ANALES DE DOCUMENTACION, Nº 7, 2004, PÁGS. 153-170

REFLEXIONES SOBRE LA EVALUACIÓN DE LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN: NECESIDAD, UTILIDAD Y VIABILIDAD.

Francisco Javier Martínez Méndez*

Departamento de Información y Documentación. Facultad de Comunicación y Documentación. Universidad de Murcia

José Vicente Rodríguez Muñoz**

Departamento de Información y Documentación. Facultad de Comunicación y Documentación. Universidad de Murcia

Resumen: La evaluación de los Sistemas de recuperación de información se ha desarrollado paralelamente a su evolución por razones de carácter científico. Así, han aparecido un gran número de propuestas para llevar a cabo este proceso. En cualquier caso y dando por sentado que cada una de las propuestas realizadas -la mayoría de las cuales se presentan en este trabajo- tratan de buscar un mecanismo robusto para medir el comportamiento de la recuperación de información, no es menos cierto que nos enfrentamos a un problema complejo y que en definitiva tienen como objetivo encontrar un método que permitiese seleccionar el mejor modo de recuperación dada una necesidad de información de un usuario. Este el es el eje donde gravita el escenario de la evaluación de la recuperación de información.
Palabras Clave: Evaluación; medidas de evaluación; recuperación de información.

Title: THOUGHTS ABOUT THE EVALUATION OF INFORMATION RETRIEVAL SYSTEMS: NECESSITY, UTILITY AND VIABILITY.

Abstract: The evaluation of Information Retrieval Systems has been developed parallelly to their evolution for reasons of scientific nature, generating the appearance of a great number of proposals to make this process. Assuming that all these proposals - most of which they appear in this work- try to look for a robust mechanism to measure the effectiveness of Information Retrieval Systems, it is not less certain information retrieval is a complex problem to resolve, due to the inherent problems related with translating the user's information needs to a formalized expression search. The implementation of an objective and simple method to evaluate the effectiveness is the axis where the scene of the evaluation of the information retrieval gravitates.
Keywords: Evaluation; Effectiveness measures; Information retrieval Systems.

1. NECESIDAD DE LA EVALUACIÓN DE LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

Los Sistemas de Recuperación de Información -SRI en adelante-, resultan susceptibles, como cualquier otro sistema, de ser sometidos a evaluación, para que sus usuarios puedan valorar su efectividad. La tradición de la evaluación es tan antigua casi como el

* javima@um.es
** jovi@um.es

desarrollo de los primeros SRI, encontrándose estrechamente vinculadas con la investigación y el desarrollo de la recuperación de información. Realmente, "la propia naturaleza de los SRI propicia su necesidad crítica de evaluación, justo como cualquier otro campo de trabajo que aspire a ser clasificado como campo científico"[1].

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". Estos criterios no deben confundirse, la *eficacia en la ejecución* es la medida del tiempo para realizar una operación, la *eficiencia del almacenamiento* es el espacio que se precisa para almacenar los datos y por último está la efectividad de la recuperación "normalmente basada en la *relevancia* de los documentos recuperados"[2] .

Borlund diferencia entre evaluar el *acceso físico* y el *acceso lógico* a los datos, considerando que las evaluaciones han de ser del segundo tipo. El acceso físico es el que concierne a cómo la información es recuperada y representada de forma física al usuario, está 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 Blair, "descubrir dónde se encuentra un libro en una biblioteca con una signatura determinada es un problema relacionado con el acceso físico al objeto -el libro-; descubrir qué libro puede informarnos sobre una determinada materia es un problema relacionado con el acceso lógico"[3]. Este segundo caso tiene que ver con la *relevancia* del objeto localizado con una determinada petición de información. Borlund distingue entre "aproximaciones al funcionamiento del sistema y aproximaciones centradas en el usuario"[4], plenamente coincidentes con acceso físico y acceso lógico.

Estas consideraciones realizadas a principio de la década de los años noventa, siguen plenamente vigentes más de una década después. La actual tendencia de mejorar el acceso físico certifica los temores de Blair quien discrepa profundamente sobre lo que debería ser evaluado con el fin de determinar con certeza que la información que un SRI proporciona es válida para sus usuarios, por medio del análisis de la *relevancia* o *no relevancia* del documento recuperado. De forma parecida, Baeza-Yates afirma que existen dos tipos de evaluaciones: la del funcionamiento del sistema y la del funcionamiento de la recuperación, siendo la segunda modalidad la que analiza cómo los documentos recuperados se clasifican de acuerdo a su *relevancia* con la pregunta efectuada[5].

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

2 Baeza-Yates, R. and Frakes, W.B. Information retrieval: data structures & algorithms Englewood Cliffs, New Yersey: Prentice Hall, 1992 504 p.

3 Blair, D.C. 1990.

4 Borlund, P. 'Information retrieval, experimental models and statistical analysis'. Journal of Documentation, vol 56, nº 1 January 2000. p. 71-90.

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

2. RELEVANCIA VERSUS PERTINENCIA

Conviene delimitar el significado de *relevancia* dentro del contexto de la recuperación de información. Según el *Diccionario de la Lengua Española*[6], *relevancia* significa "cualidad o condición de relevante, importancia, significación", y *relevante* es definida como "importante o significativo". Así, un documento será relevante cuando el contenido del mismo posea alguna significación o importancia en relación con la pregunta realizada por el usuario, es decir, con su necesidad de información. Aún así, subyacen algunos problemas estrechamente entroncados con la naturaleza cognitiva de este proceso, a la hora de determinar con exactitud cuándo un documento puede ser considerado relevante o no:

- Un mismo documento puede ser considerado relevante, o no relevante, por dos personas distintas en función de su necesidad de información o su grado de conocimiento de la materia. Llegados a un caso extremo, un mismo documento puede parecer relevante o no a la misma persona en momentos diferentes de tiempo[7]
- Resulta difícil definir establecer, a priori, unos criterios para determinar cuándo un documento es relevante e incluso resulta complicado explicitarlos de forma clara y concisa, siendo más fácil proceder a la determinación de la *relevancia* que explicar cómo la misma se lleva a cabo.
- Es muy aventurado calificar categóricamente un documento como relevante -o no relevante- con un tema, en la realidad lo normal es encontrarnos con documentos relevantes con una materia determinada en alguno de sus apartados, pero no en el resto de sus contenidos. Para subsanar esta circunstancia, algunos autores introducen el concepto de *relevancia parcial*.

Estas objeciones condicionan, en cierto grado, la viabilidad de la *relevancia* para constituirse en un criterio de evaluación de la recuperación de la información. Cooper aporta la idea de "utilidad de un documento" o *pertinencia*, considerando que es mejor definir a la *relevancia* en términos de la percepción que un usuario posee sobre la utilidad de un documento recuperado, es decir, *si el mismo le va a ser útil o no*. Este nuevo punto de vista supera alguna de las limitaciones anteriores, ya 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. Frants plantea otra acepción de *relevancia* muy similar a la anterior, en términos de *eficiencia funcional*. Así, *relevancia* queda asociada con el concepto de la relación existente entre los contenidos de un documento con una temática determinada y *pertinencia* se restringe a la *relación de utilidad* existente entre un documento recuperado y una necesidad de información individual. Por otro lado, si bien es considerable el número de problemas que presenta la *relevancia*, "en los últimos treinta años no se ha encontrado sustituto práctico para el concepto de *relevancia* como criterio de medida de la efectividad de los SRI"[8].

---

[6] DRAE *Diccionario de la Lengua Española* [En línea]. Madrid: Real Academia Española, 2004. <http://buscon.rae.es/diccionario/drae.htm> [Consulta: 29 febrero 2004].

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

[8] 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: 1 marzo 2004].

3. PRIMERAS EVALUACIONES DE LOS SRI

Casi la totalidad de la bibliografía consultada hace referencia a las evaluaciones llevadas a cabo a principios de los años cincuenta, conocidas como los Proyectos Cranfield -toman el nombre del Instituto Científico donde se llevaron a cabo-, que marcaron el rumbo de los posteriores trabajos, aportando medidas aún vigentes.

Proyectos CRANFIELD

Estos proyectos "proporcionaron una nueva dimensión a la investigación en SRI"[9] y representan 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"[10].

Son dos los estudios *Cranfield* más importantes. El primero fue dirigido por Cleverdon y comenzó en 1957, tenía como objetivos comparar la efectividad de cuatro sistemas de indización: un catálogo alfabético de materias; una clasificación CDU; un catálogo basado en una clasificación por facetas y un catálogo compilado por un índice coordinado de unitérminos. Los resultados proporcionan unos valores de *exhaustividad* altos -entre el 60 y el 90% con un promedio del 80%-, favorecidos por el tiempo dedicado a la indización, y aportaban 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"[11]. Este experimento estableció los factores que más afectan al funcionamiento de los SRI y refrendó el desarrollo de la primera metodología de evaluación, introduciendo las medidas de *exhaustividad* y *precisión*, muy utilizadas aún.

El segundo proyecto *Cranfield* consistió en un experimento destinado a fijar los efectos de los componentes de los lenguajes de indización en la ejecución de los sistemas y ofrecer información sobre la naturaleza de los fallos de un SRI. Los resultados fueron 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 a priori- Vickery comenta "que las medidas usadas en el segundo experimento *Cranfield* no caracterizaron adecuadamente los aspectos operativos de un SRI"[12], de ahí que este segundo proyecto no haya tenido tanta repercusión como el primero.

MEDLARS. Medical Literature Analysis and Retrieval System

Este sistema -sistema de recuperación de información de la Biblioteca Nacional de Medicina- fue evaluado por Lancaster observando la efectividad de la recuperación de

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

[10] 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.

[11] Chowdhury, G. G. 1999.

[12] Chowdhury, G.G. 1999.

información. Los resultados proporcionaron valores medios de *exhaustividad* más bajos que los obtenidos en el primer test de *Cranfield* -en torno al 57%-, y valores medios de *precisión* del 50%. A diferencia del anterior, este test sí proporcionaba pormenores sobre las razones de los fallos en la recuperación de información, centrándose la mayor parte 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.

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. Se distingue del resto de los SRI convencionales en cuatro aspectos fundamentales: (1) usa métodos de indización automática; (2) agrupa documentos relacionados dentro de clases comunes de materias; (3) identifica los documentos a recuperar por similitud con la pregunta realizada por el usuario y (4) incluye procedimientos automáticos para generar mejores ecuaciones de búsqueda[13].

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, al cual se le asigna el peso. Con el tesauro se asignan los términos descriptores que mejor representan a los conceptos de los documentos y se les asigna un peso. Se obtuvieron dos series de resultados principales: (1) el análisis con tesauro mejora ligeramente al de los lemas y ambos resultan bastante mejores que el de términos simples o palabras; (2) los mejores resultados de *exhaustividad* y *precisión* se obtienen en el cuarto grupo de evaluaciones, es decir, cuando tanto los usuarios que realizan las preguntas como los evaluadores ajenos a esas preguntas están de acuerdo. 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.

STAIRS. Storage And Information Retrieval System

La evaluación de STAIRS fue un proyecto desarrollado en los años ochenta por Blair y Maron[14]. Analizaron 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 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 estudios anteriores, 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 experimentado-. Los resultados mostraron que la media de las *exhaustividad*es obtenida por los abogados superaba a la media de los pasantes. En

[13] Salton, G. and Mc Gill, M.J. *Introduction to Modern Information Retrieval*. New York: Mc Graw-Hill Computer Series, 1983.
[14] Blair, D.C. 1990.

cambio, las diferencias en *precisión* son mayores. 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 se antoja poco reflexionada y algo simple, ya que las diferencias, tanto de *precisión* como de *exhaustividad* no son muy grandes.

Conferencias TREC. Text REtrieval Conferences

Las conferencias TREC, se han convertido en el foro de intercambio científico más prestigioso del campo de la recuperación de información y en consecuencia en el eje central donde gravita la evolución de los SRI. TREC reúne a creadores de diferentes sistemas y compara los resultados que éstos obtienen en diferentes pruebas, previamente estandarizadas y acordadas por todos. La primera conferencia, TREC-1 (1992), ofreció como resultado principal la existencia de una amplia similitud entre los SRI que hacen uso de técnicas basadas en lenguaje natural y los basados en el modelo 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.

4. MEDIDAS DE EVALUACIÓN TRADICIONALES

Rijsbergen se pregunta *¿qué evaluar?*, y responde a esa pregunta citando a Cleverdon, quien en 1966 presentaba seis medidas principales: "la cobertura de una colección; el tiempo de respuesta del sistema a una petición; la forma de presentación de los resultados; el esfuerzo realizado por el usuario; la *exhaustividad* del sistema y la *precisión* del sistema". Este autor opina que las cuatro primeras medidas son intuitivas y fácilmente calculables, y que las dos últimas 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"[15].

Vickery propone también 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 del sistema; 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"[16].

Además del criterio de la *relevancia*, algunos autores han empleado medidas basadas en criterios diferentes. Meadow las sintetiza en dos grupos, tal y como podemos ver en la tabla 1:

[15] Rijsbergen, C.J. *Information Retrieval*. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 29 febrero 2004].

[16] Chowdhury, G.G. 1999.

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
Núm. docs recuperados	Extensión del resultado de una búsqueda
Núm. docs revisados por el usuario	Promedio de documentos que los usuarios están dispuestos a revisar

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 1. Resumen medidas empleadas en la evaluación convencional de la recuperación de la información. Fuente: Meadow, C.T. 1992

Por regla general, se consideran de mayor importancia las medidas basadas en la *relevancia* que aquellas basadas en el *proceso*, principalmente porque estas últimas dependen mucho de las prestaciones de la aplicación informática subyacente, y no valoran adecuadamente aspectos relacionados con el contenido de los documentos. Las medidas del *resultado* son muy similares a las basadas en la *relevancia*, aunque introducen algunos aspectos diferenciadores.

Medidas basadas en la relevancia

En una operación de recuperación de información, un usuario recupera un conjunto de documentos, algunos van a ser documentos relevantes con la temática objeto de su necesidad de información, y otros van a formar parte del subconjunto de documentos que no lo van a ser. Asimismo, este usuario dejará de recuperar otros documentos que igualmente serán relevantes con esa temática, y otro conjunto de documentos que no lo van a ser. Esta distribución de resultados de una búsqueda permite la especificación de una serie de subconjuntos de la base de datos en relación con la pregunta realizada, que muestra Rijsbergen en la tabla 2, 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 2. Tabla de contingencia de Rijsbergen. Fuente: Rijsbergen, C.J. 1999.

La *precisión* mide el porcentaje de documentos recuperados que resultan relevantes con el tema de la pregunta y su cálculo es 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 su definición es clara -número de documentos relevantes recuperados dividido entre el número de documentos totales relevantes de la base de datos-, no está claro cuál es el valor de ese denominador -si el usuario conociera de antemano el número de documentos relevantes de la base de datos, ¿por qué no los recupera todos en esa búsqueda? La respuesta es simple: porque no los puede conocer de antemano, como máximo puede inferir ese valor-. Estas dos medidas tienden a relacionarse de forma inversa, ya que cuanto mayor es el valor de la *precisión*, menor va a ser el valor de la *exhaustividad*. Si un usuario lleva a cabo una operación de recuperación de información en la cual inserta condiciones muy específicas, 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.

La *tasa de fallo* refleja el porcentaje de documentos recuperados no relevantes sobre el total de documentos no relevantes de la base de datos. Otra medida relacionada es el *factor de generalidad*, "grado de documentos relevantes contenidos en una colección"[17]. 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 están estrechamente vinculadas, de manera que la *precisión* puede definirse en función de las tres restantes, tal como podemos observar en la siguiente expresión recogida:

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

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

Al considerable problema de la imposibilidad de determinar con exactitud el valor de la *exhaustividad*, Korfhage añade: "no está claro que la *exhaustividad* y la *precisión* sean medidas significativas para el usuario"[18]. De hecho, una amplia mayoría de usuarios consideran mucho más importante la *precisión*, relegando generalmente a la *exhaustividad* a un cometido secundario, mientras la búsqueda proporcione información relevante, el usuario no suele detenerse a pensar en la cantidad de documentos relevantes que no recupera,

[17] Salton, G. and Mc Gill, M.J. 1983.

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

Aunque este razonamiento no puede aplicarse como regla general en todos los SRI, por ejemplo en una base de datos jurídica es vital garantizar un alto nivel del *exhaustividad*.

Medidas orientadas al usuario

Muchos autores consideran 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 y que 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, situación no muy real. Para solucionar este problema, Salton y McGill, Baeza-Yates y Korfhage presentan una serie de 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 - situación algo difícil de asumir también en la realidad-. Son las m*edidas orientadas al usuario,* conjunto propuesto por Keen a principio de la década de los setenta:

- *Cobertura*: proporción de los documentos relevante conocidos que el usuario ha recuperado
- *Novedad*: proporción de los documentos recuperados relevantes que eran previamente desconocidos para el usuario
- *Exhaustividad Relativa*: ratio que se establece entre los documentos relevantes recuperados examinados por el usuario y el número de documentos que el usuario está dispuesto a examinar.

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. Existe 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"[19]. Esta medida presupone que "la colección contiene el número deseado de documentos relevantes y el SRI permite al usuario localizarlos todos"[20].

Medidas alternativas a E-P o medidas de valor simple

Rijsbergen presenta un amplio conjunto de medidas alternativas de la efectividad de una recuperación de información, que pretenden superar los problemas subyacentes en la *exhaustividad* y en la *precisión*. Casi todas hacen uso de técnicas probabilísticas y resultan de cálculo más complicado que las anteriores. La mayor parte de los autores presentan a estas medidas por separado, incluso no llegan a no ponerse de acuerdo en su denominación, unos las llaman *alternativas* y otros las denominan de *valor simple*; incluso algún autor, tan reconocido como Baeza-Yates, simplemente las cita en un apartado genérico denominado *otras medidas*.

Salton las denomina *de valor simple* presentan el resultado de una evaluación en función de un único valor, que puede ser objeto de clasificación. Salton presenta también sucesivamente, el *Modelo de Swet* que propicia el desarrollo de la *Medida E*, las *medidas*

[19] Baeza-Yates, R. and Ribeiro Neto, B.1999.
[20] Korfhage, R.R. 1997.

SMART -sistema de indización automática desarrollado por Salton- y la *longitud esperada de búsqueda* basada en el *Modelo de Cooper*. Es Rijsbergen el primer autor que muestra una nueva faceta de la evaluación de la recuperación de la información relacionada con las medidas de valor simple. El autor cita la *medida de satisfacción* de Borko -suma de los valores de *precisión* y de *exhaustividad* obtenidos en una búsqueda-. Otras medidas de esta naturaleza se han definido por otros autores y se recogen en la tabla 3:

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. Medidas de la calidad de una búsqueda propuestas por varios autores. Fuente: Meadow, C. T. 1992.

Estas medidas constituyen un instrumento útil para valorar la calidad de una búsqueda, aunque para algunos no resultan suficientes en todos los casos. Lo que sí está claro para la mayoría de los autores, es la necesidad de desarrollar un conjunto de medidas alternativas a las tradicionales.

Medidas mucho más complejas son el *Modelo de Swet*, el *Modelo de Robertson*, el *Modelo de Cooper* y las *medidas SMART*. Korfhage separa las *medidas simples* de la *medida longitud esperada de búsqueda -o Medida E-*, pasando a comentar luego las *medidas de satisfacción* y *frustración*. Para Salton, las medidas a emplear deberían cumplir las siguientes condiciones:

- Deben ser capaces de reflejar la efectividad de la recuperación por medio de un valor único, de forma aislada de otros criterios como puede ser el coste.
- Deben ser independientes de cualquier límite, es decir, el número de documentos recuperados en una búsqueda específica no debe influenciar a estas medidas.
- 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 esta serie de premisas se desarrolla este modelo, suficientemente explicado por Rijsbergen y Salton. Define la terna de medidas E-P-F -*exhaustividad, precisión, tasa de fallo*-, en términos probabilísticos. Así, 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* será una estimación de la probabilidad condicionada de que un documento recuperado no sea relevante.

Este modelo convierte las representaciones E-P en otras, construidas en función de disponer de un determinado número de documentos relevantes o no. Las operaciones de búsqueda producen como resultado unas funciones lineales que guardan una cierta distancia con la distribución de las probabilidades anteriormente citadas y es precisamente el valor de esa distancia -multiplicado por la raíz cuadrada de dos- el valor de la *Medida E* de Swets. El principal inconveniente de esta medida reside en que "a diferencia de la *exhaustividad* y de la *precisión*, estas medidas -la distancia y la pendiente- no resultan fácilmente descifrables por los usuarios y difícilmente harán uso de ellas"[21].

Modelo de Robertson

Este modelo es una aproximación logística a la estimación de los valores de *exhaustividad* y *precisión*, *"Robertson*, en colaboración con Teather, desarrolla un modelo que estima las probabilidades correspondientes a E-P. Este procedimiento resulta inusual ya que al calcular una estimación de ambas probabilidades para una pregunta simple, considera dos cosas: la cantidad de datos empleados para alcanzar esas estimaciones y los promedios de las estimaciones de todas las demás preguntas"[22]. El objetivo de este método es obtener un valor *delta* candidato a convertirse en una medida simple de la efectividad de un SRI.

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 cuáles de aquellos son los relevantes"[23]. Para Cooper es precisamente este *ahorro de esfuerzo* lo que se debe medir y para ello haría falta una medida simple que sólo se aplicaría a los sistemas que mostraran la salida de los documentos ordenados según un determinado criterio de alineamiento.

Longitud esperada de búsqueda

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*"[24]. Esta medida no proporciona directamente un valor simple, más concretamente 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 un valor simple.

Exhaustividad y Precisión normalizadas

Otro problema de la *exhaustividad* y de la *precisión* reside en la *secuencialidad* de la lectura de los resultados de una búsqueda. Este modo de examinar afecta al juicio de la

[21] Salton, G. and Mc Gill, M.J. 1983.
[22] Rijsbergen, C.J. 1999.
[23] Rijsbergen, C.J. 1999.
[24] Salton, G. and Mc Gill, M.J. 1983.

relevancia de los documentos siguientes, casi todos los usuarios de los SRI han sufrido este problema cuando, al consultar dos documentos más o menos igual de interesantes y relacionados con una materia, centran su atención de forma preferente en el primero de ellos, aunque el segundo no desmerezca en nada al anterior. Otra situación parecida se produce cuando un usuario realiza una búsqueda y los primeros documentos recuperados resultan relevantes con el tema de su interés. 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 ha recuperado, que puede llegar a ser muy grande. La situación contraria también se produce frecuentemente. Esta reflexión propicia el desarrollo de medidas que tomen en consideración la secuencia en la que los documentos son presentados a los usuarios.

El primer trabajo conocido de esta última línea corresponde a Rocchio[25], quien define una *exhaustividad normalizada* y una *precisión normalizada* para sistemas que presentan los documentos alineados 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 recuperación de información". En la ilustración 1 podemos ver un ejemplo de cómo la *Exhaustividad* normalizada queda comprendida entre el peor y mejor resultado posible.

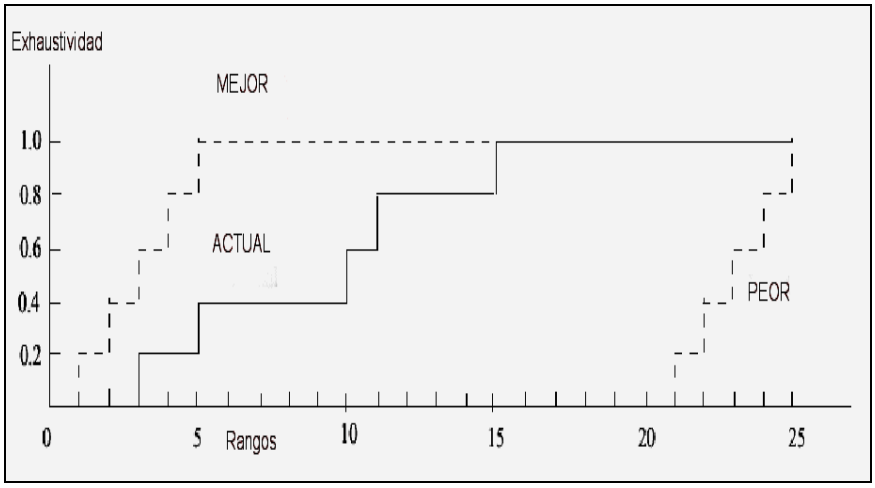

Ilustración 1. Fuente: RIJSBERGEN, C.J. 1999.

Para Korfhage, "el área comprendida entre la búsqueda actual y la gráfica ideal representa una medida de la ejecución del SRI"[26]. Rijsbergen destaca de estas medidas su 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 enten-

[25] Rijsbergen, C.J. 1999.
[26] Korfhage, R.R. 1997.

derse 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"[27].

Ratio de deslizamiento

Salton y Korfhage presentan la *ratio de deslizamiento*. Esta medida es muy similar conceptualmente a la *exhaustividad normalizada*, aunque presenta algunos matices: "se basa 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"[28].

Este modelo es algo más complejo que el anterior, porque permite la asignación de pesos a los documentos en función del grado de relevancia con la pregunta realizada por el usuario. 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 -documento relevante o no relevante-, por la asignación de un peso. 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.

Tanto esta medida, como las anteriores medidas normalizadas, pretenden cuantificar la diferencia existente entre la secuencia de documentos que entrega un sistema real y un sistema ideal, tomando en cuenta las posiciones en las que los documentos relevantes y los no relevantes aparecen en esa secuencia. En cambio, la ratio de deslizamiento presenta dos ventajas frente a la *exhaustividad normalizada*: "su uso de los pesos de *relevancia* y que sólo depende de los documentos recuperados"[29].

Satisfacción y Frustración

El esfuerzo que implica la determinación de las medidas anteriores, propicia que se establezcan otra nueva serie: *satisfacción*, *frustración* y *total*. La primera sólo considera los documentos relevantes, la segunda sólo contempla a los no relevantes y la tercera combina ponderadamente las medidas anteriores. Cuando se hace uso de los pesos, a los documentos relevantes se les suele asignar los valores más cercanos al umbral superior de la escala, y a los documentos 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 hay que considerar sus posiciones en la secuencia. De forma similar, 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* viene determinada 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.

[27] Rijsbergen, C.J. 1999.
[28] Salton, G. and Mc Gill, M.J. 1983.
[29] Korfhage, R.R. 1997.

Comparación medidas satisfacción, frustración y total para dos 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 4. Ejemplo propuesto por Korfhage para determinar la medida de la satisfacción. Fuente: Korfhage, R.R. 1997.

Korfhage propone el siguiente ejemplo para ilustrar el cálculo de esta medida: "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}"[30]. Si se entiende que la medida *total* va a ser el resultado de restarle el valor de frustración al valor de satisfacción, 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, 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 5. Diferencias de áreas para el ejemplo de determinación de la ratio de deslizamiento. Fuente: Korfhage, R.R. 1997.

[30] Korfhage, R.R. 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"[31].

Medida de Voiskunskii

El caso anterior, introduce una nueva dimensión de la evaluación de la recuperación de información, no centrada en la comparación de dos sistemas A y B, sino preocupada en intentar discernir "uno de los más importantes y complejos problemas de las Ciencias de la Información, la creación de un mecanismo de selección que permita elegir el mejor método de búsqueda -de entre las posibles variaciones- para una cuestión Q"[32]. Estos problemas residen básicamente en la ausencia de criterios sólidos y consistentes de comparación de los resultados de una búsqueda. El autor considera que estos criterios deben cumplir los siguientes requerimientos:

- Los criterios deben proveer una comparación pragmática y justificada de los resultados de la búsqueda; y
- la cantidad de trabajo precisa para determinar la información que es requerida para el establecimiento de estos criterios debe ser admisible.

La medida de valor simple más utilizada es la *satisfacción* propuesta por Borko, aunque la misma puede llevar -en algunas ocasiones muy determinadas- a una serie de conclusiones equivocadas. Frants, Shapiro y Voiskunskii[33] enuncian el siguiente contraejemplo: "en una colección de 10.000 documentos, de los cuales 100 de ellos se consideran pertinentes con una determinada materia, se llevan a cabo tres operaciones de búsqueda con los siguientes resultados.

- Se recuperan 100 documentos, 50 de ellos son pertinentes y el resto no lo es.
- Se recuperan 67 documentos, siendo pertinentes 40 de ellos.
- Por último, se recupera sólo un documento que resulta pertinente."

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

Búsqueda	E	P	I
A	0.500	0.500	1.000
B	0.400	0.597	0.997
C	0.010	1.000	1.010

Tabla 6. Ejemplo del cálculo de la medida I de Borko para el ejemplo propuesto por Frants, Shapiro y Voiskunskii. Fuente: Frants, V.I. 1997.

Una interpretación literal de estos valores indicaría que la mejor búsqueda es la *C*, al ser el valor más alto, aunque la búsqueda *C* difícilmente puede considerarse incluso admi-

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

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

[33] Frants, V.I. 1997.

sible, ya que sólo entrega al usuario un único documento pertinente de los cien que hay en la base de datos. Como solución, Frants, Shapiro y Voiskunskii proponen una nueva medida de valor simple cuya formulación analítica se corresponde con la raíz cuadrada del producto de los valores E-P. Esta medida aplicada al ejemplo anterior indicaría que la búsqueda *A* es la mejor de las tres -afirmación algo más coherente-. Posteriormente, Voiskunskii llega a desarrollar hasta nueve medidas más, que van ganando en complejidad en su cálculo y entendimiento, aunque él mismo indica que aplicando su medida en búsquedas de *precisión* mayor que 0.5, las conclusiones que se extraigan quedan suficientemente justificadas.

4. FORTALEZAS Y DEBILIDADES DE LAS MEDIDAS

A lo largo de la exposición del amplio conjunto de medidas que se han diseñado para medir la efectividad de la recuperación de información se han adelantado algunas de las ventajas e inconvenientes de algunas de ellas, que sintetizamos en la siguiente tabla de fortalezas y debilidades de cada una.

Medida	**Fortaleza // Debilidad**
Precisión	Medida intuitiva. Cálculo simple. Muy popular. // *Cierto grado de subjetividad.*
Exhaustividad	Medida intuitiva. Fácil expresión matemática. // *Cálculo aproximado. Cierto grado de subjetividad.*
Tasa de Fallo	Medida intuitiva. Fácil expresión matemática. // *Cálculo aproximado. Cierto grado de subjetividad. Poco popular.*
Generalidad	Medida intuitiva. Fácil expresión matemática. // *Cálculo aproximado. Cierto grado de subjetividad. Prácticamente desconocida.*
Cobertura	Medida intuitiva. Fácil expresión matemática. // *Muy vinculada al usuario. Alto grado de subjetividad.*
Formato de la presentación	*No está directamente relacionada con la efectividad de la recuperación de información. Expresión matemática inviable. Cierto grado de subjetividad.*
Usabilidad	Medida intuitiva *// Expresión matemática inviable. Cierto grado de subjetividad.*
Contenido	Datos técnicos necesarios para elaborar las ecuaciones de búsqueda // *No influye directamente en la efectividad de la recuperación.*
Tiempo	Medida intuitiva. Popular. // Indeterminación de su cálculo. *No influye directamente en la efectividad de la recuperación sino en la eficiencia.*
Dificultad en la operatoria	*Expresión matemática inviable. Cierto grado de subjetividad.*

Medida	Fortaleza // Debilidad
Coste de la búsqueda	*No influye directamente en la efectividad de la recuperación sino en su eficiencia.*
Nº documentos recuperados	Fácil cálculo // *No influye directamente en la efectividad de la recuperación.*

Tabla 7. Fortalezas y debilidades de las medidas propuestas para la efectividad de la recuperación de información. Fuente: Elaboración propia.

El hecho, claramente observable en la tabla anterior, de que las fortalezas de las medidas se asienten sobre su intuición y una sencilla expresión matemática (que le proporcionará cierta popularidad), y que entre las debilidades prevalezcan, de un lado, la subjetividad inherente a los juicios de relevancia y por el otro la poca popularidad de las medidas que no sufren este problema, demuestra a las claras que el debate sobre la evaluación de la recuperación de información (y por lo tanto, de los sistemas encargados de la misma), sigue abierto, ya que apenas se han producido avances efectivos desde las primeras propuestas de Cleverdon (con más de cuarenta años de antigüedad), que hayan conseguido consolidarse entre la comunidad científica y profesional.

Esto nos lleva a abogar por la necesidad de seguir investigando en este campo, aportando medidas alternativas a las tradicionalmente empleadas o, si esto no resultara posible (asumiendo que la imposibilidad actual se debe a motivos de cierta consideración e importancia), introduciendo los mecanismos correctores que sean necesarios para superar las debilidades que actualmente presentan.

5. BIBLIOGRAFÍA

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.

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.

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

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

Cleverdon, C.W., Mills, J. and Keen, E. M. *Factors determining the performance of indexing systems.* Cranfield, UK: Aslib Cranfield Research Project, College of Aeronautics., 1966. Volume 1:Design; Volume 2: Results.

Chowdhury, G. G. Introduction to modern information retrieval. London: Library Association, 1999.

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.

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

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: 1 marzo 2004].

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

Lancaster, F. W. and Warner, A.J. *Information Retrieval Today*. Arlington, Virginia: Information Resources, 1993.

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.

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

Diccionario de la Lengua Española [En línea]. Madrid: Real Academia Española, 2004. <http://buscon.rae.es/diccionario/drae.htm> [Consulta: 29 febrero 2004].

Rijsbergen, C.J. *Information Retrieval*. [En línea]. Glasgow, University, 1999. <http://www.dcs.gla.ac.uk/~iain/keith/> [Consulta: 29 febrero 2004].

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

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