaderamente simple: se divide el total de documentos relevantes recuperados entre el total de documentos recuperados.

La *exhaustividad* conlleva algunos problemas más en su cálculo, si bien la definición está clara, el número de documentos relevantes recuperados dividido entre el número de documentos totales relevantes de la base de datos, no está tan claro cuál es el valor de ese denominador, más bien no está nada claro. Lógicamente, si el usuario conociera de antemano cuántos documentos relevantes hay en la base de datos, ¿por qué no los recupera todos en esa operación de búsqueda obteniendo los valores máximos en ambas medidas?. La respuesta es simple: porque no los puede conocer de antemano, como máximo puede inferir un número aproximado pero nunca podrá afirmar esa cantidad con total seguridad.

La *tasa de fallo* equivale al porcentaje de documentos recuperados no relevantes sobre el total de documentos no relevantes de la base de datos. Esta medida es especialmente importante si se considera que la *precisión* se encuentra muy sujeta a posibles variaciones en el contenido de la base de datos, y se observa que la *tasa de fallo* no adolece tanto de esta dependencia: "los cambios en la *generalidad* de una colección afectan menos a la tasa de fallo que a la *precisión*, que resulta más sensible. En particular, si el nivel de *generalidad* decrece (bien porque el número de documentos relevantes disminuye o bien porque se acrecienta el número de documentos en la colección), el número de documentos recuperados relevantes tiende a decrecer, pero el número total de documentos recuperados así como el número total de documentos no recuperados pueden permanecer en un nivel constante" [SAL, 1983]. Estos argumentos han resaltado la importancia de esta medida.

Salton tambíen se refiere a una nueva medida, el *factor de generalidad*: "el grado de documentos relevantes contenidos en una colección" [SAL, 1983]. Una colección con un alto grado de *generalidad* es una colección donde los documentos relevantes son mayoría frente a los que no lo son. Todas estas medidas se encuentran muy vinculadas entre ellas, tanto que incluso la "*precisión* puede definirse en función de las tres restantes" [SAL, 1983], tal como podemos observar en la siguiente expresión:

$$P = \frac{(E \times G)}{(E \times G) + F \times (1 - G)}$$

P =*precisión;* E =*exhaustividad; G = generalidad y* F = *fallo*

Tabla 3.7 Enunciación de la *Precisión* con base en el resto de las tasas. . Fuente: Salton, G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983

Estas medidas suelen expresarse en un rango que oscila entre 0 y 1, aunque también podrían expresarse en tanto por ciento, como hacen [LAN, 1993] y [CHO, 1999].

La *precisión* y la *exhaustividad* tienden a relacionarse de forma inversa. A mayor valor de *precisión* menor será el valor de la *exhaustividad*. Esta situación viene recogida en casi todos los textos y manuales revisados por medio de una gráfica similar a la que reproducimos:

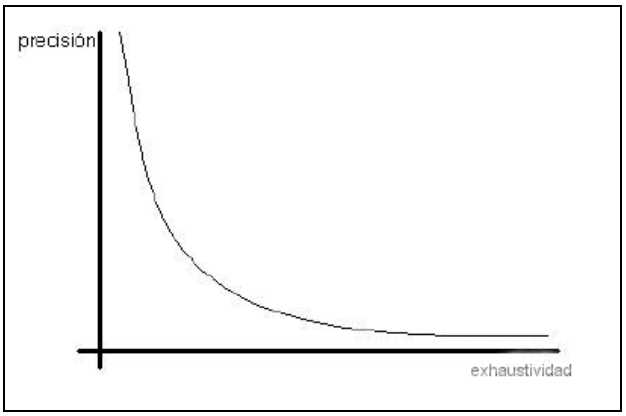


Ilustración 3.2 Evolución típica de la *precisión* y de la *exhaustividad* en un SRI

Consideramos, como primera aproximación a la evolución manifestada en esta gráfica, que un usuario lleva a cabo una operación de recuperación de información en la cual inserta condiciones muy específicas, seguramente obtendrá un conjunto de resultados muy preciso pero, de igual modo, habrá dejado de recuperar algunos documentos a causa de ese alto nivel de especificación. Como ejemplo de esta

situación tenemos estas dos operaciones de búsqueda que planteamos a continuación:

B1: "contaminación del agua en los ríos"

B2: "contaminación en los ríos"

Ambas búsquedas pretenden recuperar el mismo tipo de documento, pero, en el caso de la primera (B1), el usuario la plantea de una forma más específica que la segunda (B2). Este segundo usuario ha pensado que no es necesario emplear el término "agua" en la operación de recuperación de información, ya que, a lo mejor piensa que cuando se contamina un río, es el agua lo que se contamina y le ha parecido redundante e innecesaria tanta especificación. Con toda seguridad, la primera búsqueda (B1) va a adolecer del problema que estamos presentando, ya que basta que en un documento el autor o el indizador no haya alcanzado el nivel de especificación empleado por el usuario que plantea la búsqueda, para que sea recuperado por la segunda (B2) pero no por la primera (B1). En esta situación, la segunda búsqueda presentará unos niveles mayores de *exhaustividad* frente a la primera y unos niveles de *precisión* algo más bajos.

El caso contrario se presenta también frecuentemente: un usuario plantea una ecuación de búsqueda demasiado general, con la que seguramente recuperará la mayoría de los documentos relevantes con el tema de la cuestión, pero, al mismo tiempo recuperará muchos documentos que no resultan relevantes. Esto implicará que los valores de *precisión* se reduzcan sustancialmente. Si por ejemplo, los usuarios del ejemplo anterior, hubieran realizado estas búsquedas:

b1: "contaminación"

b2: "contaminación en los ríos"

Este es un claro ejemplo de lo que venimos comentando, en tanto que ahora el primer usuario, con su operación de búsqueda (b1), en lugar de presentar un exceso de especificidad en su planteamiento, presenta un defecto, ya que va a recuperar con toda seguridad la totalidad de documentos de la base de datos sobre contaminación fluvial, pero, de igual forma, va a recuperar documentos sobre contaminación atmosférica o contaminación acústica, documentos que, obviamente, no son relevantes con el tema de la búsqueda y cuya presencia en el conjunto de documentos recuperados provocará que el nivel de *precisión* sea bastante bajo.

Al significativo problema mencionado anteriormente de la imposibilidad de determinar con exactitud el valor de la *exhaustividad*, Korfhage añade otros dos: "en segundo lugar, no está claro que la *exhaustividad* y la *precisión* sean medidas significativas para el usuario" [KOR, 1997].

De hecho, la mayoría de los usuarios consideran mucho más importante la *precisión*, relegando a la *exhaustividad* a un cometido secundario. Así, mientras la búsqueda proporcione información relevante con la materia objeto de la necesidad informativa, el usuario no se detiene a pensar en la cantidad de documentos relevantes que no recupera, aunque este razonamiento no puede aplicarse como regla general en todos los SRI (como es el caso de las bases de datos jurídicas, donde es necesario garantizar un alto nivel del *exhaustividad*, si queremos estar en posesión de todos los precedentes legales sobre un tema objeto de litigio con el fin de que un abogado fundamente su argumentación ante el tribunal en las debidas condiciones y con las mayores posibilidades de éxito).

Medidas orientadas al usuario.

El tercer problema apuntado por Korfhage, reside en el hecho de que las medidas basadas en la *relevancia* están excesivamente vinculadas con la persona que lleva a cabo la evaluación y resultan de difícil traslado a otra persona, tal como alude Baeza-Yates: "ambas medidas se basan en el supuesto de que el conjunto de documentos relevantes para una respuesta es siempre el mismo, independientemente del usuario que lleva a cabo la evaluación. Esta situación, evidentemente, no responde adecuadamente a la realidad, en tanto que diferentes usuarios pueden tener una interpretación desigual de qué documento es relevante y cuál no lo es" [BAE, 1999]. Para solucionar este problema, [BAE, 1999], [KOR, 1997] y [SAL, 1983] presentan una serie de nuevas medidas en las cuales se parte del supuesto de que los usuarios forman un grupo homogéneo, de similar respuesta en el proceso de determinación de la *relevancia* del resultado de una operación de búsqueda. Esta situación, aunque algo difícil de asumir en el mundo real, permite el desarrollo de una serie de nuevas medidas denominadas "*Medidas Orientadas al Usuario*".

Korfhage indica que este nuevo conjunto de medidas "fue propuesto por Keen a principio de la década de los setenta. Hay tres comunes:

1. *Cobertura*: proporción de los documentos relevante conocidos que el usuario ha recuperado
2. *Novedad*: proporción de los documentos recuperados relevantes que eran previamente desconocidos para el usuario
3. *Exhaustividad Relativa*: la ratio de los documentos relevantes recuperados examinados por el usuario entre el número de documentos que el usuario está dispuesto a examinar." [KOR, 1997].

De esta forma, la obtención de un valor alto de cobertura indica que el sistema ha localizado la mayoría de los documentos relevantes que el usuario esperaba encontrar. Un valor alto de novedad indica que el

sistema ha mostrado al usuario una considerable cantidad de documentos, los cuales desconocía previamente.

Ambos autores sugieren una cuarta medida orientada al usuario: la conocida como "*esfuerzo de exhaustividad*", entendida como "la ratio entre el número de documentos relevantes que el usuario espera encontrar y el número de documentos examinados en un intento de encontrar esos documentos relevantes" [BAE, 1999]. Esta medida, para Korfhage, parte de dos supuestos primordiales: "la colección contiene el número deseado de documentos relevantes y el SRI permite al usuario localizarlos todos" [KOR, 1997].

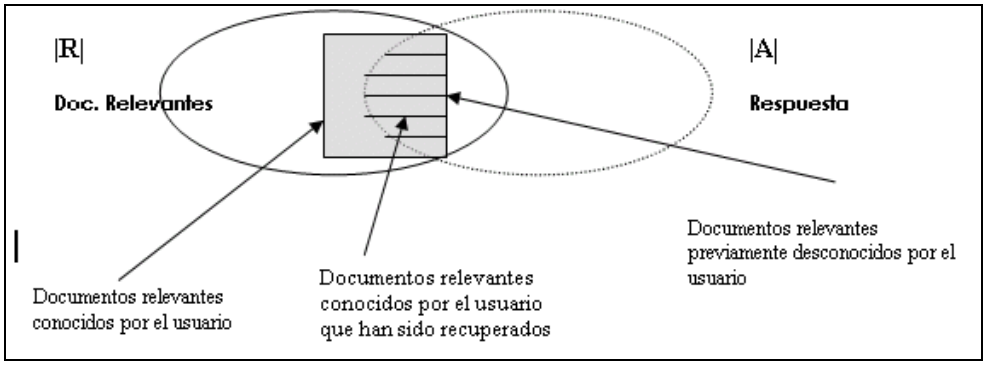


Ilustración 3.3 Distribución de documentos relevantes y de documentos integrantes de una respuesta a una operación de búsqueda. Fuente: Baeza-Yates, R. and Ribeiro-Neto, B. Modern information retrieval. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999.

Cálculo de la *Precisión* y de la *Exhaustividad*.

Si bien la *precisión* de una operación de recuperación de información puede ser calculada fácilmente, el cálculo de la *Exhaustividad* se presenta inviable, "solamente puede ser estimado" [BLA, 1990].

La estimación de esta medida ha constituido el objeto de trabajo de varios investigadores. Blair lleva a cabo una amplia revisión de algunos de estos trabajos, que vamos a ir presentando de forma sintetizada. Uno de los primeros métodos empleados consistió en limitar el tamaño de la base de datos, por ejemplo, a algo menos de 1000 documentos (cantidad que se entiende como factible para localizar la información documento a documento), y calcular el valor de la *exhaustividad* una vez analizados todos los documentos. Este método ha resultado de gran aceptación y su posible justificación procede del campo de la experimentación en Física, donde los resultados a pequeña escala producían resultados generalmente extrapolables al mundo real. Aunque, por desgracia, éste no es el caso de los SRI, donde un cambio cuantitativo de la colección provoca un cambio cualitativo en el modo que la recuperación de información ha de llevarse a cabo.

Si por ejemplo, tenemos un SRI cuya base de datos posee un tamaño de 1000 documentos, y un usuario recupera el 10% de los documentos de la base de datos (es decir, 100), a este usuario le puede resultar tedioso y aburrido analizar esta cantidad de documentos uno a uno (para verificar la *relevancia* o no del mismo con la materia objeto de su necesidad de información), pero esa operación resulta posible.

En cambio, si se supone que la misma técnica de recuperación y el mismo vocabulario controlado de términos, es aplicado a una colección de 100.000 documentos, es lógico presumir que el resultado rondará la cantidad de 10.000 documentos recuperados (diez veces mayor que el tamaño de la base de datos anterior), extensión absolutamente inmanejable para el usuario, quien se vería obligado a introducir nuevas expresiones en su ecuación de búsqueda con el fin de limitar la extensión del conjunto de documentos resultado (es decir, tendría que refinar su búsqueda), esta nueva serie de expresiones implica una variación de la estrategia de recuperación que ya no será igual que la efectuada sobre la búsqueda.

Blair cita a Resnikoff, para quien "las pruebas a pequeña escala no dicen mucho sobre el rendimiento del SRI o sobre las estrategias óptimas de recuperación para sistemas del mismo tipo pero mayores en tamaño" [BLA, 1990], y a Swanson, quien opina que "un argumento estadístico simple mostrará que las pruebas sobre colecciones de unos pocos cientos de documentos no son lo suficientemente sensitivas ni siquiera para permitir una aproximación del rendimiento en colecciones de cientos de miles de documentos, que son sobre los que debemos realizar las pruebas" [BLA, 1990].

Lancaster, en el desarrollo de la evaluación de MEDLARS, determina de antemano los valores de la *exhaustividad* [LAN, 1968], procediendo a lo que se conoce como una "estimación informal" de la *exhaustividad*. Este estudio de Lancaster aporta también un copioso análisis de las causas de los errores que se cometen en las operaciones de recuperación de información.

Otro procedimiento de cálculo de la *exhaustividad,* consiste en asignar a varias personas la tarea de analizar la documentación recuperada. Este procedimiento, además de más complejo y costoso en recursos humanos, contradice el sentido de la utilidad del documento recuperado para el ser humano que realiza la búsqueda (o *pertinencia*), ya que dos personas emiten, de forma ineludible, distintos juicios de valor y lo que puede ser bueno para uno no tiene por qué serlo para el otro.

Más lógica parece la idea de calcular la *exhaustividad* a partir de la toma de una muestra aleatoria de la colección documental, donde el usuario evaluará la *pertinencia* de los mismos y entonces, empleando técnicas estadísticas fiables, procederá a la estimación del número de documentos útiles de la colección.

Esta técnica, que parece viable y de sencilla realización, también presenta problemas, porque no está muy claro qué tamaño debe tener la muestra a analizar. Blair cita a Tague [BLA, 1990] quien avisa de la dificultad de llevar a cabo esta tarea en base de datos de muy baja *generalidad* (es decir, donde el porcentaje de documentos relevantes es muy bajo); en este caso el tamaño de la muestra debería ser muy grande y el análisis se complicaría.

Como ejemplo de lo anterior se puede presentar el siguiente caso: si la materia que deseamos localizar es demasiado específica, por ejemplo del orden de 1 documento pertinente por cada 100, la muestra que deberíamos manejar rondaría la cantidad de 5.000 documentos. Aunque este problema subyace, esta técnica ha resultado ser la más empleada por la mayoría de los estudios que hemos repasado, asumiendo todos ellos, que los niveles de *generalidad* no alcanzarán cotas tan bajas.

También se apunta una última técnica: realizar sondeos exploratorios en el resto de los documentos no recuperados. Este conjunto de estimaciones será más o menos fiable en función de la profundidad y alcance de los sondeos.

Blair prosigue presentado ampliamente esta problemática, incorporando las reflexiones de Sparck-Jones (quien recopila trabajos desarrollados a lo largo de treinta años en este campo), Swanson y Salton.

En todos los casos surge la duda, razonable por otra parte, del tamaño de la muestra, la cual, además de los problemas inherentes a su determinación, presenta otra problemática: difícilmente garantizará una similar distribución de la temática de los documentos recuperados fuera de ella.

Sparck-Jones critica también la realización de estudios donde los fallos resultan más achacables a un inadecuado (e incluso inexistente, a veces) método de estimación estadística de estos valores.

Con la idea de proporcionarnos algo de luz en este oscuro panorama, Salton [SAL, 1983] apuesta por calcular los valores de *exhaustividad* y *precisión* sobre una muestra de documentos de la colección total.

Realmente Salton[50] no llega a afirmar en ningún momento que los resultados de este análisis puedan trasladarse sin problema alguno a la globalidad de una gran base de datos, pero como afirma que no existen evidencias de lo contrario, sugiere que puede hacerse.

Esta actitud positivista, es absolutamente contrarrestada por la opinión de Swanson, quien aboga por encontrar pruebas de su corrección

[50] Salton defiende esta postura en el curso del estudio de comparación del rendimiento de los sistemas de recuperación de información STAIRS y SMART, que hemos citado anteriormente.

antes de aplicarla [BLA, 1990], actitud loable, sin duda alguna, pero que no aporta solución a este problema.

Aceptando que el cálculo de la *precisión* y *exhaustividad* debe llevarse a cabo sobre una muestra pequeña de la amplia colección de documentos de la base de datos, se va a exponer a continuación cómo se realiza este cálculo. En primer lugar, se supone que se elige una muestra constituida por los primeros ocho documentos (d1, d2, ... , d8) recuperados en una búsqueda Q, en la que resultan pertinentes los documentos {d1, d2, d4, d6, d8}.

Salton entiende que los cálculos *Exhaustividad-Precisión* (E-P en adelante), deben realizarse documento a documento recuperado, es decir, no son iguales el par de valores E-P en el primer documento que en el segundo. Siguiendo lo indicado por Salton los valores de *exhaustividad* y *precisión* calculados son los siguientes:

Exhaustividad – Precisión			
N	Relevante	E	P
d1	X	0.2	1
d2	X	0.4	1
d3		0.4	0.66
d4	X	0.6	0.75
d5		0.6	0.60
d6	X	0.8	0.66
d7		0.8	0.57
d8	X	1	0.625

Tabla 3.8 Cálculo de pares de valores E-P de la búsqueda ejemplo. Fuente: elaboración propia.

Cuando realizamos los cálculos en el primer documento (d1), se ha recuperado un único documento que es pertinente y, por tanto, la *precisión* valer uno (un acierto en un intento) y la *exhaustividad* (resultado de dividir el valor de uno entre el *total de documentos relevantes de la muestra*, valor que sí conocemos de antemano y es cinco), vale 0.2. Así, el documento d1 tiene asignado el par de valores E-P (0.2, 1). A continuación, procedemos a calcular el par de valores E-P de d2, también relevante, aquí la *precisión* será el resultado de dividir el valor de dos documentos relevantes recuperados (d1 y d2) entre el total de documentos recuperados hasta el momento (dos también), por lo que adquiere de nuevo el valor de la unidad; la *exhaustividad* será el resultado de dividir el valor de dos (ambos son relevantes) entre el *total de*

documentos relevantes de la muestra (cinco), obteniéndose un valor de 0.4, por lo que al documento d2 se le asignaría el par de valores E-P (0.4,1). Siguiendo este método se determinan el resto de los pares de valores E-P para los seis restantes documentos recuperados. Este conjunto de ocho pares de valores caracterizará, en principio, a la búsqueda Q, y se podría construir un gráfico E-P similar al reflejado en la ilustración 3.4:

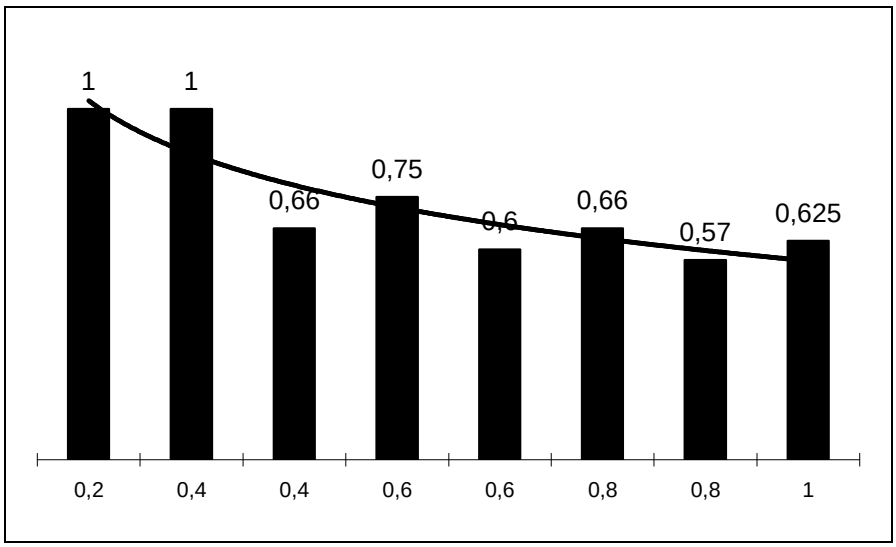


Ilustración 3.4 Evolución distribución pares de valores E-P del ejemplo. Fuente: elaboración propia.

Gráficos como el de la anterior ilustración son muy criticados por no representar claramente algunos parámetros, tales como "el tamaño del conjunto de documentos recuperados y el tamaño de la colección" [SAL, 1983]. Junto a esta serie de deficiencias, se añade que los problemas aumentan al tratarse de un gráfico que muestra una sucesión discreta de valores E - P en lugar de una sucesión continua de los mismos.

Sirva como ejemplo que la Ilustración 3.4 no indica qué valor de *precisión* corresponde a un valor de *exhaustividad* de 0.4, ya que el mismo varía desde el valor inicial de 1 hasta el de 0.66. Este problema aumentaría bastante si en lugar de representar el resultado de una búsqueda individual, se quisiera representar los resultados de varias, con el fin de determinar el comportamiento promedio del sistema. Salton opina que "es conveniente sustituir las curvas con 'dientes de sierra' de las preguntas individuales, por líneas atenuadas que simplifiquen el proceso de representar los promedios" [SAL, 1983]. Además, en el gráfico anterior, se detecta una cierta divergencia entre los valores de la línea de regresión de los pares E-P y los representados en el gráfico. Resulta vital que la representación de la evolución de los pares de valores E-P muestre un único valor de *precisión* para cada valor de *exhaustividad*, truncando (si es preciso), a un valor intermedio de la *precisión* las situaciones de similar valor de la *exhaustividad*. Así, Salton renombra la tradicional formulación

de las medidas *exhaustividad* y *precisión*, proponiendo una "*exhaustividad* instantánea" y una "*precisión* instantánea", distintas para cada operación.

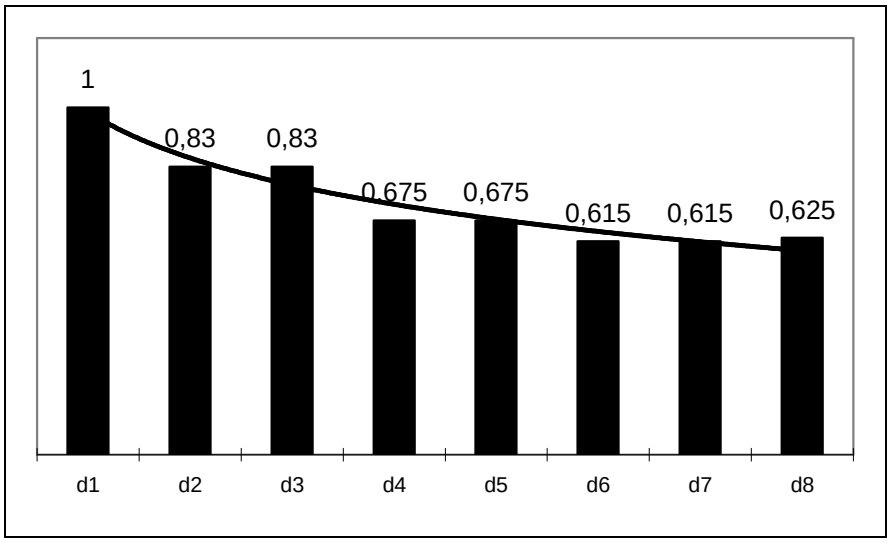


Ilustración 3.5 Evolución distribución valores E-P del ejemplo corregida. Fuente: elaboración propia.

En esta ilustración se determinan valores similares de *precisión* para valores análogos de *exhaustividad*, y la línea de regresión E-P se aproxima mucho más a los valores representados en el gráfico de barras.

Medidas Promedio E-P.

Por estos problemas, Salton propone calcular las medidas E-P en términos de promedio, reflejando que "el promedio que el usuario puede esperar de la realización de las búsquedas por parte del sistema, puede ser calculado tomando la media aritmética sobre un número de N búsquedas de la *exhaustividad* y de la *precisión* individuales de cada una de ellas".

***Exhaustividad* = 1/N * $\Sigma_{i=1}$ (RecRel$_i$ / RecRel$_i$ + NoRecRel$_i$)** ***Precisión* = 1/N * $\Sigma_{i=1}$ (RecRel$_i$ / RecRel$_i$ + RecNoRel$_i$)**	
RecRel: documentos recuperados relevantes	**NoRecRel**: documentos no recuperados relevantes
RecNoRel: documentos recuperados no relevantes	**NoRecNoRel**: documentos no recuperados no relevante

Tabla 3.9 Formulación de las Medidas Promedio E-P. Fuente: Salton , G. and Mc Gill, M.J. *Introduction to Modern Information Retrieval*. New York: Mc Graw-Hill Computer Series, 1983.

Esta formulación permite representar una curva E-P con valores diferentes de *exhaustividad* para cada valor de *precisión* siendo la función ahora continua y la representación gráfica coincide con la tradicional curva de Rijsbergen.

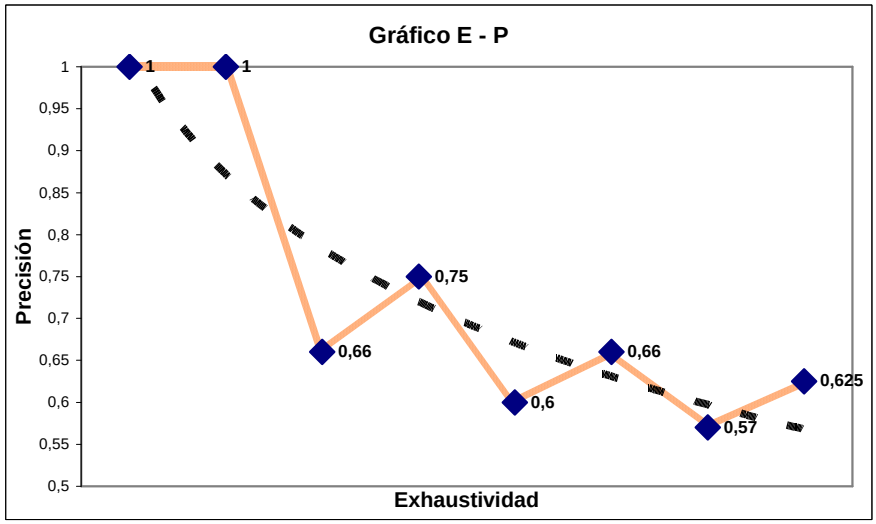


Ilustración 3.6 Representación de la evolución E-P real y óptima del ejemplo.
Fuente: elaboración propia.

Ahora, la curva E-P de la búsqueda Q presenta la evolución real de los valores obtenidos, más cercana a la tendencia natural de los pares de valores E-P (línea discontinua). Este método de cálculo se conoce como *exhaustividad y precisión relativas*, entendiéndose estas medidas como aproximaciones a los verdaderos valores de ambas ratios, sencillas de calcular y base sólida para la evaluación de los SRI.

Korfhage presenta dos maneras de calcular el promedio de la *exhaustividad* y de la *precisión*. La primera consiste en "calcular los promedios de la *precisión* para un conjunto de tres o de once valores previamente establecidos de la *exhaustividad*" o "promedio en tres puntos" y "promedio en once puntos" [KOR, 1997]. El segundo de los procedimientos aportados por Korfhage es "una forma completamente diferente de obtener los promedios. Se parte del supuesto de que en una evaluación, los documentos relevantes para cada conjunto de preguntas son conocidos a priori. Si también se supone que cada pregunta no se realiza hasta que una determinada condición sea satisfecha (como puede ser el recuperar un determinado número de documentos). Tanto la *precisión* como la *exhaustividad* pueden ser medidas en ese punto, obteniéndose un par de valores para cada pregunta. A partir de estos valores, se puede proceder a la construcción de una tabla E-P aumentando ambas medidas en un valor de 0.1" [KOR, 1997]. Ambos procedimientos proporcionan, generalmente, diferentes resultados, en tanto que analizan el funcionamiento del sistema de diferente manera, llegándose, a veces, a conclusiones harto dispares.

Medidas alternativas a E-P o medidas de valor simple.

Rijsbergen presenta un amplio conjunto de medidas alternativas a la relevancia. Todas pretenden superar los problemas antes estudiados de las medidas *E-P*. Casi todas usan técnicas probabilísticas y resultan difíciles de calcular [RIJ, 1999]. La mayoría de los autores presentan a estas medidas por separado, no llegando a ponerse de acuerdo en su denominación ("alternativas", "valor simple" u "otras medidas"). Salton las denomina como "de valor simple", porque no representan el resultado de una evaluación en función de un par de valores sino de un único valor, que puede ser objeto de clasificación. Posteriormente presenta el *Modelo de Swet* (que desarrolla la "Medida E"), las *medidas SMART* (sistema de indización automática desarrollado por Salton) y la *longitud esperada de búsqueda* basada en el *Modelo de Cooper* [SAL, 1983].

Rijsbergen es el primero que muestra una medida relacionada con el valor simple, la *satisfacción*. El autor cita a Borko, para quien este factor de satisfacción es el resultado de sumar los valores de *precisión* y de *exhaustividad*.

Otras medidas de esta naturaleza se han definido en diversos trabajos de otros autores, siendo todas ellas un instrumento útil para valorar la calidad de una búsqueda, aunque luego veremos que diversos autores opinan que no es suficiente (en todos los casos) realizar una operación aritmética para conseguirlo, y que harán falta otras medidas para estimar esta satisfacción.

Autor	**Expresión**
Borko	$I = E + P$
Meadow	$M = 1 - \frac{\sqrt{(1-P^2)+(1-R^2)}}{\sqrt{2}}$
Heine	$D_1 = 1 - \frac{1}{\frac{1}{P}+\frac{1}{R}-1}$
Vickery	$V = 1 - \frac{1}{\frac{2}{P}+\frac{2}{R}-3}$

Tabla 3.10 Medidas de la calidad de una búsqueda propuestas por varios autores. Fuente: Meadow, C. T. *Text Information retrieval Systems*. San Diego: Academic Press, 1993.

Los críticos a las medidas de valor simple basan sus opiniones fundamentalmente en la escasa capacidad que confieren a la *exhaustividad* y a la *precisión* en la representación de la efectividad de

una búsqueda. Tampoco se debe olvidar que cada medida tiene una interpretación diferente, en las propuestas por Borko y Meadow, un valor mayor equivale a una búsqueda mejor (o más efectiva), mientras que en las medidas de Heine y Vickery, ocurre justo al contrario. Como justificación sirve el siguiente ejemplo: un usuario realiza una búsqueda "ideal" en la que recupera diez documentos de los cuales todos son relevantes y en toda la base de datos sólo están estos diez documentos relevantes. En este caso "ideal", el valor de E es igual a 1 y el valor de P es igual a 1, siendo la suma igual a 2 (lo máximo que alcanza la medida de Borko), mientras que la medida de Meadow valdrá 1 (su máximo) y las otras dos medidas valdrían cero. Si se supone a continuación, que la cantidad de documentos relevantes en la base de datos es de 20, el valor de P seguirá siendo 1 pero el valor de E descenderá a 0.5; en este supuesto (que tampoco corresponde con una búsqueda que pueda calificarse de mala), la medida de Borko valdrá ahora 1.5, la medida de Meadow valdrá 0.38 (mucho menos que antes), la medida de Heine tendrá un valor de 0.5 y la medida de Vickery tendrá un valor de 0.66.

En último lugar, si se imagina que sólo la mitad de los documentos recuperados son relevantes, P valdría ahora 0.5 y E valdría 0.25 (supuesto algo más desfavorable que en los dos casos anteriores), obteniéndose un valor de 0.75 para la medida de Borko, un valor de 0.07 para la medida de Meadow, de 0.8 para la medida de Heine y de 0.88 para la de Vickery. Así, a mayor efectividad de la búsqueda, los valores de las medidas de Borko y de Meadow serán mayores y, por el contrario, los valores de las medidas de Heine y Vickery decrecerán.

Quedan otras medidas alternativas o de valor simple propuestas. Rijsbergen presenta el *Modelo de Swet*, el *Modelo de Robertson*, el *Modelo de Cooper* y las medidas SMART [RIJ, 1999]. Korfhage separa las *medidas simples* de la medida *longitud esperada de búsqueda*, pasando a comentar luego las medidas de *satisfacción* y *frustración* [KOR, 1997]. Baeza-Yates es mucho más sintético en la presentación de estas medidas. Considera que se engloban dentro de la determinación de valores de E-P [BAE, 1999]. Posteriormente se refiere a la "Medida E" de Swet (dedicándole un párrafo) y cita la existencia de otras medidas (sin explicar cómo se obtienen), entre ellas la medida de *satisfacción* y de *longitud esperada de búsqueda*. En lo que coinciden casi todos los autores es en desarrollar unas medidas alternativas a las tradicionales, "algunos observadores repudian por completo la tabla de contingencia como base para la construcción de parámetros capaces de reflejar la efectividad de la recuperación de información" [SAL, 1983]. Para este autor, las medidas a emplear deberían cumplir las siguientes condiciones:

1. Deben ser capaces de reflejar la efectividad de la recuperación únicamente.
2. Deben ser independientes de cualquier límite, es decir, el número de documentos recuperados en una búsqueda específica no debe influenciar a estas medidas.

3. Deben ser expresadas en un número simple, en lugar de hacer uso de un par de valores (tales como E-P).

Modelo de Swets.

Bajo estas premisas se desarrolla el *Modelo de Swets*, explicado en [RIJ, 1999], [SAL, 1983]. Este modelo define la terna E-P-F (*exhaustividad-precisión*-tasa de fallo), en términos probabilísticos. La *exhaustividad* será una estimación de la probabilidad condicionada de que un documento recuperado sea relevante; la *precisión* será una estimación de la probabilidad condicionada de que un documento relevante sea recuperado y la tasa de fallo una estimación de la probabilidad condicionada de que un documento recuperado no sea relevante. Este modelo convierte las representaciones E-P en una serie de funciones basado en el cómputo de un número de documentos relevantes o no. Las búsquedas dan como resultado unas funciones lineales que guardan una cierta distancia con la distribución de las probabilidades anteriormente citadas. Es precisamente el valor de esa distancia (multiplicado por la raíz cuadrada de dos) el valor de la "Medida E de Swets".

El ejemplo de la Ilustración 3.7 muestra dos líneas (A y B) que propician el cálculo de dos distancias (DIST2 y DIST1 respectivamente), que derivarán en dos valores distintos de la medida de Swets. También resultan determinables las pendientes de ambas líneas. La principal ventaja de este modelo y de esta medida reside en "que la misma se basa en una teoría estadística suficientemente reconocida y aceptada" [SAL, 1983].

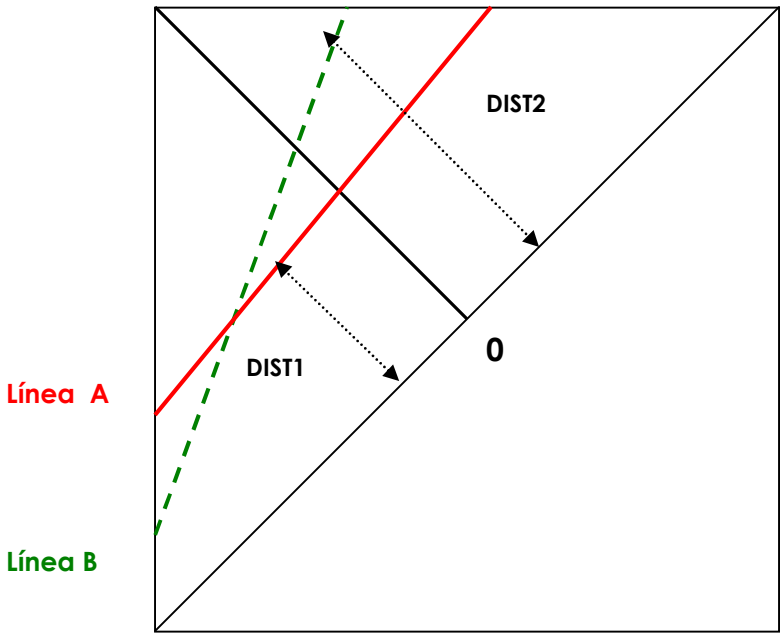


Ilustración 3.7 Ejemplo del Modelo de Swets. Fuente: Salton, G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983..

Esta medida, "a diferencia de la E-P, estas medidas (la distancia y la pendiente) no son fácilmente descifrables por los usuarios y difícilmente las usarán. Asimismo, el valor de E[51] no puede obtenerse a partir de un simple umbral que distinga documentos relevantes de no relevantes, tal como se hace en las evaluaciones convencionales; en cambio, un simple par de valores E-P es fácil de calcular en todo momento" [SAL, 1983]. Esta opinión coincide con la realidad, "otras medidas globales basadas en sólidas teorías, especialmente en la teoría de la probabilidad y en la teoría de la información, han sido descritas por diversos autores pero no han recibido consideración en la práctica" [SAL, 1983].

Modelo de Robertson.

Rijsbergen presenta el *Modelo de Robertson,* una aproximación logística a la estimación de los valores E-P". Robertson y Teather desarrollan un modelo estimando las probabilidades correspondientes a E-P basado en un procedimiento inusual ya que al calcular una estimación de ambas probabilidades para una pregunta simple, toman en consideración dos cosas: la cantidad de datos empleados para alcanzar esas estimaciones y los promedios de las estimaciones de todas las demás preguntas" [RIJ, 1999]. El objetivo de este método no es otro que obtener un valor, denominado "delta" que puede considerarse como candidato para ser una medida simple de la efectividad de un SRI [RIJ, 1999].

Modelo de Cooper.

En 1968, Cooper estableció que "la función primaria de un SRI es poner a salvo a sus usuarios, en la medida que esto sea posible, de la tarea de leer detenidamente todo el conjunto de documentos recuperados en la búsqueda, para discernir los relevantes" [RIJ, 1999]. Este "ahorro de esfuerzo" es lo que se debe medir y para ello sólo haría falta una medida simple. Esta medida sólo se aplicaría a los sistemas que mostraran la salida de los documentos ordenados según un criterio de alineamiento. Salton indica que "algunas medidas adicionales de valor único, emplean también las diferencias entre los rangos de los documentos relevantes recuperados y, o bien los rangos ideales (aquellos casos donde los documentos relevantes son recuperados antes que los no relevantes), o bien los rangos aleatorios (donde los documentos relevantes son aleatoriamente incluidos en la salida entre los no relevantes). Una de estas medidas es la *longitud esperada de búsqueda*." [SAL, 1983].

En [SAL, 1983] y [KOR, 1997] recurren al mismo ejemplo de cálculo de esta medida. Se supone que el conjunto de documentos recuperados se divide en K subconjuntos (S_1, ..., S_k), tal que los elementos de cualquiera de los subconjuntos sean equivalentes pero de manera que los documentos del subconjunto S_i sean mejores que los documentos del subconjunto S_{i+1}. Si la búsqueda realizada es una *booleana disyuntiva*

[51] El valor de la medida E de Swets.

(unión de dos o más expresiones o condiciones), que produce un resultado { t_1 OR t_2 OR t_3 OR .... t_n }, entonces S_1 podría consistir en aquellos documentos que contienen todos los términos de la expresión disyuntiva, S_2 en aquellos que contuvieran sólo n-1 términos, S_3 sería el conjunto de los documentos con sólo n-2 términos y así sucesivamente.

Para presentar estos documentos en orden, se desarrolla un conjunto débilmente ordenado de documentos. Esto es, todos los documentos de cualquier subconjunto de igual rango, pero peor que el precedente y mejor que el siguiente[52]. En estos términos la longitud esperada de búsqueda puede ser definida como el número promedio de documentos no relevantes que pueden ser llegar ser examinados por el usuario, antes de acceder al número deseado de documentos. Si se supone que el conjunto de documentos recuperados se divide en tres subconjuntos {S_1, S_2 y S_3}. S_1 contiene tres documentos, de los cuales sólo uno es relevante; S_2 contiene cuatro documentos relevantes y uno que no lo es, y, finalmente, S_3 contiene dos documentos relevantes y tres que no lo son. Si el usuario sólo desea recuperar un documento relevante, al usuario le basta con el subconjunto S_1. Como el documento relevante de S_1, puede ser el primero, el segundo o el tercero de los examinados y, asumiendo que todas las combinaciones son iguales, el usuario tiene una probabilidad de 1/3 de acceder al documento deseado al instante. Así, la longitud esperada de búsqueda, sería el resultado de sumar las probabilidades: (1 * 1/3) + (2 * 2/3) + (3 * 3/3), lo que totaliza un valor de dos, que significa: "como promedio, el usuario necesitará examinar un documento no relevante antes de encontrar el documento relevante que desea" [KOR, 1997].

Si ahora se supone que el usuario desea recuperar seis documentos relevantes, el examen de los subconjuntos S_1 y S_2 sólo le proporcionaría 5 documentos relevantes, por lo que ha de localizar el sexto en el subconjunto S_3. Forzosamente, el usuario tendrá que examinar los ocho documentos incluidos en S_1 y S_2 más los que le sean necesarios en S_3. Considerando dónde pueden encontrarse en S_3 los dos documentos relevantes, hay cuatro situaciones en los que el primero de ellos (el único de interés) está en la primera posición, tres situaciones en segunda, dos situaciones en tercer lugar y una única oportunidad donde está en el cuarto lugar, lo que totaliza un total de 10 posibles situaciones diferentes. En este caso, la *longitud esperada de búsqueda* sería el valor resultante de las operaciones: (9 * 4/10) + (10 * 3/10) + (11 * 2/10) + (12 * 1/10); "lo que resulta un valor de 10, es decir, que el usuario, como término medio, va a consultar cuatro documentos no relevantes antes de encontrar el sexto relevante deseado" [KOR, 1997].

[52] Muy bien puede aparecer sólo un documento en cada subconjunto o bien pueden llegar a aparecer todos los documentos en el mismo subconjunto.

Esta medida no proporciona directamente un valor simple. Proporciona una serie de valores que muestran qué puede esperar el usuario del sistema bajo distintos requerimientos de *exhaustividad*. No obstante, esta serie de valores pueden sintetizarse en una valor simple obtenido al dividir la longitud esperada de búsqueda correspondiente a cada valor distinto de documentos relevantes entre ese valor de documentos relevantes, sumar todas los cocientes y hallar la media aritmética.

Documentos Relevantes	**Longitud esperada de búsqueda**	**DocRel / Long**
1	2,0	2,0
2	4,2	2,1
3	5,4	1,8
4	6,6	1,65
5	7,8	1,56
6	10,0	1,67
7	12,0	1,71
	Suma	12,49
	E	12,49/7 = 1.78

Tabla 3.11 Ejemplo de determinación de la medida E de longitud esperada de búsqueda. Fuente: Korfhage, R.R. Information Retrieval and Storage. New York: Wiley Computer Publisher, 1997.

Exhaustividad y Precisión normalizadas.

Otro de los problemas que conlleva el uso de las medidas E-P es la secuencialidad de la lectura de los resultados: "los SRI típicos presentan los resultados al usuario formando una secuencia de documentos. Incluso en aquellos sistemas que no presentan así la información, el usuario suele examinar los documentos secuencialmente. Este modo de examinar va a afectar al juicio que el usuario ha de llevar a cabo sobre la *relevancia* o no de los documentos siguientes" [KOR, 1997].

Todos los usuarios de los SRI han sufrido este problema alguna vez, cuando, al consultar dos documentos más o menos igual de interesantes, centran su atención preferentemente en el primero de ellos, aunque el segundo no desmerezca en nada al anterior. Otra situación frecuente ocurre cuando un usuario realiza una búsqueda y los primeros documentos recuperados resultan relevantes. En esta circunstancia, el usuario va a tener una sensación positiva y se considerará satisfecho (no preocupándose por el número de documentos no relevantes que también recupera que puede llegar a ser muy grande). La situación contraria también se produce cuando los documentos no relevantes al principio de

la secuencia de documentos entregados son mayoría, entonces la sensación de frustración va a ser de cierta consideración, independientemente de que se entreguen muchos más documentos relevantes que no relevantes.

Esto propicia el desarrollo de medidas que tomen en consideración la secuencia en la que los documentos son presentados a los usuarios. La primera se debe a Rocchio [RIJ, 1999], [KOR, 1997], quien define una *exhaustividad normalizada* y una *precisión normalizada* para sistemas que presentan los documentos alineados según un criterio de clasificación y donde no afecte el tamaño de la muestra analizada. Rocchio define un "sistema ideal donde los documentos relevantes se recuperan antes que los documentos no relevantes y se puede representar en un gráfico la evolución de la *exhaustividad* de esta operación de búsqueda".

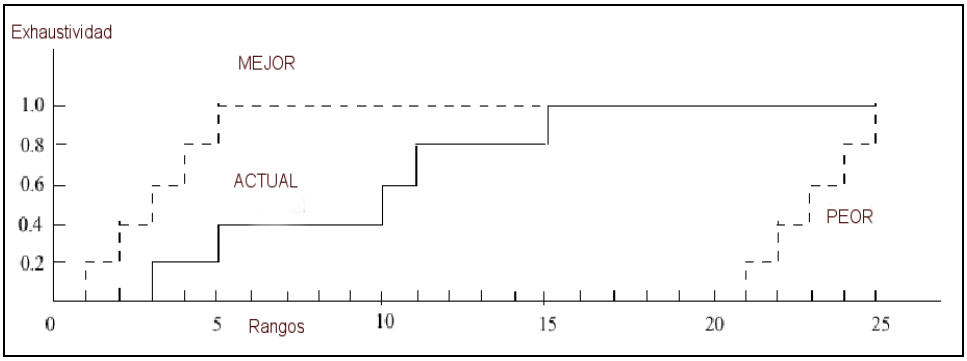


Ilustración 3.8 Ejemplo de cómo la *Exhaustividad* normalizada queda comprendida entre el pero y mejor resultado posible. Fuente: Rijsbergen, C.J. Information Retrieval. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 21 de marzo de 2002]

Si en una base de datos de 25 documentos se sabe que cinco de ellos son relevantes, y que han sido devueltos por el sistema al realizar una búsqueda en las posiciones {3, 5 , 10, 11, 15}, se puede representar el gráfico etapa a etapa como muestra la anterior Ilustración 3.8. La *exhaustividad* de esta búsqueda alcanza el valor de 0.2 (1 documento relevante recuperado dividido entre el total de 5 documentos relevantes de la colección) al analizar el tercer documento (hasta entonces vale cero).

Cada vez que analizamos un documento relevante, el valor de la *exhaustividad* aumenta hasta llegar a la unidad (en este caso, en el documento 15), permaneciendo constante hasta el último documento recuperado, porque todos los relevantes han sido recuperados. La comparación con la mejor búsqueda posible (cinco documentos relevantes en las cinco primeras posiciones de la secuencia) o con la peor búsqueda posible (correspondiente a presentar los cinco documentos relevantes en las cinco últimas posiciones de la secuencia), resulta muy intuitiva, tal como se puede ver en la anterior ilustración.

Para Korfhage, "el área comprendida entre la búsqueda actual y la gráfica ideal representa una medida de la ejecución del SRI" [KOR, 1997]. Esta medida se calcula restando al valor de la unidad el resultado de dividir el valor de esta área por (n1 * (N – n1))[53]. En el ejemplo anterior, el valor del área es 29, por lo que el valor de la *exhaustividad normalizada*, aplicando la anterior fórmula, será el resultado de la siguiente operación: (1 – 21/(5 * (25 – 5)) = (1 – 0.21) = 0.79

La *precisión normalizada* se define de manera análoga: "mientras la *precisión* ordinaria es una medida exactamente definida, esta medida depende del conocimiento del número total de documentos relevantes." [KOR, 1997]. Rijsbergen propone como método para su cálculo "restar a la unidad el resultado de dividir el valor de esta área por el valor del área existente entre la búsqueda ideal y la peor búsqueda" [RIJ, 1999]. En el ejemplo propuesto, el resultado de esta *precisión normalizada* sería (1 – 21/(95 – 15)) = (1 – 21/80) = 0.7375.

Rijsbergen destaca algunos aspectos de estas medidas: "ambas presentan un comportamiento consistente, es decir, cuando una se aproxima a cero la otra se aproxima a la unidad. Ambas medidas asignan valores distintos de peso a los documentos recuperados en la secuencia, la *precisión* los asigna a los iniciales y la *exhaustividad* asigna un valor uniforme a todos los documentos relevantes.

En tercer lugar, estas medidas pueden entenderse como una aproximación de la *precisión* y *exhaustividad* promedio (estudiadas anteriormente) y, por último, los problemas que surgían en la determinación de la longitud esperada de búsqueda (por la posición de los documentos relevantes), son inexistentes en este caso" [RIJ, 1999].

Ratio de deslizamiento.

Salton y Korphage presentan la *ratio de deslizamiento,* medida muy similar conceptualmente a la *exhaustividad normalizada y basada en* "la comparación de dos listas ordenadas de documentos recuperados (es decir, el SRI devuelve los resultados según un criterio de rango). Una lista es la salida de nuestro sistema actual, y la otra representa un sistema ideal donde los documentos recuperados se muestran en orden descendente" [SAL, 1983].

Este modelo es más complejo que el anterior, porque asigna pesos a los documentos en función de su *relevancia* con la pregunta realizada. La ratio se establece como el resultado de dividir la suma de los pesos de los documentos recuperados por nuestro sistema entre la suma de los pesos de los documentos que hubiera devuelto el sistema ideal. Este modelo sustituye la asignación binaria de *relevancia* de un documento, por la asignación de un peso.

[53] **n1** es el número de documentos relevantes y **N** el número total de documentos

La situación más favorable para un sistema evaluado es que la búsqueda realizada sea exactamente igual que la que ofreciera el sistema ideal, adquiriendo la ratio de deslizamiento un valor de uno.

Para ilustrar su cálculo, Korfhage propone el siguiente ejemplo: "supongamos que un sistema ha recuperado 10 documentos, con el siguiente peso: 7.0, 5.0, 0.0, 2.5, 8.2 4.5, 3.7, 1.1, 5.2 y 3.1, en el orden de recuperación. En un sistema ideal, estos documentos habrían sido recuperados y presentados en el orden descendente de pesos, y se podría calcular la tabla 3.12 de ratios:

Ratio de Deslizamiento			
N	**ΣPesos Reales**	**Σpesos Ideales**	**Deslizamiento**
1	7.0	8.2	0.85
2	12.0	15.2	0.79
3	12.0	20.4	0.59
4	14.5	25.4	0.57
5	22.7	29.9	0.76
6	27.2	33.6	0.81
7	30.9	36.7	0.84
8	32.0	39.2	0.82
9	37.2	40.3	0.92
10	40.3	40.3	1

Tabla 3.12 Ejemplo de determinación de la ratio de deslizamiento. Fuente: Korfhage, R.R. Information Retrieval and Storage. New York: Wiley Computer Publisher, 1997.

Esta medida, como las anteriores normalizadas, cuantifica la diferencia entre la secuencia de documentos que entrega un sistema real y la que entregaría un sistema ideal, tomando en cuenta las posiciones de los documentos relevantes y los no relevantes. En cambio, la ratio de deslizamiento aporta dos ventajas: "su uso de los pesos de *relevancia* y que sólo depende de los documentos recuperados" [KOR, 1997]).

Satisfacción y Frustración.

El esfuerzo que implica la determinación de las medidas anteriores, propicia que se establezcan otra nueva serie: *satisfacción*, *frustración* y *total* [KOR, 1997], [BAE, 1999].

La primera de estas medidas sólo considera los documentos relevantes, la segunda únicamente contempla a los no relevantes y la tercera combina ponderadamente las medidas anteriores. Cuando se usan los pesos, a los documentos relevantes se les suele asignar los valores

más cercanos al umbral superior de la escala, y a los no relevantes se les asigna valores cercanos al cero.

Para calcular la medida de la *satisfacción*, el peso de los documentos no relevantes se simplifica a cero, aunque se debe considerar sus posicionen en la secuencia.

Similarmente, para calcular la ratio de *frustración*, los documentos relevantes tendrán peso cero. La elección de un esquema de peso en la definición de la medida *total* se determina por el nivel de satisfacción que alcance el usuario cuando reciba los documentos relevantes pronto y su tolerancia a la presencia de documentos no relevantes.

Comparación medidas satisfacción, frustración y total para sistemas A y B									
	Ideal			Sistema A			Sistema B		
N	***S***	***F***	***T***	***S***	***F***	***t***	***S***	***f***	***T***
1	4	0	4	3	0	3	0	2	-2
2	8	0	8	7	0	7	4	2	2
3	11	0	11	9	0	9	6	2	4
4	14	0	14	9	2	7	9	2	7
5	17	0	17	11	2	9	11	2	9
6	19	0	19	14	2	12	11	4	7
7	21	0	21	17	2	15	11	5	6
8	21	1	20	21	2	19	14	5	9
9	21	3	18	22	3	19	17	5	12
10	21	5	16	22	5	17	21	5	16

Tabla 3.13 Ejemplo propuesto por Korfhage para determinar la medida de la satisfacción. Fuente: Korfhage, R.R. Information Retrieval and Storage. New York: Wiley Computer Publisher, 1997.

El modo más simple de determinar esta medida es calcular la diferencia entre las otras dos anteriores, aunque se han propuesto algunas expresiones más refinadas.

Korfhage propone el siguiente ejemplo: "supongamos que vamos a comparar dos SRI, cada uno de los cuales recupera los mismos 10 documentos. Estos documentos serán juzgados siguiendo una escala de 5 puntos, donde 0 y 1 representan documentos no relevantes; y 2, 3 y 4 representan documentos relevantes.

El sistema A recupera los documentos en el orden {3, 4, 2, 0, 2, 3, 3, 3, 4, 1, 0} y el sistema B recupera los documentos en el orden {0, 4, 2, 3, 2, 0, 1, 3, 3, 4} Si se entiende que la medida *total* va a ser el resultado de

restarle el valor de frustración al valor de satisfacción, vamos a proceder a representar estos valores en la siguiente tabla[54]" [KOR, 1997].

Aunque, a primera vista, los valores de satisfacción del sistema A están más cerca que los del sistema B del sistema ideal, se puede resumir la tabla anterior en la siguiente tabla 3.14, que Korfhage denomina: "tabla de diferencias de áreas".

Tabla de diferencias de área						
	dS		***dF***		***dT***	
N	***A***	***B***	***A***	***B***	***A***	***B***
1	1	4	0	2	1	6
2	1	4	0	2	1	6
3	2	5	0	2	2	7
4	5	5	2	2	7	7
5	6	6	2	2	8	8
6	5	8	2	2	7	12
7	4	10	2	5	6	15
8	0	7	1	4	1	11
9	0	4	0	2	0	6
10	0	0	0	0	0	0

Tabla 3.14 Tabla de "diferencias de áreas" para el jemplo de determinación de la ratio de deslizamiento. Fuente: Korfhage, R.R. Information Retrieval and Storage. New York: Wiley Computer Publisher, 1997.

En este ejemplo, se observa que el sistema A es, al menos igual de bueno que el sistema B, aunque realmente es bastante mejor en casi todos los niveles para las tres medidas. Todo este conjunto de medidas que pretenden calcular el nivel de satisfacción del usuario de un SRI resultan, para Meadow, "las de mayor utilidad de todas las empleadas normalmente si el objetivo es establecer el promedio del proceso de la recuperación de información" [MEA, 1997].

Medida de Voiskunskii.

El ejemplo anterior introduce una nueva dimensión en la evaluación de la recuperación de información, no centrada únicamente en la comparación de dos sistemas A y B sino preocupada también en intentar discernir "uno de los más importantes y complejos problemas de las Ciencias de la Información, la creación de un mecanismo de selección

[54] La representación de la tabla propuesta por Korfhage es la Tabla 3.13.

que permita elegir el mejor método de búsqueda (de entre las posibles variaciones) para una cuestión Q" [VOI, 1997].

Los grandes problemas a los que alude Voiskunskii, residen básicamente en la ausencia de criterios sólidos y consistentes de comparación de los resultados de una búsqueda, a pesar de los esfuerzos de muchos autores centrados en ello en los últimos treinta años. El autor considera que estos criterios deben cumplir los siguientes requerimientos:

a. "Los criterios deben proveer una comparación pragmática y justificada de los resultados de la búsqueda; y

b. la cantidad de trabajo precisa para determinar la información que es requerida para el establecimiento de estos criterios debe ser admisible" [VOI, 1997].

Tradicionalmente se ha empleado la medida de valor simple propuesta por Borko $\mathbf{I_0 = E + P}$ (es decir, la suma de los valores de la *exhaustividad* y de la *precisión*) [RIJ, 1999], aunque estas dos medidas (*exhaustividad* y *precisión*) no cumplen totalmente los criterios comentados, debido, fundamentalmente, a que infieren el valor de la *exhaustividad*.

Para esta medida $\mathbf{I_0}$, una búsqueda será mejor que otra cuanto mayor sea el resultado de la suma, aunque la misma puede llevar, a veces, a una serie de conclusiones equivocadas. Como ejemplo de una situación de esta naturaleza, Frants, Shapiro y Voiskunskii enuncian el siguiente caso: "supongamos que sobre una colección de 10.000 documentos, de los cuales se consideran pertinentes para una determinada materia aproximadamente 100, se llevan a cabo tres operaciones de búsqueda con los siguientes resultados.

a. Se recuperan 100 documentos, 50 de ellos son pertinentes y el resto no lo es.

b. Se recuperan 67 documentos, siendo pertinentes 40 de ellos.

c. Por último, se recupera sólo un documento que resulta pertinente.

Si calculamos los valores de *exhaustividad* y de *precisión* vamos a obtener los siguientes valores de la medida $\mathbf{I_0}$:

Búsqueda	**E**	**P**	$\mathbf{I_0}$
A	0.5	0.5	1
B	0.4	0.597	0.997
C	0.01	1	1.01

Tabla 3.15 Ejemplo del cálculo de la medida I1 de Borko para el ejemplo propuesto por Frants, Shapiro y Voiskunskii. Fuente: Frants , V.I. et al. Automated information retrieval : theory and methods. San Diego [etc.] : Academic Press, cop.1997. XIV, 365 p.

La interpretación literal de estos valores indica que la mejor búsqueda es la "c", al ser el valor más alto" [FRA, 1997]. Si bien el cálculo está bien realizado, estos resultados resultan algo incongruentes cuando el observador se aleja un poco de la mera interpretación matemática, reflexiona y se da cuenta de que difícilmente la búsqueda "c" puede considerarse incluso "admisible", máxime cuando sólo entrega al usuario un único documento y es casi seguro que el usuario preferirá cualquiera de las otras dos búsquedas que le entregan más documentos, independientemente de lo que nos diga esta fórmula.

Para solucionar este problema, Frants, Shapiro y Voiskunskii proponen una nueva medida de valor simple $\mathbf{I_1}$ [FRA, 1997] [VOI, 1997], a partir de la ratio entre el cuadrado de documentos relevantes recuperados y el número de documentos que conforman el resultado, "ratio cuya formulación analítica se corresponde con la *raíz cuadrada del producto de los valores* E-P" [VOI, 1997]. Si en el ejemplo anteriormente planteado se aplicara esta nueva medida a las tres búsquedas realizadas, los resultados serían:

Búsqueda	**E**	**P**	$\mathbf{I_1}$
A	0.5	0.5	0.25
B	0.4	0.597	0.2388
C	0.01	1	0.01

Tabla 3.16 Ejemplo del cálculo de la medida I2 de Voiskunskii para el ejemplo propuesto por Frants, Shapiro y Voiskunskii. Fuente: Frants , V.I. et al. Automated information retrieval : theory and methods. San Diego [etc.] : Academic Press, cop.1997. XIV, 365 p.

En este caso, la medida $\mathbf{I_1}$ de Voiskunskii propondría a la búsqueda "a" como la mejor de las tres realizadas, conclusión algo más lógica y coherente que la anterior. Frants, Shapiro y Voiskunskii desarrollan otro ejemplo donde muestran una búsqueda en la que coinciden ambas medidas y otra en la que difieren (en la determinación de cuál es mejor búsqueda). Aunque, tal como el mismo Frants indica posteriormente: "es evidente que dos ejemplos no resultan suficientes para justificar plenamente las bondades de las medidas $\mathbf{I_0}$ y $\mathbf{I_1}$, aunque escapa al objeto de nuestro trabajo considerar todos los casos adecuados para dicha justificación. Limitamos nuestra exposición a los ejemplos previamente presentados, ejemplos que, en nuestra opinión, proporcionan una clara y suficiente ilustración de estas medidas" [FRA, 1997].

Unos años más tarde, Voiskunskii llega a desarrollar hasta nueve medidas más, medidas que, evidentemente van ganando en complejidad en su cálculo y entendimiento, aunque él mismo indica que "aplicando la medida I_2 en búsquedas de *precisión* mayor que 0.5, las conclusiones que se extraigan quedan suficientemente justificadas" [VOI, 1997].

Caso práctico de cálculo de medidas de valor simple.

En la práctica, la medida $\mathbf{I_0}$ de Borko y la medida $\mathbf{I_1}$ de Voiskunskii, suelen coincidir en sus resultados, excepto en casos extraordinarios (como puede ser el del ejemplo anterior, donde cabría pensar que ni siquiera vale la pena plantearse el determinar la calidad de una búsqueda, si los niveles de *exhaustividad* o de *precisión* disminuyen demasiado).

Para verificar los ejemplos que proporcionan los distintos autores que han abordado esta materia, se propone la realización de un pequeño ensayo consistente en intentar localizar en el motor de búsqueda *Google* información sobre "alquiler de apartamentos en Málaga". Para ello se han formulado tres preguntas, cuya sintaxis se transcribe a continuación:

1. B1: alquiler apartamentos Málaga
2. B2: "alquiler de apartamentos" Málaga
3. B3: ("alquiler de apartamentos" OR "alquiler de viviendas" OR "alquiler de alojamientos") AND Málaga

Como se desprende fácilmente de la sintaxis de estas expresiones, a medida que se han realizado las búsquedas, se ha aumentado la complejidad de la ecuación de búsqueda, haciendo uso de las prestaciones que este SRI proporciona. A continuación se han analizado los treinta primeros documentos recuperados en B1, B2 y B3, observando que aumentan los niveles de E y P relativos calculados para cada uno de ellas, lo que propicia la siguiente evolución de las distintas medidas estudiadas.

	E	P	I_0	I_1	Meadow	Heine	Vickery
B1	0,5159	0,8045	1,32	0,64	0,26	0,54	0,70
B2	0,5282	0,9018	1,43	0,69	0,32	0,50	0,66
B3	0,5512	0,9462	1,49	0,72	0,36	0,46	0,63

Tabla 3.17 Ejemplo del cálculo de las medidas de valor simple estudiadas para el ejemplo propuesto. Fuente: elaboración propia.

Estos datos confirman que, en situaciones normales, todas las medidas responden del mismo modo, indicando que la mejor búsqueda es B3 ($\mathbf{I_0}$ de Borko, $\mathbf{I_1}$ de Voiskunskii y medida de Meadow, aumentando el valor frente a los valores de B1 y B2; mientras que las medidas de Heine y Vickery disminuyen su valor en B3 frente a los valores de las otras dos búsquedas analizadas).

Esta coincidencia permite afirmar, como hacen todos los autores que abogan por el uso de medidas simples, que la efectividad de una operación de búsqueda en un SRI puede medirse en términos de una medida simple, tal como puede ser cualquiera de las expuestas, o bien alguna que se desarrolle "ad hoc" para la evaluación de SRI que presenten algunas variantes que le diferencien de los empleados por

norma general, tal como puede ser el caso de los sistemas que permiten la recuperación de información en la web, donde tienen lugar fenómenos que no suelen encontrarse en los SRI convencionales y que debemos tener en consideración a la hora de reflexionar sobre la viabilidad de una medida de esta naturaleza.

4 Casos prácticos.

1. En un SRI basado en el modelo del espacio vectorial, calcula la similitud de los siguientes vectores de documentos A y B por medio de la función del coseno.

Vector	T1	T2	T3	T4	T5	T6	T7
A	1	1	0	0	1	1	1
B	1	0	0	0	1	0	1

Solución.

La función de similitud del coseno se define por la siguiente expresión:

$$Cos(P,N) = \sum_{i=1}^{n}(p_i * n_i) \Big/ \sqrt{\sum_{i=1}^{n} p_i^2} * \sqrt{\sum_{i=1}^{n} n_i^2}$$

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

Equivale a calcular el producto escalar de los vectores A y B y dividirlo por la raíz cuadrada del sumatorio de los componentes del vector A multiplicada por la raíz cuadrada del sumatorio de los componentes del vector B. El producto escalar de los vectores A y B se calcula multiplicando los componentes y sumando los productos, así (A ⭘ B) es igual a:

(A ⭘ B) = 1*1 + 1*0 + 0*0 + 0*0 + 1*1 + 1*0 + 1*1 = 1 + 0 + 0+ 0 +1 + 0 + 1 = 3

La suma de los cuadrados de los vectores es igual a:

$\Sigma(Ai)^2$ = 1*1 + 1*1 + 0*0 + 0*0 + 1*1 + 1*1 + 1*1 = 1 + 1 + 0 + 0 + 1 + 1 + 1 = 5

$\Sigma(Bi)^2$ = 1*1 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 1*1 = 1 + 0 + 0 + 0 + 1 + 0 + 1 = 3

Sólo resta dividir el producto escalar de los dos vectores (3) entre la raíz cuadrada de 5 multiplicada de la raíz cuadrada de 3 (o raíz de quince, ya que el producto de raíces cuadradas es la raíz cuadrada del producto). La raíz cuadrada la vamos a denotar como sqrt. Así tendremos como resultado:

Sim(A,B) = Cos(A,B) = 3 / sqrt (15) = 3 / 3.87 = **0.77**

2. En un SRI basado en el modelo vectorial, determina la similitud media de todos los vectores almacenados.

Vector	T1	T2	T3	T4	T5	T6
A	1	1	0	0	1	1
B	1	0	0	0	1	0
C	0	1	1	1	1	1
D	0	0	0	0	0	1

Solución.

La similitud media existente entre todos los vectores se obtiene calculando la media aritmética de las similitudes de cada par de vectores. Se usará la función del coseno para calculas los valores de similitud.

(A,B)

(A ⊙ B) = 1*1 + 1*0 + 0*0 + 0*0 + 1*1 + 1*0 = 1 + 0 + 0+ 0 +1 + 0 = 2

$\Sigma(Ai)^2$ = 1*1 + 1*1 + 0*0 + 0*0 + 1*1 + 1*1 = 1 + 1 + 0 + 0 + 1 + 1 = 4

$\Sigma(Bi)^2$ = 1*1 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 = 1 + 0 + 0 + 0 + 1 + 0 = 3

Sim(A,B) = 2 / sqrt (12) = 2 / 3.46 = 0.57

(A,C)

(A ⊙ C) = 1*0 + 1*1 + 0*1 + 0*1 + 1*1 + 1*1 = 0 + 1 + 1 + 0 + 0 + 1 + 1 = 4

$\Sigma(Ai)^2$ = 1*1 + 1*1 + 0*0 + 0*0 + 1*1 + 1*1 = 1 + 1 + 0 + 0 + 1 + 1 = 4

$\Sigma(Ci)^2$ = 0*0 + 1*1 + 1*1 + 1*1 + 1*1 +1*1 = 0 +1 + 1 + 1 + 1 + 1 = 5

Sim(A,C) = 4 / sqrt(20) = 4 / 4.47 = 0.89

(A,D)

(A ⊙ D) = 1*0 + 1*0 + 0*0 + 0*0+ 1*0+ 1*1 = 0 + 0 + 0 + 0 + 0 + 1 = 1

$\Sigma(Ai)^2$ = 1*1 + 1*1 + 0*0 + 0*0 + 1*1 + 1*1 = 1 + 1 + 0 + 0 + 1 + 1 = 4

$\Sigma(Di)^2$ = 0*0 + 0*0 + 0*0 + 0*0 + 0*0 +1*1 = 0 +0 + 0+ 0 + 0 + 1 = 1

Sim(A,D) = 1 / sqrt(4) = 1 / 2 = 0.50

(B,C)

(B ⊙ C) = 1*0 + 0*1 + 0*1 + 0*1+ 1*1+ 0*1 = 0 + 0 + 0 + 0 + 1 + 0 = 1

$\Sigma(Bi)^2$ = 1*1 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 = 1 + 0 + 0 + 0 + 1 + 0 = 3

$\Sigma(Ci)^2$ = 0*0 + 1*1 + 1*1 + 1*1 + 1*1 +1*1 = 0 +1 + 1 + 1 + 1 + 1 = 5

Sim(B,C) = 1 / sqrt(15) = 1 / 3.87 = 0.26

(B,D)

(B ⊙ D) = 1*0 + 0*0 + 0*0 + 0*0+ 1*0+ 0*1 = 0 + 0 + 0 + 0 + 1 + 0 = 0

No hace falta proseguir ya que el numerador de la expresión es cero, lo que implica que la similitud entre los vectores B y D es nula.

Sim(B,D) = 0

(C,D)

(C ⊙ D) = 0*0 + 1*0 + 1*0 + 1*0+ 1*0+ 1*1 = 0 + 0 + 0 + 0 + 0 + 1 = 1

$\Sigma(Ci)^2$ = 0*0 + 1*1 + 1*1 + 1*1 + 1*1 +1*1 = 0 +1 + 1 + 1 + 1 + 1 = 5

$\Sigma(Di)^2$ = 0*0 + 0*0 + 0*0 + 0*0 + 0*0 +1*1 = 0 +0 + 0+ 0 + 0 + 1 = 1

Sim(C,D) = 1 / sqrt(5) = 1 / 2.23 = 0.44

Si se suman las seis similitudes obtenidas y se calcula la media aritmética, se establecerá la similitud media del SRI.

SIM_MEDIA = (0.57 + 0.89 + 0.50 + 0.26 + 0 + 0.44) / 6 = 2.66 / 6 = **0.44**

3. En un SRI basado en el modelo vectorial cuya tabla de vectores de documentos podemos ver a continuación, calcula el alineamiento de la respuesta proporcionada por el sistema para responder a un vector pregunta **Q = (2,0,0,0,1,1)**.

Vector	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
A	1	1	0	0	1	1
B	1	0	0	0	1	0
C	0	1	1	1	1	1
D	0	0	0	0	0	1

Solución.

Entendiendo que en la matriz origen que nos proporcionan, los términos son todos no vacíos y se encuentran ya agrupados por una entrada común, el primer paso a seguir para resolver este problema es calcular las frecuencias inversas del conjunto de términos (T1, T2, ..., T6).

Frecuencias Inversas

Idf(T1) = log (4/2) = log 2 = 0.301

Idf(T2) = log (4/2) = log 2 = 0.301

Idf(T3) = log (4/1) = log 4 = 0.602

Idf(T4) = log (4/1) = log 4 = 0.602

Idf(T5) = log (4/3) = log 1.33 = 0.124

Idf(T6) = log (4/3) = log 1.33 = 0.124

Matriz de pesos

Vector	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
A	0.301	0.301	0	0	0.124	0.124
B	0.301	0	0	0	0.124	0
C	0	0.301	0.602	0.602	0.124	0.124
D	0	0	0	0	0	0.124
Q	2*0.301=0.602	0	0	0	0.124	0.124

Productos vectoriales.

(A • Q) = 0.301*0.602 + 0.301*0 + 0*0 + 0*0 + 0.124*0.124 + 0.124*0.124 = 0.181 + 0 + 0 + 0 + 0.015 + 0.015 = 0.211

(B • Q) = 0.301*0.602 + 0*0 + 0*0 + 0*0 + 0.124*0.124 + 0*0.124 = 0.181 + 0 + 0 + 0 + 0.015 + 0 = 0.196

(C • Q) = 0.301*0.602 + 0.301*0 + 0.602*0 + 0.602*0 + 0.124*0.124 + 0.124*0.124 = 0.181 + 0 + 0 + 0 + 0.015 + 0.015 = 0.211

(D • Q) = 0*0.602 + 0*0 + 0*0 + 0*0 + 0*0.124 + 0.124*0.124 = 0 + 0 + 0 + 0 + 0 + 0.015 = 0.015

Resultado

Ordenando los documentos de mayor a menor similitud, la respuesta del motor será: (A, C, B, D)

4. En un SRI basado en el modelo del espacio vectorial compuesto por los siguientes documentos:

D1: "el Río Segura pasa por Murcia y desemboca en el Mediterráneo"

D2: "Murcia es una región mediterránea seca con gran producción agrícola"

D3: "el Río Mundo es afluente del Río Segura"

D4: "los ríos Turia, Jucar y Segura, son ríos que desembocan el Mediterráneo"

D5: "el río Segura riega las huertas de Alicante y Murcia"

Determina el alineamiento del motor de búsqueda a la siguiente pregunta: "¿por qué región pasa el Río Segura?"

Solución

Se establece la matriz de frecuencias absolutas y se incluye la pregunta.

	Río	Segura	pasa	Murcia	Desem-boca	Medite-rráneo	región	seca	Produ-cción	agrícola	Mundo	afluente	Turia	Jucar	riega	huertas	Alicante
D1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
D2	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0
D3	2	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
D4	2	1	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0
D5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Q	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0

Se calculan las frecuencias inversas de todos los términos

Idf(río) = Idf(Segura) = log (5/4) = log 1.25 = 0.096

Idf(pasa) = log (5/1) = log 5 = 0.698

Idf(Murcia) = Idf(desemboca) = Idf(Mediterráneo) = log (5/2) = log 2.5 = 0.397

Como el resto de términos de la matriz sólo aparecen una vez, su Idf será igual a Idf(pasa), es decir, el resto de las frecuencias inversas es igual a 0.698. Con estos valores se elabora la matriz de pesos y se calculan los productos escalares.

	Río	Segura	pasa	Murcia	Desem-boca	Medite-rráneo	región	seca	Produ-cción	agrícola	Mundo	afluente	Turia	Jucar	riega	huertas	Alicante
D1	0.09	0.09	0.69	0.39	0.39	0.39	0	0	0	0	0	0	0	0	0	0	0
D2	0	0	0	0.39	0	0	0.69	0.69	0.69	0.69	0	0	0	0	0	0	0
D3	0.19	0.09	0	0	0	0	0	0	0	0	0.69	0.69	0	0	0	0	0
D4	0.19	0.09	0	0	0.39	0.39	0	0	0	0	0	0	0.69	0.69	0	0	0
D5	0.09	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0.69	0.69	0.69
Q	0.09	0.09	0.69	0	0	0.39	0.69	0	0	0	0	0	0	0	0	0	0

(D1 • Q) = 0.09*0.09 + 0.09*0.09 + 0.69*0.69 + 0 + 0 + 0.39*0.39 + 0 + ... + 0 = 0.008 + 0.008 + 0.476 + 0 + 0 + 0.152 + 0 ... + 0 = 0.796

(D2 • Q) = 0 + 0 + 0 + 0 + 0 + 0 + 0.69*0.69 + 0 + 0 + ... + 0 = 0.476

(D3 • Q) = 0.19*0.09 + 0.19*0.09 + 0 + + 0 + ... + 0 = 0.034

(D4 • Q) = 0.19*0.09 + 0.09*0.09 + 0 + 0 + 0.39*0.39 + 0 + 0 + ... + 0 = 0.034 + 0.008 + 0 + 0 + 0.152 + 0 + ... + 0 = 0.194

(D5 • Q) = 0.09*0.09 + 0.09*0.09 + 0 + 0 + ... + 0 = 0.008 + 0.008 = 0.016

En función de estos valores obtenidos, la respuesta será (D1, D2, D4, D3, D5)

5. Aplicando el método de cálculo aproximado de la Exhaustividad y Precisión de Salton, calcula los valores de la operación de búsqueda que se plasma en la siguiente tabla.

D1	D2	D3	D4	D5	D6	D7	D8
X	X	X		X	X		X

X ⇒ documento relevante

Solución:

Esta búsqueda ha obtenido 6 documentos relevantes de un total de 8 recuperados. Aplicando el método de Salton se considera que el total de documentos relevantes del sistema es 6 y esto permite calcular los pares de valores E-P de la siguiente forma:

Exhaustividad – Precisión			
N	Relevante	E	P
d1	X	1/6 = 0.16	1/1 = 1
d2	X	2/6 = 0.33	2/2 = 1
d3	X	3/6 = 0.50	3/3 = 1
d4		3/6 = 0.50	3/4 = 0.75
d5	X	4/6 = 0.66	4/5 = 0.80
d6	X	5/6 = 0.83	5/6 = 0.83
d7		5/6 = 0.83	5/7 = 0.71
d8	X	6/6 = 1	6/8 = 0.75
Sumatorios		4.81	6.84
Medias		0.60	0.85

Los valores medios E-P que se obtienen son (0.60 – 0.85)

6. Aplicando el método de cálculo aproximado de valores E-P de Salton, calcula cuál de las tres búsquedas siguientes es más precisa.

	D1	D2	D3	D4	D5	D6	D7	D8
B1	X	X	X		X	X		X
B2	X	X			X	X	X	X
B3		X	X	X	X	X	X	X

X ⇒ documento relevante

Solución:

La primera operación de búsqueda es la misma que se ha realizado en el ejercicio 5, por lo tanto ya conocemos su exhaustividad. Queda ahora por calcular la exhaustividad media de las búsquedas B2 y B3. Ambas recuperan 8 documentos, siendo 6 los relevantes en B2 y 7 los relevantes en B3.

Precisión						
		B1		***B2***		***B3***
N	Rel	P	Rel	P	Rel	P
d1	X	1/1 = 1	X	1/1 = 1		0/1 = 0
d2	X	2/2 = 1	X	2/2 = 1	X	1/2 = 0.50
d3	X	3/3 = 1		2/3= 0.66	X	2/3 = 0.66
d4		3/4 = 0.75		2/4= 0.50	X	3/4 = 0.75
d5	X	4/5 = 0.80	X	3/5 = 0.60	X	4/5 = 0.80
d6	X	5/6 = 0.83	X	4/6 = 0.66	X	5/6 = 0.83
d7		5/7 = 0.71	X	5/7 = 0.71	X	6/7 = 0.85
d8	X	6/8 = 0.75	X	6/8 = 0.75	X	7/8 = 0.87
Sumatorios		6.84		5.88		5.26
Medias		0.85		0.73		0.65

Los valores medios obtenidos indican claramente que la búsqueda más precisa es la búsqueda B1.

7. Aplicando el método de cálculo aproximado de valores E-P de Salton, calcula cuál de las tres búsquedas siguientes es más exhaustiva.

	D1	D2	D3	D4	D5	D6	D7	D8
B1	X	X			X	X		X
B2	X				X	X	X	X
B3			X	X	X	X	X	

X ⇒ documento relevante

Solución:

La búsquedas B1 y B2 recuperan 5 documentos relevantes de un total de 8 documentos recuperados. La tercera búsqueda recupera 5 documentos relevantes de un total de 7 documentos recuperados. Aplicando a continuación el método de Salton, obtendremos los valores medios de exhaustividad.

Exhaustividad						
		B1		***B2***		***B3***
N	Rel	E	Rel	E	Rel	E
d1	X	1/5 = 0.2	X	1/5 = 0.2		
d2	X	2/5 = 0.4		1/5 = 0.2		
d3		2/5 = 0.4		1/5 = 0.2	X	1/5=0.2
d4		2/5 = 0.4		1/5 = 0.2	X	2/5=0.4
d5	X	3/5 = 0.6	X	2/5 = 0.4	X	3/5=0.6
d6	X	4/5 = 0.8	X	3/5 = 0.6	X	4/5=0.8
d7		4/5 = 0.8	X	4/5 = 0.8	X	5/5=1
d8	X	5/5 = 1	X	5/5 = 1		
Sumatorios		4.4		3.6		3
Medias		0.55		0.45		0.42

8. Aplicando la fórmula de Borko, indica cuál de las siguientes tres búsquedas es mejor.

	D1	D2	D3	D4	D5
B1	X	X	X	X	
B2		X	X		X
B3		X	X	X	

X ⇒ documento relevante

Solución:

Las búsquedas B1 y B2 recuperan cinco documentos, resultando relevantes 4 en la primera y 3 en la segunda. La búsqueda B3 recupera 4 documentos, siendo relevantes 3 de ellos. Aplicando el método reducido de Salton para calcular los valores medios E-P se obtiene lo siguiente:

Exhaustividad-Precisión									
		B1			***B2***			***B3***	
N	Rel	E	P	Rel	E	P	Rel	E	
d1	X	1/4 = 0.25	1/1 = 1		0	0		0	0
d2	X	2/4 = 0.5	2/2 = 1	X	1/3 = 0.33	1 / 2= 0.5	X	1/3 = 0.33	1 / 2= 0.5
d3	X	3/4 = 0.75	3/3 = 1	X	2/3 = 0.66	2/3 = 0.66	X	2/3 = 0.66	2/3 = 0.66
d4	X	4/4 = 1	4/4 = 1		2/3 = 0.66	2/4 = 0.5	X	3/3 = 1	¾ = 0.75
d5		4/4 = 1	4/5 = 0.8	X	3/3 = 1	3/5 = 0.6			
Sumatorios		3.5	4.8		2.65	2.26		1.99	1.91
Medias		0.7	0.96		0.53	0.45		0.49	0.47

Aplicando la fórmula de Borko (suma de los valores medios E-P obtenidos en cada búsqueda), se obtienen los siguientes resultados.

Io (B1) = 0.7 + 0.96 = 1.66

Io (B2) = 0.53 + 0.45 = 0.98

Io (B3) = 0.49 + 0.47 = 0.96

Con estos valores, resulta evidente que la búsqueda B1 es la mejor de las tres, quedando en segundo lugar la búsqueda B2 y en tercer lugar la búsqueda B3, aunque con unos niveles de efectividad muy similares.

9. Aplicando la fórmula de Voiskunskii, indica cuál de las siguientes tres búsquedas es mejor.

	D1	D2	D3	D4	D5
B1	X	X	X		
B2		X	X	X	X

X ⇒ documento relevante

Solución:

Las búsquedas B1 y B2 recuperan cinco documentos, resultando relevantes 3 en la primera (los tres primeros) y 4 en la segunda (todos menos el primero). Aplicando el método de Salton vamos a calcular los valores medios E-P de cada una de estas búsquedas.

		B1			***B2***	
N	Rel	E	P	Rel	E	P
d1	X	1/3 = 0.33	1/1 = 1		0	0
d2	X	2/3 = 0.66	2/2 = 1	X	1/4 = 0.25	1 / 2= 0.5
d3	X	3/3 = 1	3/3 = 1	X	2/4 = 0.5	2/3 = 0.66
d4		3/3 = 1	3/4 = 0.75	X	3/4 = 0.75	3/4 = 0.75
d5		3/3 = 1	3/5 = 0.6	X	4/4 = 1	4/5 = 0.8
Sumatorios		3.99	4.35		2.5	2.71
Medias		0.79	0.87		0.5	0.54

Aplicando la fórmula de Voiskunskii (raíz cuadrada del producto de los valores medios E-P obtenidos) a las búsquedas B1 y B2 se obtienen los siguientes valores.

I1 (B1) = sqrt (0.79*0.87) = 0.829

I1 (B2) = sqrt (0.5*0.54) = 0.519

Con estos valores, la fórmula de Voiskunskii indica que la mejor búsqueda es la operación B1.

5 Referencias bibliográficas y fuentes de información.

[AGU, 2002] Aguilar González, R. *Monografía sobre motores de búsqueda* [En línea]. Yahoo Geocities, 2002. <http://www.geocities.com/motoresdebusqueda/inicio.html> [Consulta: 3 de abril de 2002]

[ALA, 1983] *Glosario A.L.A. de Bibliotecología y Ciencias de la Información*. Madrid: Díaz de Santos, 1983.

[ARA, 2000] Arauzo Galindo, M. et al. *Nuevas técnicas de búsqueda en Internet* [En línea]. Madrid: Universidad Complutense, 2000. <http://www.fdi.ucm.es/asignaturas/ssii_sup/SI/grupo1/resumenes/Nuevas%20tecnicas%20de%20busqueda.htm> [Consulta: 2 de marzo de 2002]

[BAE, 1992] Baeza-Yates, R. and Frakes, W.B. *Information retrieval : data structures & algorithms* Englewood Cliffs, New Yersey : Prentice Hall, 1992 504 p. ISBN 0-13-463837-9

[BAE, 1999] Baeza-Yates, R. and Ribeiro-Neto, B. *Modern information retrieval*. New York : ACM Press ; Harlow [etc.] : Addison-Wesley, 1999 XX, 513 p. ISBN 0-201-39829-X

[BER, 1992] Berners-Lee, T. et al World-*Wide Web: The Information Universe* [En línea]. World Wide Web Consortium: MIT, Massachusetts, 1992. <http://www.w3.org/History/1992/ENRAP/Article_9202.pdf> [Consulta: 18 de enero de 2002]

[BLA, 1990] Blair, D.C. *Language and representation in information retrieval*. Amsterdam [etc.]: Elsevier Science Publishers, 1990.

[BOR, 2000] Bors, N. 'Information retrieval, experimental models and statistical analysis'. *Journal of Documentation*, vol 56, nº 1 January 2000. p. 71-90

[BOW, 1994] Bowman, C. M. *The Harvest Information Discovery and Access System* [En línea]. Illinois: NCSA Information Techn. Division, 1994 <http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/schwartz.harvest/schwartz.harvest.html> [Consulta: 4 de febrero de 2002]

[BRA, 2000] *Multi-search Engines - a comparison* [En línea]. <http://www.philb.com/msengine.htm> [Consulta: 2 de abril de 2002]

[BRI, 1998 Brin, S. and Page, L. 'The anatomy of a large-scale hypertextual web search engine'. *Computer Netsorks and ISDN Systems*, 30, 1998. p. 107-117. También: [En línea] Stanford: University, 1999. <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm> [Consulta: 21 de noviembre de 2001]

[CAN, 1990] Canals Cabiró, I. "El concepto de hipertexto y el futuro de la documentación". *REDC*, 13(2), 685-709, 1990. p. 49-73.

[CHA, 2001] Chang, G. et al. *Mining the World Wide Web: an information search approach"*. Norwell, Massachusetts: Kluwer Acad. Publishers, 2001.

[CHO, 1999] Chowdhury, G. G. *Introduction to modern information retrieval*. London: Library Association, 1999.

[CON, 1987] Conklin, J. 'Hypertext: and introduction and survey'. *IEEE Computer* 20, 9, 1987. p.17-41.

[COO, 1973] Cooper, W.S. 'On selecting a Measure of Retrieval Effectiveness'. *Journal of the American Society for Information Science*, v. 24, March-April 1973. p.87-92

[CRO, 1987] Croft, W. B. 'Approaches to intelligent information retrieval'. *Information Proccesing & Management*, 23, 4,1987. p. 249-254

[DAV, 1996] Davis, E. *A Comparison of Seven Search Engines* [En línea]. Kent University, 1996. <http://www.iwaynet.net/~lsci/Search/paper.htm> [Consulta: 29 de diciembre de 2000]

[DEL, 1998] Delgado Domínguez, A. *Mecanismos de recuperación de Información en la WWW* [En línea]. Mallorca, Universitat Illes Balears, 1998. <http://dmi.uib.es/people/adelaida/tice/modul6/memfin.pdf> [Consulta: 18 de septiembre de 2001]

[DEL, 2001] Delgado Domínguez, A. *Herramientas de búsqueda para la WWW* [En línea]. Mallorca, Universitat Illes Balears, 1998. <http://dmi.uib.es/people/adelaida/CIVE/adecive.htm> [Consulta: 6 de enero de 2002]

[DOM, 2000] Dominich, S. 'A unified mathematical definition of classical information retrieval'. *Journal of the American Society for Information Science*, 51 (7), 2000. p. 614-624.

[FOS, 1972] Foskett, D.J. 'A Note on the Concept of Relevance'. Information Storage and Retrieval, 8 (2):77-78, April 1972

[FRA, 1997] Frants , V.I. et al. *Automated information retrieval : theory and methods*. San Diego [etc.] : Academic Press, cop.1997. XIV, 365 p.

[GIL, 1999] Gil Leiva, I. *La automatización de la indización de documentos*. Gijón: Trea, 1999. 221 p. ISBN 84-95178-11-7

[GOR, 1999] Gordon, M., Pathak, P. 'Finding information on the World Wide Web: the retrieval effectiveness of search engines'. *Information Processing and Management* 35, 1999. p. 141-180

[GRA, 2000] Grado-Caffaro, M. *Mecanismos/motores de búsqueda: ¿qué es lo que buscan?* Valencia: Quadernsdigital.net, 2000. <http://www.quadernsdigitals.net/articles/idg/mecanismos.htm> [Consulta: 11 de enero de 2002]

[GRE, 2000] Greisdorf, H. 'Relevance: An interdisciplinary and Informacion Science perspective'. Informing Science: Special Issue on Information Science Research. Vol 3 No 2, 2000. [También accesible en línea en <http://inform.nu/Articles/Vol3/v3n2p67-72.pdf> [Consulta: 16 de octubre 2001]

[GRO, 1998] Grossman, D.A. and Frieder, O. *Information retrieval: algorithms and heuristics*. Boston: Kluwer Academia Publishers, 1998.

[HU, 2001 Hu W-C et al. 'An overview of World Wide Web search technologies". *Proceedings of the 5th World Multi-Conference on Systemics, Cybernetics and Informatics, SCI 2001.* , Orlando, Florida, 2001. p. 356-361. También accesible [En línea] <http://www.eng.auburn.edu/users/wenchen/publication/overview.ps> [Consulta: 12 de marzo de 2002]

[IEI, 1997] *International Encyclopedia of Information & Library Science*. London: Rotledge, 1997

[INQ, 2002] *Welcome to Inquirus* [En línea]. Princeton: NEC Research Inst, 2002. < http://inquirus.nj.nec.com/> [Comsulta: 12 de marzo de 2002]

[KOR, 1997] Korfhage, R.R. *Information Retrieval and Storage*. New York: Wiley Computer Publisher, 1997.

[LAN, 1968] Lancaster, F.W. *Evaluation of the MEDLARS Demand Search Service*. Library of Medicine, Betsheda, Md, 1968.

[LAN, 1993] Lancaster, F. W. and Warner, A.J. *Information Retrieval Today*. Arlington, VIrginia : Information Resources, 1993.

[LON, 1989] Longley, D. and Shain M. *Mac Millan Dictionary of IT*. London and Basingstoke: The MacMillan Press, 1989.

[LOP, 1998] López Huertas, M.J. "La representación del usuario en la recuperación de la información". *Actas de las VI Jornadas de Documentación Automatizada*. Valencia: FESABID 98. p.

[MAN, 2002] Manchón, E. *Navegación jerárquica o categorial frente al uso de buscador* [En línea]. Ainda.info: Barcelona, 2002. <http://www.ainda.info/navegacion_vs_buscador.html> [Consulta: 21 de febrero de 2002]

[MEA, 1992] Meadow, C. T. *Text Information retrieval Systems*. San Diego: Academic Press, 1993.

[MIZ, 1998] Mizzaro, S. 'How many relevances in information retrieval?' [En línea]. Udine: Università degli Studi, 1998. <http://www.dimi.uniud.it/~mizzaro/papers/IwC/> [Consulta: 6 de septiembre de 2001]

[NIE, 1990] Nielsen, J. *Hypertext and hypermedia*. Oxford: Oxford Academia Press, 1990.

[NOT, 2000] Notess, G. R. *Search Engine Statistics: Dead reports*. [En línea]. Bozeman, MT: Notes.com, 2000. <http://www.searchengineshowdown.com/stats/dead.shtml> [Consulta: 12 marzo 2002]

[PER, 2000] Pérez-Carballo, J. and Strzalkowski, T. 'Natural language information retrieval: progress report'. *Information Processing and Management* 36, 2000. p. 155-178

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

[SAL, 1983] Salton , G. and Mc Gill, M.J. *Introduction to Modern Information Retrieval*. New York: Mc Graw-Hill Computer Series, 1983.

[TRA, 1997] Tramullas Sáez, J. *Introducción a la Documática* [En línea]. Zaragoza: Kronos, 1997.

<http://www.tramullas.com/nautica/documatica/3-1.html> [Consulta: 18 de noviembre de 2001]

[VIL, 1997] Villena Román, J. *Sistemas de Recuperación de Información* [En línea]. Valladolid: Departamento Ingeniería Sistemas Telemáticos, Universidad. <http://www.mat.upm.es/~jmg/doct00/RecupInfo.pdf> [Consulta: 20 de febrero de 2002]

[VOI, 1997] Voiskunskii, V. G. 'Evaluation of search results'. *Journal of the American Society for Information Science*. 48(2) 1997. p.133-142

[WAN, 2001] WANG, S. 'Toward a general model for web-based information systems'. *International Journal of Information Management* 21, 2001. p. 385–396

[WES, 2001] WESTERA, G. *Comparison of Search Engine User Interface Capabilities* [En línea]. Curtin: University of Technology, july, 2001. <http://lisweb.curtin.edu.au/staff/gwpersonal/compare.html> [Consulta: 9th-july-2001]

[WIN, 1995] Winsip, I. 'World Wide Web searching tools - an evaluation'. *Vine* 99, 1995. p. 49-54 También disponible en: [En línea] <http://gti1.edu.um.es:8080/javima/World-Wide-Web-searching-tools-an-evaluation.htm> [Consulta: 5 de noviembre de 2000]

[ZOR 1996] Zorn, P., Emanoil, M., Marshall, L. and Panek, M. *Advanced Searching: Tricks of the Trade*. [En línea]. Wilton, CT: Online Inc, 1996. <http://www.onlinemag.net/MayOL/zorn5.html > [Consulta: 15 de julio de 2001]